UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Métodos de Clasificación Supervisada aplicados a datos de Radares Meteorológicos

Sofia Ruiz Suarez

Directora: Dra. Mariela Sued
Codirector: Dr. Luciano Vidal

5 de Julio 2016

Resumen

Predecir una variable cualitativa a partir de una observación (información disponible) puede ser considerado como un problema de clasificación o discriminación, en cuanto a que implica asignar una categoría a una cierta observación. Este problema puede plantearse de diferentes maneras y se encuentra presente en varios contextos de la vida humana. En particular, en el presente trabajo de tesis se abordará dicha problemática a partir del empleo de técnicas de *clasificación supervisada*. Las mismas parten de muestras pre-clasificadas con las que se aspira a *aprender* como discriminar (o clasificar) nuevas observaciones. Los métodos abordados en este trabajo toman como premisa la *Regla de Clasificación de Bayes*. la cual se basa en asignar la nueva observación a la clase con mayor probabilidad condicional estimada.

En esta tesis, nos abocaremos a buscar una respuesta para un problema de clasificación concreto: la discriminación entre ecos meteorológicos y no meteorológicos en imágenes de radar. Los radares meteorológicos son una herramienta fundamental para el monitoreo y diagnóstico de fenómenos meteorológicos de alto impacto social y económico como tormentas severas, caída de granizo, vientos destructivos e incluso la ocurrencia de tornados. Sin embargo, los datos captados por los mismos requieren de procesos de adecuación previos de la información para poder ser utilizados correctamenten en el diagnóstico e identificación de dichos fenómenos.

En particular, uno de los principales problemas es la presencia de datos generados a partir de ecos no meteorológicos (insectos, pájaros, edificios, suelo, entre otros). Trabajaremos con datos generados por un radar Doppler doble polarización en banda C (5.6cm) ubicado en la Estación Experimental Agropecuaria INTA Anguil en la provincia de La Pampa. Sobre una selección de casos aplicaremos diversos métodos de clasificación supervisada para luego analizar y comparar los resultados obtenidos.

Palabras claves: *clasificación supervisada, Regla de Bayes, radar meteorológico, eco no meteorológico.*

Agradecimientos

A mi papá, mi mamá y mi hermana, por acompañarme y apoyarme durante toda mi carrera.

A mis amigas Ani y Ro, con quienes compartimos poca cursada pero mucha vida. Gracias por ser parte de mi realidad facultativa, por alentarme siempre, por escucharme y por dejarme ser parte de las suyas.

A Mariano, porque siempre estuviste, me escuchaste, me acompañaste, me alentaste, y confiaste en mí.

A Juli y Vero, que hicieron los últimos años más amenos, las cursadas más divertidas, los tps menos tediosos y las tardes de estudio más relajadas y placenteras.

A Yami Barrera por darme aliento y energía en esta etapa final de la tesis.

A Daniela por involucrarse en el proyecto. Gracias por sus ideas y aportes.

Al equipo del Servicio Meteorológico, por darme esta oportunidad y por confiar en mí. Gracias a mis compañeros de oficina, por estar dispuetos a enseñarme el mundo de la meteorología, y por apoyarme en todo este proceso.

A Luciano por su pasciencia infiníta. Gracias por responder siempre a mis preguntas y dudas, por enseñarme, por regalarme tu tiempo y dedicación. Un lujo trabajar con un meteorólogo semejante.

A Mariela por animarse a este desafío. Por ser una excelente guía y docente. Porque su entusiasmo y dedicación son realmente motivadores y contagiosos. Gracias por tu apoyo, motivación y confianza en mí.

Índice general

1. Introducción **7**

2. Aplicación real: Datos de Radar Meteorológico **9**
2.1. Estructura de los datos y variables 11
2.2. Control de calidad de datos de radar y clasificación: síntesis breve de trabajos previos 13
2.3. Metodología 15

3. Regla de Bayes **19**
3.1. Estimación paramétrica de la densidad 21
3.1.1. Discriminante Lineal (LDA) 21
3.1.2. Discriminante Cuadrático (QDA) 24
3.2. Estimación no paramétrica de la densidad 27
3.2.1. Estimador núcleo de la densidad 27
3.2.2. Estimador Bayes Naive (BN) 31
3.2.3. K vecinos más cercanos 33
3.3. Métodos de validación cruzada 35
3.4. Laboratorio R 38
3.4.1. Modelo LDA 39
3.4.2. Modelo QDA 40
3.4.3. Estimador núcleo 40

4. Modelos de Regresión **43**
4.1. Modelo Logit 43
4.1.1. Estimación de los parámetros 44
4.1.2. Relación con modelo LDA 46
4.1.3. Modelo Multilogic 48
4.2. Estimación no paramétrica de regresión: K vecinos próximos (KNN) 48
4.3. Laboratorio R 51
4.3.1. Modelo Logit 51
4.3.2. K vecinos próximos: KNN 53

5. Resultados en datos meteorológicos y conclusiones **55**
5.1. Simulación 1 59
5.2. Simulación 2 61
5.3. Conclusiones 66

Bibliografía **67**

Capítulo 1

Introducción

El problema general de clasificación se puede plantear de la siguiente manera: se tiene un elemento que puede provenir de alguna categoría de entre dos o mas alternativas, y sobre el mismo se han medido una cierta cantidad de variables. Se busca entonces predecir a partir de los valores de estas variables, a qué categoría pertenece dicho elemento.

El problema de discriminación o clasificación se encuentra presente en varias situaciones de la vida humana. Se aplica en contextos muy diversos tales como la medicina, las finanzas, la biología, la psicología y la meteorología. Algunos ejemplos concretos podrían ser: reconocer un paciente como enfermo o sano a partir de resultados de estudios médicos, discriminar diferentes especies de pájaros según sus características físicas, clasificar a una empresa como en riesgo de quiebra o no, según sus valores de ventas, gastos y servicios, discriminar billetes falsos de verdaderos midiendo valores de textura, espesor y tamaño, determinar tipologías de personalidad a partir de dimensiones comportamentales, entre otras.

Existen varios enfoques distintos para este problema, por lo tanto propondremos revisar algunas de las técnicas existentes para resolverlo. Estudiaremos las técnicas que reciben el nombre de *técnicas de clasificación supervisada*. Las mismas parten de un conocimiento previo, es decir parten de una muestra de elementos pre clasificados que sirven de modelo para la discriminación de nuevas observaciones.

Según indica Breiman en su libro "Classification and regression trees"(1984) [2], una regla de clasificación es una forma sistemática de predecir en qué clase estará un nuevo individuo. Al momento de hacer una clasificación, dependiendo del problema, el propósito puede ser producir un clasificador lo más exacto posible, o bien, entender la estructura predictiva del mismo. Es claro que estos objetivos pueden en algunos casos no ser exclusivos, aunque muchas veces uno tendrá más énfasis que el otro.

El planteo estadístico es el siguiente: tenemos N elementos provenientes de J clases distintas $C_1, ..., C_J$ y, para cada uno de ellos hemos observado p características. Es decir, contamos con $x_n = (x_{n1}, ..., x_{np}, c_n)$ para $1 \leqslant n \leqslant N$ observaciones. A estos datos los llamaremos muestra de entrenamiento. Notaremos como n_j a la cantidad de elementos observados de la clase C_j ($\sum_{j=1}^{J} n_j = N$). Se trata entonces de encontrar una regla que permita asignar a una nueva observación $x = (x_1, ..., x_p)$, cuya procedencia se desconoce, alguna de las J clases.

En esta tesis, nos abocaremos a buscar una respuesta para un problema de clasificación concreto: *la discriminación de ecos meteorológicos y no meteorológicos en imágenes de radar meteorológico*. Los radares meteorológicos son una herramienta fundamental para el monitoreo y diagnóstico de fenómenos meteorológicos severos como lluvias intensas, granizo y tornados. Pero para poder hacer un uso apropiado de los datos captados por éstos, resulta fundamental adecuar la información recibida ya que los datos muchas veces están afectados por ecos provenientes de objetos que no son meteorológicos (insectos, pájaros, edificios, entre otros) y que introducen errores en los productos derivados de los radares que luego serán empleados en un sin número de aplicaciones (estimación de precipitaciones, seguimiento de tormentas, asimilación de datos en modelos numéricos, entre otros).

El presente trabajo de tesis se organiza como sigue: en el **Capítulo 2** presentaremos una breve descripción del funcionamiento de un radar meteorológico, los datos generados y utilizados, y una revisión bibliográfica sobre el problema de clasificación con radar. En el **Capítulo 3** estudiaremos técnicas de clasificación que son modeladas a partir de las funciones de densidad de las variables predictoras de cada clase, mientras que en el **Capítulo 4** discutiremos sobre los modelos de clasificación analizados como modelos de regresión. Por último en el **Capítulo 5** expondremos los resultados obtenidos, las conclusiones y los pasos a seguir posteriores a este trabajo.

Capítulo 2

Aplicación real: Datos de Radar Meteorológico

Los radares meteorológicos junto con otros sensores remotos como los satélites y los detectores de descargas eléctricas atmosféricas, son una herramienta primordial en todo el mundo para la detección, seguimiento y pronóstico de tormentas severas que pueden conducir a episodios de alto impacto social y económico en la población, debido a los fenómenos meteorológicos adversos que las acompañan como pueden ser el granizo, las lluvias intensas, las ráfagas de viento destructivas e incluso los tornados.

Un **radar** es un instrumento que permite detectar y ubicar objetos en espacio y tiempo por medio de ondas electromagnéticas. Su nombre deriva del inglés **RA**dio **D**etection **A**nd **R**anging. El *radar meteorológico* está adaptado para la detección de blancos distribuidos y difusos como áreas de precipitación (lluvia) o granizo, a diferencia de los radares de vigilancia que están diseñados para observar blancos puntuales como un avión o un barco.

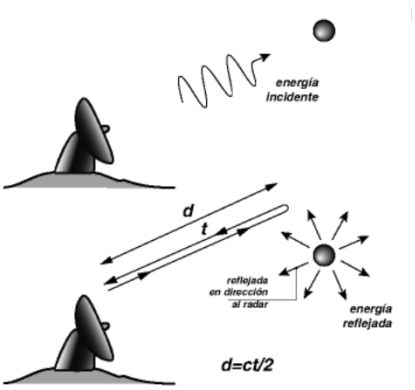

Figura 2.1: Esquema del funcionamiento del radar. *Imagen adaptada de www.crahi.upc.edu*

Un radar es un sensor denominado activo ya que posee su propia fuente de energía. El funcionamiento del mismo es similar al de un faro: transmite pulsos de energía electromagnética (ondas) en intervalos de tiempo regulares, concentrados en un haz angosto a través de una antena que rota. Los objetos "iluminados" por el haz de radar pueden absorber, dispersar y re-emitir parte de esa energía. La fracción de energía que regresa al radar, ingresa por la antena y es medida por el receptor. Conociendo el tiempo t entre la emisión y el retorno del eco, es posible determinar la distancia r del objeto (Figura 2.1). Por otro lado, mediante el análisis de la energía recibida se puede recuperar información sobre la naturaleza de los

objetos con los cuales el haz de radar está interactuando, mediante la ecuación de radar. Finalmente, esta información se utiliza para generar imágenes.

La señal (potencia) recibida por el radar se relaciona mediante una ecuación (comúnmente llamada "Ecuación de radar") con el denominado "factor de reflectividad" o "reflectividad" (Z). La reflectividad es una medida de la capacidad de los blancos de interceptar y devolver la energía del radar, y depende de los parámetros físicos del objeto, como su tamaño, forma, orientación, composición, etc.

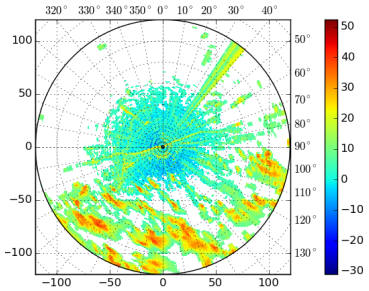

Figura 2.2: Campo de Reflectividad de la primera elevación

El factor de reflectividad radar es la suma de la sexta potencia de los diámetros de todos los blancos que dispersan la energía transmitida en la unidad de volumen muestreada. La dependencia de la sexta potencia significa que las partículas grandes predominan en el valor de reflectividad calculado.

Los valores típicos de reflectividad radar para llovizna o nubes que no producen precipitación oscilan entre 10^{-5} y 10; en el caso de lluvias muy intensas y granizo, el factor puede alcanzar 10^7. Como estos valores abarcan varios órdenes de magnitud y son difíciles de representar gráficamente con suficiente detalle, normalmente los productos radar emplean una escala logarítmica en decibelios de $Z(dBZ)$ que resulta más fácil de interpretar.

En la Figura 2.2 mostramos un ejemplo de cómo se grafica este campo de reflectividad. El sistema de escaneo es bastante sencillo: el radar gira sobre su propio eje escaneando el sitio a su alrededor, luego aumenta su elevación y vuelve a girar, y así una determinada cantidad de veces (Figura 2.3). Entonces se organizan los datos por elevaciones, pudiendo de esta manera reproducir un "volumen" con la información.

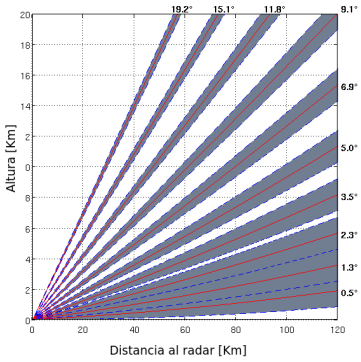

Figura 2.3: Estrategia de escaneo del radar metorológico de Anguil. En línea roja continua se indica el centro del haz para cada elevación de antena. En sombreado gris se indica el área cubierta por cada haz del radar cuyo ancho es de 1 grado tanto en la vertical como en la horizontal.

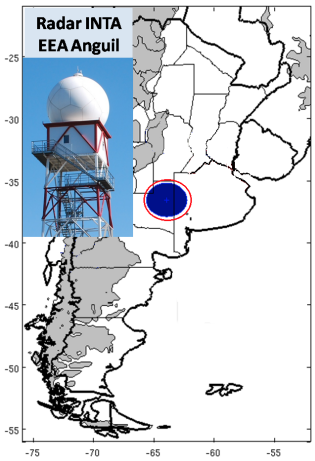

Figura 2.4: Ubicación geográfica del Radar Anguil. En sombreado azul se indica el área de cobertura para el rango máximo de 120km. Sobre el marco superior izquierdo: fotografía de la torre y radomo del radar.

El haz del radar durante su trayectoria no sólo interactúa con ecos meteorológicos (nieve, lluvia, granizo, etc) sino que también lo hace con ecos NO meteorológicos como edificios, aves, isectos terreno o polvo que terminan "contaminando" las imagenes resultantes. Además, las variaciones de temperatura y humedad en la parte baja de la atmósfera afectan la forma en la cual se propaga el haz del radar. Estos cambios hacen que el haz no se propague de forma rectilínea sino describiendo una curva, que puede ser mas o menos pronunciada. Este fenómeno se conoce como *Propagación anómala* (AP por sus siglas en inglés) y bajo determinadas condiciones atmosféricas puede generar ecos falsos en los datos de radar.

2.1. Estructura de los datos y variables

Para este trabajo de tesis se utilizaron datos generados por un radar meteorológico Doppler doble polarización banda C (frecuencia de 5 GHz) instalado en el predio de la Estación Experimental Agropecuaria INTA Anguil (La Pampa) que está ubicado en 36°22'22,9''S y 63°59'23,6''O. La Figura 2.4 muestra la ubicacion geográfica del radar.

Los datos de radar se organizan por elevación de antena (ángulo respecto del plano horizontal). Para cada una de ellas los mismos se disponen según "rango" y "azimuth", es decir según la distancia al radar y el ángulo de giro de la antena (Figura 2.5). Podriamos pensar que por cada elevación parametrizamos los datos en coordenadas polares, en donde r sería la coordenada radial, y θ la coordenada azimutal. En lo que respecta al caso específico del radar de Anguil, se tiene que éste genera un conjunto de datos cada 10 minutos para un total de 10 elevaciones de antena que varían entre 0,5 y 19,2 grados (Figura 2.3). La resolución en rango es de 0,25 km, con una distancia máxima de 120 km, y de 1° en azimuth.

Entonces, podemos pensar que por cada "volumen"de datos tenemos 10 matrices (una por cada elevación) de 360×480: por fila, el ángulo azimuth y por columna, el rango. A cada elemento de estas matrices lo llamaremos *pixel*.

Es importante aclarar que a medida que nos alejamos del radar, el ancho de las celdas aumenta debido a que aumenta el ancho del haz.

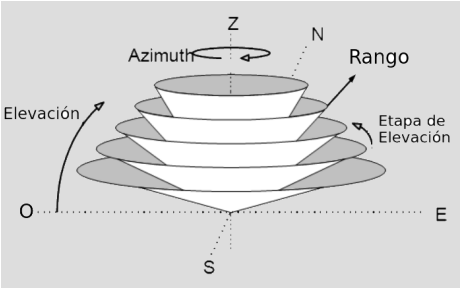

Figura 2.5: Escaneo volumétrico del Radar. *Imagen adaptada de cfa.aquila.infn.it*

Ahora bien, aun si la reflectividad sirve como un indicador de la estructura de los ecos, muchas veces no es concluyente por sí sola en la clasificación de los mismos. Sin embargo, los radares meteorológicos Doppler de doble polarización generan otras variables a parte de la reflectividad como son la velocidad Doppler, el ancho espectral, la reflectividad diferencial, el coeficiente de correlacion copolar, el diferencial de fase, entre otras. En particular, para el desarrollo del presente trabajo de tesis utilizamos junto con la reflectividad, las siguientes variables:

- **Velocidad Doppler (V)**: componente radial del movimiento de los objetos que está observando el radar. Pemite determinar solo si un blanco se está acercando o alejando del radar.

- **Reflectividad diferencial (Z_{DR})**: es la resta entre Z_H (reflectividad horizonal) y Z_V (reflectividad vertical), ambas en unidades de $[dBZ]$. Da información sobre la forma y orientación de los blancos. Se tiene que valores positivos de Z_{DR} están relacionados con objetivos de orientación horizonal (objetivos chatos). Por el contrario, los valores negativos de Z_{DR} se relacionan con objetos verticales (objetivo alargado). Si en cambio, el valor de Z_{DR} es cercano a cero, se asocia a un objetivo circular.

- **Coeficiente de correlación copolar (ρ_{HV})**: es una medida de la correlación entre la reflectividad horizontal y vertical devuelta por el radar. Este factor provee información sobre la diversidad en tamaño, forma y orientación de los blancos dentro del volumen iluminado por el radar. El mismo toma valores entre cero y uno. Valores cercanos a 1 están asociados a regiones con una distribución (en términos de tamaño y forma) mas o menos homogénea de los blancos. En cambio cuando su valor decae, indica un aumento de la heterogeneidad en tamaños y formas.

Conforme a trabajos anteriores sobre el tema (Schur (2003), Cho(2006), Gourley(2007), Berenguer(2006), entre otros), y según un análisis previo de los datos, basándonos en la capacidad de discriminación de cada una de éstas variables y teniendo en cuenta la calidad de los datos con los que contábamos, decidimos usar como variables predictoras:

- Velocidad doppler (Not: `V`). Se espera que para ecos de terreno y para AP sea cercana a cero.

- Coeficiente de correlación copolar (Not: `Rho`). Es un gran discriminador entre ecos meteorológicos y no meteorológicos. Valores bajos de ρ_{HV} se suelen relacionar con ecos no meteorológicos. De todas formas, objetivos como hielo o granizo grande podrían llegar a tener valores bajos de este factor.

- Desvio estándar de la reflectividad diferencial sobre un entorno de cada pixel (Not: `SZDR`): tomamos sobre una ventana de 5×5 (en rango y azimuth) el desvío estándar del Z_{DR}. Se esperan tener valores mayores para los ecos no meteorológicos.

2.2. Control de calidad de datos de radar y clasificación: síntesis breve de trabajos previos

Para poder utilizar cuantitativamente los datos generados por un radar meteorológico de una manera consistente y significativa, es necesario en primera instancia aplicar metodologías tendientes a minimizar las fuentes de error presentes en los datos con el objetivo de poder emplear los mismos rutinariamente en aplicaciones relacionadas con estimaciones cuantitativas de precipitación, detección de granizo, elaboración de pronósticos a muy corto plazo, asimilación de datos en modelos numéricos de pronóstico del tiempo, diagnóstico de tiempo severo, entre otras.

Villarini y Krajewski (2010) [15] plantean que las fuentes de error más comunes en los datos de radares (Figura 2.6) pueden originarse en:

1. Efectos relacionados con la distancia al radar (el ancho del haz aumenta con la distancia radial y por ende aumenta el volumen considerado).

2. Variabilidad en la distribución de los diámetros de las gotas (cambios en la relación Z-R utilizada para estimación de precipitación).

3. Bloqueo del haz

4. Presencia de ecos no meteorológicos (insectos, pájaros, interferencias electromagnéticas, entre otras).

5. Atenuación de la señal de radar por lluvia (pérdida de energía que sufre el haz de radar debido a la absorción y dispersión por parte de las gotas).

6. Errores de calibración del radar

7. Propagación anómala y ecos de terreno

8. Evaporación y deriva de la precipitación (La presencia de aire seco cercano a la superficie terrestre puede evaporar parte de la precipitación estimada por el radar. Además el viento cerca del suelo puede desvíar las gotas precipitantes e introducir errores al momento de comparar las estimaciones del radar con las observaciones en superficie).

9. Estructura vertical del sistema precipitante

10. Errores de muestreo temporal

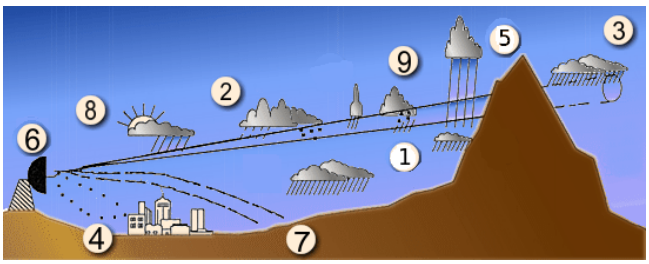

Figura 2.6: Esquema de fuentes de errores mas comunes en los datos de radar. Los números hacen referencia a la lista anterior. *Imagen adaptada de www.eumetcal.org*

En la actualidad existen numerosos y variados algoritmos de control de calidad de datos de radar orientados a la corrección de la mayoría de los errores antes mencionados. La mayor parte se basa en técnicas de lógica difusa, generación de índices de calidad de dato y combinación de variables polarimétricas. Todos ellos se concentran principalmente en abordar el problema de la identificación y separación de ecos no meteorológicos de aquellas regiones con presencia de eco meteorológico (precipitación, ya sea líquida o sólida).

En el ámbito de la meteorología este tema resulta de gran importancia, y es por eso que ya ha sido estudiado por varios autores en las últimas décadas. Se han utilizado diferentes técnicas, las cuales pueden dividirse en las que hacen un filtro directamente en el procesador de señal y aquellas que lo hacen sobre los datos procesados (nuestro caso). Sobre estas últimas a su vez se pueden distinguir tres tipos de abordajes: las que utilizan las llamadas técnicas de lógica difusa (Cho 2006 [3], Gourley 2006 [6] , Rico-Ramírez 2008 [11], entre otros), las que utilizan técnicas de árboles de decisión (Stein y Smith 2002 [14]), y las que lo hacen a partir de las llamadas redes neuronales (Lakshmanan 2010, Greku y Krajewski 2000). En su mayoría los trabajos realizados hacen un análisis pixel por pixel y se basan, según el tipo de radar y las variables disponibles, en el comportamiento de las mismas en las dos clases: meteorológica y no meteorológica.

El tema de la clasificación en radar meteorológico tal como ya mencionamos, no es reciente. Muchos autores, entre ellos Moszkowicz (1993) [9], Berenguer (2006) [1], Rico-Ramirez (2008), Gourley (2007) buscan estimar la probabilidad de cada punto de provenir de eco no meteorológico, y proponen clasificadores basados en la regla de clasificación de Bayes.

Moszkowicz (1993) propuso un clasificador basado en la regla de Bayes para detectar la presencia de propagación anómala. El autor tomó como campos de entrada en su modelo la elevación con mayor valor de reflectividad, el valor de ese máximo, y el gradiente horizontal de la reflectividad entre otros. Supone que las funciones de probabilidad condicional de cada campo sujeto a cada tipo de eco son gaussianas. A partir de allí estima los parámetros de dichas distribuciones con datos previamente seleccionados y clasificados de forma manual por un experto.

Mas recientemente, Stein y Smith (2002) proponen un algoritmo basado en un árbol de decisión que toma como variables predictoras ECHOtop (extensión vertical del eco del radar), SPINchange (variabilidad espacial de la reflectividad: mide cuán seguido el gradiente de este campo cambia a lo largo de una dirección en el espacio) y VertGRAD (gradiente vertical de la reflectividad) fo-

calizado mayormante en la eliminación de los ecos generados por el fenómeno de la *propagación anómala*.

El trabajo presentado por Schur (2003) [12] apunta a una clasificación de hidrometeoros (gotas de agua, nieve, cristales de hielo, granizo, etc), para la cual es necesario un filtrado previo de los ecos. Para ello sugiere un modelo de clasificación por lógica difusa en donde las variables analizadas provienen de radares de doble polarización. Estudia los campos de la reflectividad horizontal, la reflectividad diferencial, el coeficiente de correlación y la textura espacial de la reflectividad horizontal y de la diferencia de fase.

Cho (2006) también plantea apoyarse en la información espacial y temporal del eco de terreno y propagación anómala, asi como también la continuidad y textura del campo de reflectividad. Arma entonces un modelo de lógica difusa cuyas variables son: SDZ(desvío estándar de la reflectividad), VGZ (análogo a vertGRAD) y Vr (Valor absoluto de la velocidad radial). Para el armado de las denominadas "funciones de pertencia"(MFs) considera las frecuencias de cada una de las variables provenientes de 5 eventos precipitantes y 5 no precipitantes seleccionados con anterioridad. Para la validacion del método, utiliza medidas provenientes de radares de doble polarización.

Siguiendo esta línea, Berenguer (2006) presenta un trabajo focalizado hacia la eliminación de propagación anómala. Para cada pixel evalúa la posibilidad de que la medición haya sido afectada por dicho fenómeno asociando un valor entre 0 y 1. Con este objetivo, analiza las distribuciones de las frecuencias de cada campo estudiado condicional al tipo de eco. A partir de la regla de clasificación de Bayes, deriva la probabilidad condicional de cada pixel de estar afectado según el valor de los campos. Propone dos configuraciones distintas del algoritmo dependiendo si se trata de zonas cercanas al mar o no.

Gourley (2007) también trabaja sobre datos provenientes de radares de doble polarización. El autor estudia como es el comportamiento de ρ_{HV} y la variabilidad espacial (textura) de Z_{DR} y del diferencial de fase para cada tipo de eco. Además, tiene en cuenta la velocidad radial y la continuidad en el campo de la reflectividad horizontal. Sugiere un modelo de lógica difusa cuyas MFs se derivan a partir de una estimación no paramétrica de la densidad de cada campo en cada eco.

Rico-Ramirez (2008) presenta un clasificador en donde nuevamente la regla de Bayes y la lógica difusa son protagonistas. Trabaja con radares de banda C de doble polarización. Este modelo se diferencia de los algoritmos presentados por Gourley (2007) y por Schur (2003) en cuanto a la forma de calcular la textura de las variables. La propuesta es computar el desvío estándar de las mismas sobre una ventana fija de 3×3 o 5×5. Esta idea tiene como objetivo disminuir los efectos de borde en las cercanías de las regiones de precipitación.

2.3. Metodología

Como se mencionó en la sección anterior, los enfoques utilizados al día de hoy para encarar esta problemática son variados, pero la gran mayoría de ellos se basan en entender el comportamiento de ciertas variables en los diferentes blancos escaneados por el radar. Muchos de los modelos planteados en la literatura basan su conocimiento en análisis estadísticos y mucho otros lo hacen a partir de reglas establecidas según la experiencia de un experto (por ejemplo

se suelen implentar umbrales fijos de ciertas variables para clasificar entre las distintas clases). En muchos casos, se plantea combinar ambos abordajes en un mismo modelo (métodos de lógica difusa). En vista de los trabajos realizados previamente y según los resultados obtenidos por los mismos, para el presente trabajo de tesis decidimos trabajar con algunos métodos de clasificación netamente estadísticos, intentando en principio comprender el funcionamiento de los mismos, para luego poder comparar los distintos resultados provenientes de diferentes metodologías.

Según los métodos de clasificación que desarrollaremos en los capítulos 3 y 4, asignaremos a cada *pixel* en una de las dos clases: meteorológica (M) o no meteorológica (NM). Si bien la variable `SZDR` otorga una noción de la estructura espacial al análisis, consideramos necesario incorporar un proceso final a la clasificación obtenida por los modelos, de manera de darle mayor importancia a la caracterización espacial de los datos. Partiendo de la clasificación inicial del modelo, reasignamos algunos puntos teniendo en cuenta la clase de sus vecinos. Más especificamente, sobre cada *pixel* miramos una ventana de 3×3, si la mayoría de sus vecinos (la mitad $+1$) fueron asignados a su misma clase, entonces no se hace ningún cambio, sino, se cambia la clase previamente asignada.

Tal como introdujimos en la sección anterior, para la aplicación de estas técnicas de clasificación supervisada, es necesario contar con una muestra pre clasificada que permita brindar la información necesaria para armar el modelo. Con este objetivo, analizamos un conjunto de casos y realizamos una clasificación manual asistida por un meteorólogo experto en el área. De estos casos, seleccionamos algunos de ellos para ser utilizados como muestra de entrenamiento, y otros para ser utilizados como datos test (consideramos únicamente primera y segunda elevación).

Uno de los inconvenientes con el que nos hemos encontramos fue que muchas veces los datos con los que contábamos estaban incompletos. Es decir, a algunos *pixeles* les faltaba la información sobre alguna de las variables. Si esto se daba en la muestra de entrenamiento, simplemente se eliminaron aquellos puntos. En cambio, cuando el pixel con una variable faltante era parte del conjunto test, se clasificó dicho punto considerando el modelo con las covariables disponibles. Es decir, se ajustaron cuatro modelos, uno con las tres covariables y los otros tres con las covariables tomadas de a dos. Luego dependiento de la disponibilidad de variables en cada pixel se predijo su clase según alguno de estos cuatro modelos ajustados. Por ejemplo si para el pixel p, se contara con las variables (Rho,V), y $SZDR$ no estuviese disponible, se clasificaría el punto a partir del modelo que utliza como covariables (Rho,V). Si la falta de dato fuera en dos o más variables, se clasificará al punto como NaN (sin clase). En la Figura 2.7 esquematizamos de manera sencilla la metodología a seguir.

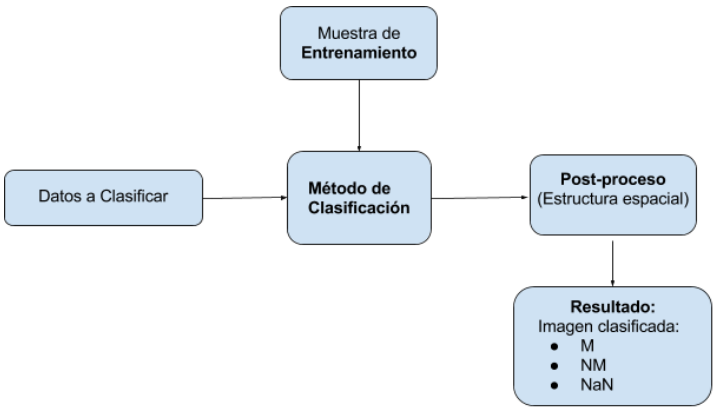

Figura 2.7: Diagrama de flujo sobre la metodología a seguir para el problema de clasificación en imágenes de radar

Cabe destacar que para la selección de los datos a analizar, tuvimos en consideración que los días pertenecieran a un período donde se contara con las variables polarimétricas y que las mismas fueran lo suficientemente confiables para su procesamiento. Además se buscó que estos casos representaran, dentro de las posibilidades, diferentes situaciones con configuraciones diferenciadas en relación a la distribución espacial de los ecos de radar. Las fechas analizadas fueron:

- 20 de Junio 2009
- 22 de Noviembre 2009
- 01 de Enero 2010
- 06, 04, 09, 20, 21 y 27 de Febrero 2010

Capítulo 3

Regla de Bayes

Para definir un buen clasificador, será necesario determinar previamente una medida para cuantificar la bondad del mismo. Consideremos el vector $Z \in R^{p+1}$, $Z = (X, C)$, con $X \in R^p$ y $C \in \{C_1, C_2, ..., C_J\}$. Entonces X corresponderá a p características de de cada individuo y C indicará la clase o población que le corresponde. Supongamos que se tiene una regla de clasificación dada por $h : R^p \to \{C_1, C_2, ..., C_J\}$ la cual asigna a cada elemento $x \in R^p$ alguna de las J clases existentes.

Como podemos suponer, una buena regla de clasificación será aquella que cometa la menor cantidad de equivocaciones posible. Dadas J clases $C_1, ..., C_J$ con probabilidades a priori $\pi_j = P(C = C_j)$ definimos la **Tasa de Error** para la regla h como:

$$L(h) = P(h(X) \neq C)$$

Observemos que la expresión anterior equivale a a considerar

$$L(h) = \sum_{j=1}^{J} P(h(X) \neq C|C = C_j)P(C = C_j) = \sum_{j=1}^{J} P(h(X) \neq C_j|C = C_j)\pi_j$$

Un buen clasificador, será aquel que posea Tasa de Error baja. Supongamos ahora que conocemos las probabilidades a priori de cada una de las clases. Entonces dado $x \in R^p$, *El Teorema de Bayes* nos permitiría calcular la probabilidad condicional de que $C = C_j$ dado que X toma valor x (para X discreto) mediante la fórmula

$$P(C = C_j|X = x) = \frac{\pi_j P(X = x|C = C_j)}{\sum_{i=1}^{J} \pi_i P(X = x|C = C_i)}. \tag{3.1}$$

$P(C = C_j|X = x)$ se denomina probabilidad a posteriori (de la clase C_j dada la observación x).

Si suponemos que $X|C = C_j$ tiene densidad f_j resulta entonces:

$$P(C = C_j|X = x) = \frac{\pi_j f_j(x)}{\sum_{i=1}^{J} \pi_i f_i(x)} \tag{3.2}$$

Clasificaremos entonces a x en la clase cuya probabilidad a posteriori sea máxima. Esta regla es conocida como la **"Regla de Bayes"**.

Teorema 3.1. *La Regla de Bayes, definida como*

$$h_B(x) = C_j \quad si \quad f_j(x)\pi_j = \max_i\{f_i(x)\pi_i : 1 \leq i \leq J\}, \tag{3.3}$$

minimiza la tasa de error de clasificación definida en (3.1), con valor: $L(h_B) = 1 - \int \max_j\{f_j(x)\pi_j\}dx$

Demostración. Sea h_B la Regla de clasificacion de Bayes definida en (3.3) y sea h cualquier otra regla.

$$\begin{aligned} P(h(X) = C) &= \sum_{j=1}^{J} P(h(X) = C_j | C = C_j)\pi_j \\ &= \sum_{j=1}^{J} \int_{h(x)=C_j} f_j(x)dx\pi_j \\ &= \int \sum_{j=1}^{J} I_{\{h(x)=C_j\}} f_j(x)\pi_j dx \end{aligned}$$

Para cada x, tenemos que

$$\sum_{j=1}^{J} I_{\{h(x)=C_j\}} f_j(x)\pi_j \leq \max_j\{f_j(x)\pi_j\}$$

La igualdad se alcanza si $h(x)$ iguala a aquella C_j para la cual $f_j(x)\pi_j$ es máxima. Entonces la Regla de Bayes, tiene la propiedad que dada cualquier otra regla

$$P(h(X) = C) \leq P(h_B(X) = C) = \int \max_j\{f_j(x)\pi_j\}dx$$

Luego,

$$L(h_B) \quad = \quad 1 - \int \max_j\{f_j(x)\pi_j\}dx \quad \leq \quad L(h)$$

□

A partir de la expresión (3.2), es posible notar que la Regla de Bayes puede ser formulada en términos de la densidad condicional de X dado $C = C_j$ (f_j), o bien puede ser caracterizada en término de las probabilidades a posteriori $P(C = C_j|X)$. Notemos que esta última resulta ser la esperanza condicional de la indicadora

$$P(C = C_j|X) = E(I_{\{C=C_j\}}|X)$$

En muchos problemas de clasificación, las probabilidades a priori ($\pi_j = P(C = C_j)$) son conocidas, y sino lo son, es fácil estimarlas tomando la proporción de cada una de las clases sobre los datos de entrenamiento. De esta manera, podemos dividir los métodos en dos, aquellos que abordan esta problemática modelando la función de densidad condicional de X en cada categoría y aquellos que lo hacen modelando la esperanza condicional es decir, modelando la esperanza de la indicadora dadas las covariables (métodos de regresión). En ambos casos es posible plantear modelos basados en técnicas *paramétricas* o en técnicas *no paramétricas.*

- **Métodos Paramétricos**: parten de suposiciones sobre la forma de la función f que se desea estimar y reducen el problema a estimar un conjunto de parámetros.

- **Métodos No Paramétricos**: no hacen suposiciones explícitas sobre la forma de f. Buscan una estimación que se aproxime lo más posible a los datos de muestra. El hecho de no hacer suposiciones con respecto a la forma de la f en principio es bueno, ya que si la forma verdadera de f es muy distinta de la supuesta, los resultados del modelo pueden ser muy malos. La desventaja que suelen tener estos métodos, es que es necesario un largo número de observaciones para obtener una buena aproximación de f.

Tenemos entonces los modelos paramétricos que son menos flexibles y los no paramétricos que son menos restrictivos. La razón para usar un tipo de modelo u otro viene de la mano del propósito de la clasificación. Si la idea es poder hacer una inferencia en la estructura del modelo, en general, los modelos paramétricos suelen ser mucho más interpretables. Si en cambio el interés viene de la mano de la exactitud en la predicción, sería lógico pensar que modelos menos restrictivos serán de mayor utilidad. Esto, no siempre es así, suele suceder que modelos más restrictivos dan predicciones más exactas. La razón de esto último tiene que ver con la sobre-estimación que aportan los métodos más flexibles.

En general, elegir el nivel adecuado de flexibilidad para un modelo, no es una tarea sencilla. Dependiendo mucho del problema y de los datos, un modelo se adaptará mejor que otro.

En lo que resta de esta sección, estudiaremos algunos de los métodos que apuntan a la estimación de la función de densidad condicional.

3.1. Estimación paramétrica de la densidad

3.1.1. Discriminante Lineal (LDA)

Comencemos suponiendo que la distribución de X en cada una de las clases C_j es normal, con media μ_j para cada clase pero con la misma matriz de varianzas Σ. Entonces la función de densidad de X condicionada a que $C = C_j$ resulta:

$$f_j(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j)\} \tag{3.4}$$

Utilizando esta función de densidad y aplicando logaritmo en la expresión (3.2), se pone de manifiesto que la Regla de Bayes asignará a una nueva observación la clase para la cual

$$\delta_j(x) = -\frac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j) + \log \pi_j \tag{3.5}$$

sea máxima.

Notemos que $(x-\mu_j)'\Sigma^{-1}(x-\mu_j)$ es la distancia de Mahalanobis entre x y la media de la clase C_j. Entonces, si las probabilidades a priori son iguales, asignaremos x a la clase cuya media esté mas próxima en el sentido de esta distancia.

Observemos también que si $\Sigma = I\sigma^2$, entonces, la regla anterior equivale a usar la distancia euclidea.

Bajo el supuesto de igual matriz de varianzas, la expresión 3.5 equivale a:

$$\delta_j(x) = x'\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \log \pi_j \tag{3.6}$$

Podemos intentar escribir este método interpretándolo geométricamente. Para ello analicemos la zona de la frontera, es decir cuando $\delta_k = \delta_i$ (asumiendo $\pi_i = \pi_k \forall i, k$), tenemos entonces, álgebra de por medio:

$$\Sigma^{-1}(\mu_i - \mu_k)'x = \Sigma^{-1}(\mu_i - \mu_k)'\frac{1}{2}(\mu_i + \mu_k)$$

Si consideramos las direcciones:

$$v_{ik} = \Sigma^{-1}(\mu_i - \mu_k) \tag{3.7}$$

entonces las frontera puede escribirse como:

$$v_{ik}'x = v_{ik}'\frac{1}{2}(\mu_i + \mu_k)$$

Ahora proyectemos las medias y el punto x que tratamos de clasificar sobre los v_{ik}. Nos encontraremos en la zona de la frontera cuando el punto x proyectado se encuentre a la misma distancia de las medias proyectadas. De lo contrario, asignaremos a x la clase cuya media proyectada esté mas próxima.

En la práctica, más alla de que hagamos suposiciones sobre la forma gaussiana de la función de distribución de X, será necesario estimar los parámetros de la misma usando la muestra de entrenamiento. Para ello llamaremos x_{ij} al vector fila que contiene las p variables del elemento i de la clase C_j. Entonces, el estimador de medias para cada una de las poblaciones será:

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i}^{n_j} x_{ij} \tag{3.8}$$

Para estimar la matriz de varianzas y covarianzas Σ, miraremos primero sobre cada una de las clases:

$$\hat{\Sigma}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \hat{\mu}_j)(x_{ij} - \hat{\mu}_j)' \tag{3.9}$$

Como supusimos que todas las poblaciones poseían la misma matriz de varianzas y covarianzas, su mejor estimación entonces será una combinación convexa de las $\hat{\Sigma}_j$ con pesos asociados a la cantidad de elementos n_j presentes de cada población:

$$\hat{\Sigma} = \sum_{j=1}^{J}\frac{n_j - 1}{n - J}\hat{\Sigma}_j \tag{3.10}$$

Al reemplazar por los estimadores en la expresión (3.6) obtenemos que $\hat{\delta}_j$ es una función lineal en x, siendo precisamente este hecho el que le otorga el nombre de "Discriminante Lineal" al clasificador. En la Figura 3.1 ilustramos el método para un caso muy sencillo. Dibujamos del lado izquierdo los histogramas de dos

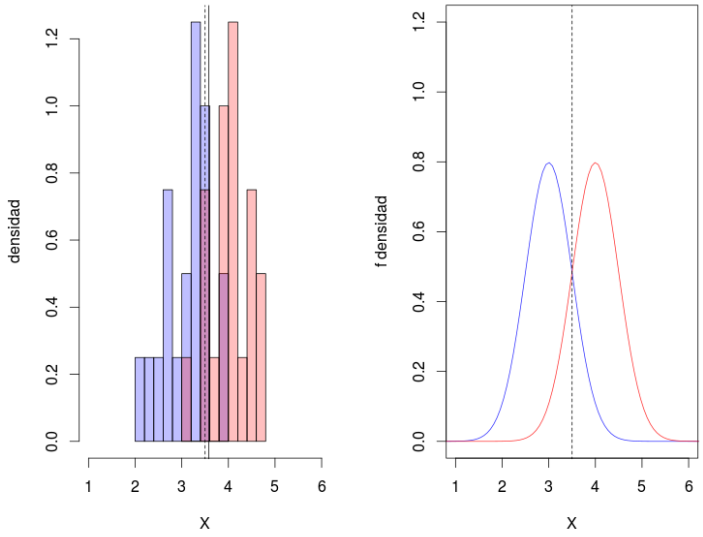

Figura 3.1: Derecha: histogramas de dos muestras (tamaño $N = 20$) con distribución Normal. Ambas poseen la misma varianza ($\sigma = 0{,}5$) pero distinta media ($\mu_1 = 3$ y $\mu_2 = 4$). Izquierda: las funciones de densidad. En ambos casos con línea punteada se indica la zona fronteriza de decisión de Bayes.

muestras de tamaño 20 provenientes de dos gaussianas con distintas medias e igual varianza. La línea completa indica la frontera de decisión obtenida a partir del modelo lineal. De lado derecho graficamos la densidad de cada de ellas. En ambos casos dibujamos con línea punteada la zona fronteriza de decisión dada por la regla de Bayes.

Según indica Peña en su libro "Análisis de Datos Multivariantes" [10] es conveniente testear si las medias de las clases difieren realmente o no, previo a utilizar esta regla.

Veamos ahora otra manera de obtener esta regla para el caso donde las probabilidades a priori son iguales. Supongamos primero que contamos únicamente con dos poblaciones ($J = 2$). El siguiente razonamiento fue presentado primero por Fisher en 1936. Éste, parte de la idea de disminuir la dimensión de las covariables considerando:

$$z = \alpha' X$$

En otras palabras, propone tomar como variable explicativa una combinanción lineal de los elementos de X. El objetivo entonces será hallar el vector α que separe lo mas posible a ambas clases. Nos referimos con esto a encontrar una dirección de forma que las medias proyectadas sobre la misma estén lo más separadas posible entre sí con relación a la variabilidad resultante de la proyección. Una vez hecho esto se procederá a realizar la clasificación teniendo en cuenta la variable unidimensional z en vez de vector p-dimensional X.

Tenemos que la media estimada de esta nueva variable z en cada una de las clases es $\hat{m}_i = \alpha' \hat{\mu}_i$ $(i = 1, 2)$, y que la varianza dada por $\alpha' \Sigma \alpha$ será la misma

para ambas clases. Estimamos esta última como $s^2 = \alpha'\hat{\Sigma}\alpha$. El criterio para hallar la mejor dirección de proyección, será el de maximizar la distancia de las medias comparadas con la variabilidad dada por los mismos datos. Con este propósito analizaremos el comportamiento de la siguiente función:

$$\phi(\alpha) = (\frac{\hat{m}_1 - \hat{m}_2}{s})^2 = \frac{(\alpha'(\hat{\mu}_1 - \hat{\mu}_2))^2}{\alpha'\hat{\Sigma}\alpha} \tag{3.11}$$

Derivando e igualando a cero, obtenemos:

$$\alpha = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

Notemos que el valor de α hallado, es el mismo que el de la dirección obtenida previamente en (3.7). Además tenemos que:

- $\alpha'\hat{\Sigma}\alpha = (\hat{\mu}_2 - \hat{\mu}_1)'\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$
- $\alpha'(\hat{\mu}_2 - \hat{\mu}_1) = (\hat{\mu}_2 - \hat{\mu}_1)'\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$

Por consiguiente al reemplazar en (3.11) se verifica que $\phi(\alpha)$ resulta ser la distancia de Mahalanobis entre las medias. Observemos que en este caso, no hicimos suposiciones previas con respecto a la forma de la distribución de X, pero igualmente llegamos a la misma regla que presentamos previamente.

Es posible extender este mismo razonamiento para el caso de mas clases ($J > 2$). En esta situación la propuesta es la de trabajar con un vector Z de dimensión dada por el mínimo entre la cantidad de clases y la dimensión de X, tomado como combinación lineal de las originales: $Z_i = \alpha_i X$. Nuevamente el objetivo será hallar la combinación lineal que permita obtener nuevas variables con máximo poder discriminante. Para ello se busca encontrar los α_i de forma que las medias proyectadas sobre los mismos estén lo más separadas posible entre sí (mayor variabilidad entre-grupos) y, al mismo tiempo, que cada una de ellas esté lo menos dispersa (menor variabilidad intra-grupos).

3.1.2. Discriminante Cuadrático (QDA)

En el método anterior supusimos que las observaciones dentro de cada una de las clases se distribuían como normales con distintas medias pero con la misma matriz de varianzas y covarianzas. Podemos ahora "relajar" este supuesto considerando el caso en que la matriz de varianzas y covarianzas sea diferente para cada población. El razonamiento es el mismo que antes, simplemente que ahora será necesario estimar dicha matriz por separado para cada clase. Bajo esta condición clasificaremos a una nueva observación en la clase que maximice la función:

$$\delta_j(x) = -\frac{1}{2}(x - \mu_j)'\Sigma_j^{-1}(x - \mu_j) - \frac{1}{2}\log|\Sigma_j| + \log\pi_j$$

Una vez más será necesario estimar los parámetros desconocidos. Esta vez: un vector de medias $\hat{\mu}_j$ (según (3.8)), y una matriz de varianzas $\hat{\Sigma}_j$ (según (3.9)) para cada clase.

Uno se podría preguntar en qué casos es mejor utilizar el método lineal que el cuadrático, es decir cuándo preocuparnos por distinguir diferentes matrices de covarianzas para las clases. La respuesta viene de la mano de la relación

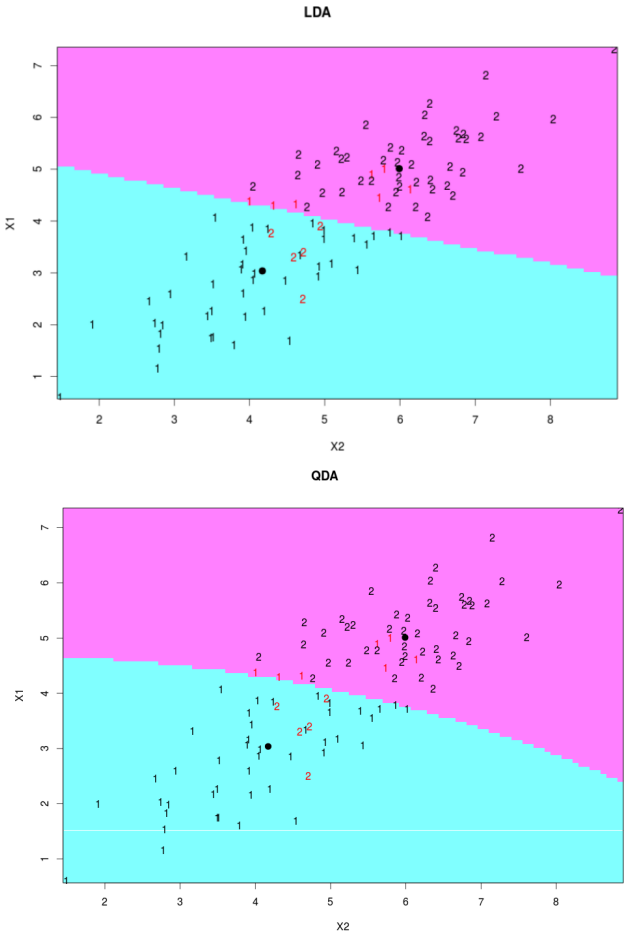

Figura 3.2: Problema de dos clases. A partir de una muestra normal bi-variada con medias $\mu_1' = (0, -1)$ y $\mu_2' = (0, 1)$ y $\Sigma_1 = \Sigma_2$, se computaron ambos modelos. Primera fila: modelo lineal. Segunda fila: modelo cuadrático.

sesgo-varianza presente en cada uno de los métodos. Por un lado, tenemos que el clasificador cuadrático es más flexible, pero a la vez necesita de la estimación de una mayor cantidad de paramétros que el lineal. Con J clases y p covariables es necesario estimar $Jp + Jp(p+1)/2$ parámetros, mientras que para el lineal sólo son necesarios $Jp + p(p+1)/2$. Así con 4 clases y 10 covariables debemos estimar 260 parámetros para el clasificador cuadrático, en cambio para el lineal alcanza solamente con 95. Por otro lado, el modelo lineal si bien es menos flexible es más estable en lo que refiere a cambios en el resultado por modificaciones en la muesta de entrenamiento. Es decir, el clasificador lineal posee menor varianza pero mayor sesgo, contrariamente con lo que ocurre en el caso del cuadrático. Es por esto que cuando la muestra de entrenamiento no es muy grande suele ser mejor considerar el modelo lineal, permitiendo así que el asunto de la varianza no sea un obstáculo.

En las Figuras 3.2 y 3.3 ilustramos el comportamiento del modelo cuadrático

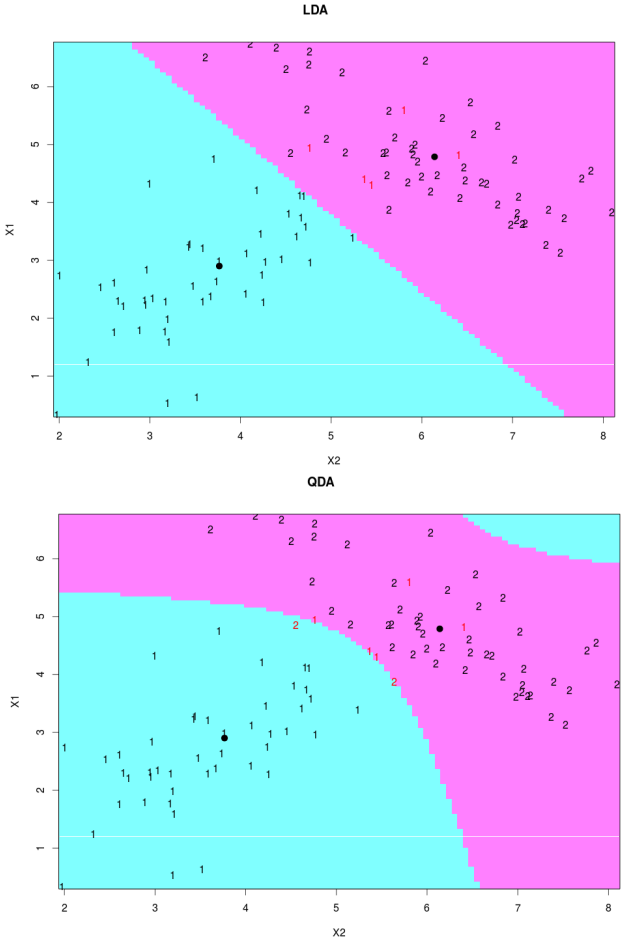

Figura 3.3: Problema de dos clases. A partir de una muestra normal bi-variada con medias $\mu_1' = (0, -1)$ y $\mu_2' = (0, 1)$ y $\Sigma_1 \neq \Sigma_2$, se computaron ambos modelos. Primera fila: modelo lineal. Segunda fila: modelo cuadrático.

y del modelo lineal en dos escenarios distintos. En la Figura 3.2, partimos de dos muestras gaussianas bi-variadas con la misma matriz de covarianzas para ambas, y con una correlación de 0,7 entre X_1 y X_2. Para el caso de la Figura 3.3, partimos también de muestas gaussianas, pero esta vez con matrices de covarianzas diferentes. En este caso, para la clase C_1 existe una correlación de 0,7 entre X_1 y X_2, mientras que para la clase C_2 la correlación es de $-0{,}7$. Es claro para que en ambos escenarios, el modelo que parte de suposiciones verdaderas es el que se adapta mejor a los datos. De todas formas, los dos parecen funcionar bastante bien cuando los supuestos de los mismos no son correctos.

3.2. Estimación no paramétrica de la densidad

Vamos a querer ahora estimar las funciones de densidad $f_j(x)$ con aún menos suposiciones. Consideremos el caso en que nada sabemos de las funciones de densidad del vector X en cada una de las clases. Buscaremos entonces estimarlas utilizando únicamente la muestra de entrenamiento a partir de técnicas no paramétricas. Así pues, una vez hecho esto, podremos reemplazar estas estimaciones en (3.2).

Planteamos entonces el problema de estimar la función de densidad f de una cierta variable aleatoria X a partir de una muestra de tamaño n proveniente de la misma. Llamaremos x_i a los elementos de dicha muestra ($1 \leq i \leq n$). Con este objetivo, presentamos algunas técnicas de estimación no paramétrica de la densidad.

3.2.1. Estimador núcleo de la densidad

Para comenzar, supongamos $p = 1$. Como primera aproximación uno podría plantear una expresión de la forma:

$$\hat{f}(x) = \frac{\#x_i \in \Lambda(x)}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I_{[-1,1]} \left(\frac{x - x_i}{h} \right)$$

En donde por $\Lambda(X)$ nos referimos a un entorno de X de tamaño h. A éste último lo llamaremos "parámetro de suavizado". Notemos que en esta expresión figura la función de densidad de una variable aleatoria con distribución $U(-1, 1)$. Es evidente que este estimador no es suave y por esta razón sustituimos el término correspondiente a la densidad uniforme por alguna otra función de densidad a la que llamaremos núcleo K:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i}{h} \right) \tag{3.12}$$

Por la manera en que está definido el estimador, es claro que éste heredará las propiedades del núcleo. Debido a esto, se suele tomar K unimodal, simétrico y suave. Una forma interesante de interpretar el estimador núcleo es pensarlo como una convolución entre la función empírica de los datos y la distribución cuya densidad se sigue por $K_h(x) = \frac{K(x/h)}{h}$.

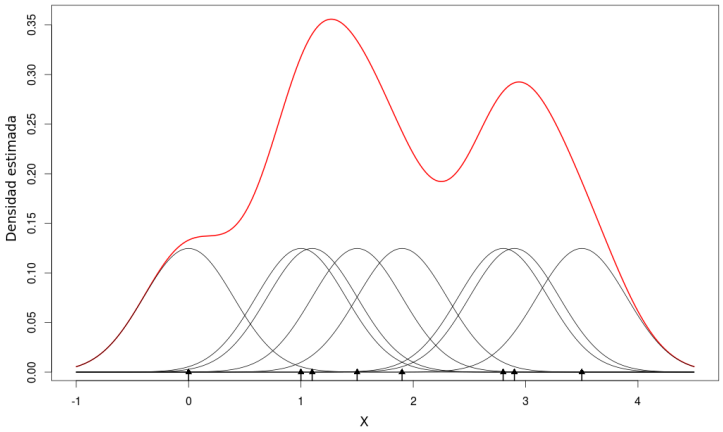

Figura 3.4: Estimación no paramétrica de la densidad a partir de una muesta de tamaño 8 con núcleo gaussiano. En rojo, el resultado de la estimación, en gris la función núcleo. Los puntos sobre el eje x indican los elementos de la muestra.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = (\hat{F} * K_h)(x)$$

Para entender mejor esta interpretación, consideremos la variable aleatoria $X_K = x_c + h\xi$. Donde x_c es el resultado de elegir con equiprobabilidad un elemento de la muestra, y ξ un ruido generado a partir de una variable aleatoria con densidad K. Tenemos entonces que la v.a. X_K tiene función de densidad igual a $\hat{f}(x)$. Así pues, $\hat{f}$ resulta una mezcla de n densidades con peso $\frac{1}{n}$, con la forma del núcleo K, reescaladas por h y centradas en cada una de las observaciones x_i. En la Figura 3.4 ilustramos este estimador para el caso de una muestra de tamaño ocho con núcleo gaussiano.

Notemos entonces que es preciso definir dos parámetros: uno la función núcleo K, y otro el parámetro de suavizado h. Usualmente se suele tomar como K un núcleo gaussiano.

La elección del parámetro de suavizado es de suma importancia para el buen funcionamiento del estimador. Como h controla la concentración de peso alrededor de cada x_i, se tiene que valores chicos de h darán lugar a que únicamente las observaciones mas cercanas al punto donde queremos estimar sean relevantes en la estimación. De lo contrario, si tomamos valores grandes de h, observaciones más lejanas influirán también en la estimación. Sucede entonces que un parámetro de suavizado pequeño le dará poco sesgo pero mayor varianza al estimador, mientras que si lo tomamos grande tendremos menor varianza pero aumentará el sesgo. Nuevamente nos encontramos ante un compromiso del tipo sesgo-varianza. En la Figura 3.5 graficamos el efecto que produce la variación del parámetro de suavizado en el comportamiento del estimador.

Se han desarrollado varias técnicas para la selección del parámetro de suavizado, que no sólo pueden ser utilizadas para el caso del estimador núcleo, sino que también pueden ser fácilmente aplicables para otros estimadores. Varias de estas técnicas utilizan los llamados *métodos de resampleo*. Estos se basan, tal como su nombre lo indica, en tomar nuevas muestras de las originales de modo

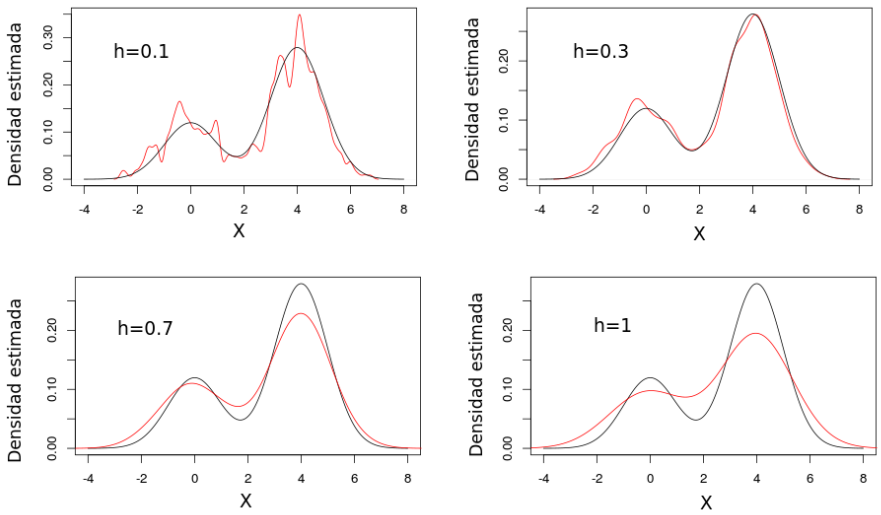

Figura 3.5: Efecto del parámetro de suavizado en el comportamiento del estimador. Partimos de una muestra de tamaño 500 de una mixtura de normales. Tomamos distintos valores de h=0.1,,0.3, 0.7, 0.1. En color negro la verdadera función de densidad, en rojo las estimaciones hechas según cada valor de h.

de poder ajustar algún modelo de interés sobre cada una de ellas y así obtener información adicional acerca del modelo en cuestión. Si bien los métodos de resampleo tienen un costo computacional alto, son realmente útiles, y los grandes avances en materia de poder de cómputo permiten un buen uso de los mismos.

Con el objetivo de hallar un valor adecuado para el parámetro de suavizado será necesario cuantificar la bondad de un estimador. Para ello definamos algunos criterios que nos permitan hacerlo de forma más precisa. Una forma de evaluar localmente un estimador de $f(x)$ es el *Error Cuadrático Medio*:

$$MSE(\hat{f}(x)) = E[(\hat{f}(x) - f(x))^2] = (Sesgo(\hat{f}(x)))^2 + Var(\hat{f}(x))$$

Para medir de alguna forma la distancia entre el estimador y la función que queremos estimar, integrar el MSE sobre el soporte de f (suponemos acotado), puede ser una buena opción. De esta manera obtenemos el valor de lo que se conoce como *Error Cuadrático Integrado*(ISE):

$$ISE(\hat{f}(x)) = \int_{sop(f)} (\hat{f}(x) - f(x))^2 dx$$

Observemos que el ISE es una variable aleatoria ya que depende de la muestra. Tomando valor esperado, podemos definir el *Error Cuadrático Integrado Medio*(MISE) como:

$$MISE(\hat{f}(x)) = E\left[\int_{sop(f)} (\hat{f}(x) - f(x))^2 dx\right]$$

A los términos principales de la expresión asintótica del MISE se les llama AMISE, y será éste valor, el que nos ayude a seleccionar el valor del parámetro de

suavizado. A continuación presentamos algunas de las técnicas más utilizadas, que se basan esencialmente en la minimizacion del valor del AMISE.

- **Regla de Referencia a la Normal**: supone que f proviene de una distribución normal. Primero utilizando un núcleo gaussiano y luego utilizando un núcleo distinto K, se calcula el h que minimice el AMISE. Se tiene que el cociente entre ambas ventanas óptimas no depende de la función de densidad desconocida. El valor de h para el caso del núcleo gaussiano es $h_{K_N} = 1{,}059\sigma n^{-1/5}$.

- **Método Plug-in**: hoy en día es el método que da mejores resultados prácticos. Al calcular la expresión del h que minimiza el AMISE se tiene que ésta depende en parte de la derivada de la función de densidad desconocida. Se propone entonces sustituir este fragmento por una estimación realizada de forma iterativa sobre las derivadas de f y que parte de la regla de referencia a la normal.

- **Método de validación cruzada:** parte directamente de la minimización del valor del MISE. Minimizar el valor del MISE, equivale a minimizar la expresión:

$$E_{\mathbf{X}}\left(\int \hat{f}(x,h)^2 dx - 2\int \hat{f}(x,h)f(x)dx\right)$$

  en donde $\mathbf{X}$ es una m.a.s de X de tamaño n y cuya densidad conjunta es f^n. Es claro que esta expresión depende de la función desconocida f, por lo tanto no es posible evaluarla. De todas maneras, lo que si es posible es construir un estimador que dependa de h. Se puede probar que $E_{\mathbf{X}}\left(\int \hat{f}(x,h)f(x)\right) = E_{\mathbf{X}X}(\hat{f}(X,h))$. Se propone entonces estimar $\int \hat{f}(x,h)f(x) = E_X(\hat{f}(X,h))$ mediante el estadístico $\frac{1}{n}\sum_{i=1}^{n}\hat{f}(X_i,h)$ utilizando técnicas de validación cruzada. Éstas se basan en separar los datos de forma que para cada iteración se tenga un sólo dato de prueba y se utilice todo el resto para la estimación. A partir de lo anterior se arma un nuevo estadístico que depende solo de h:

$$\int \hat{f}(x,h)dx - 2\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i,h)$$

  Donde $\hat{f}_{-i}$ es el estimador núcleo que considera como muestra de entrenamiento todos los elementos menos el i-ésimo. Se toma entonces como parámetro de suavizado aquel que minimice la expresión anterior.

Como bien ya dijimos, la elección de h es realmente significativa en lo que respecta al buen funcionamiento del estimador. Se han realizado una gran cantidad de trabajos que estudian este problema y lo analizan desde distintas ópticas[1]. Como nuestro objetivo final está en la clasificación, no ahondaremos mucho más en este tema. Es necesario aclarar igualmente que la elección del parámetro de

[1]Entre otros se puede consultar: "Density Estimation for Statistics and Data Analysis"B.W.Silverman [13] y, "Curso de Moldelos no Paramétricos"Pedro Delicado [4].

suavizado es de mayor importancia que la del núcleo ya que para éste último, la misma obedece esencialmente a cuestiones computacionales o de suavidad.

Pasemos ahora al caso de más dimensiones ($p > 1$). Generalizamos el estimador núcleo de la expresión (3.12):

$$\hat{f}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K_p(H^{-1}(x - x_i))$$

En donde ahora K_p es un núcleo p-dimensional y H una matriz de $p \times p$ no singular, ambos de dimensión p. Notamos con $|H|$ al módulo del determinante de esta matriz. En este caso, H representa una rotación de los datos y cambios de escala en cada variable. Al igual para el caso de una dimensión, la elección de H es de suma importancia. Hay tres formas usuales de elegir esta matriz: tomarla como un múltiplo de la identidad, tomarla como una matriz diagonal, o bien tomarla como un múltiplo de la matriz de covarianzas de los datos. En lo que respecta a la elección del núcleo, suele ser frecuente la opción del núcleo producto, en cuyo caso el estimador resulta:

$$\hat{f}(x) = \frac{1}{n \prod_{k=1}^{p} h_k} \sum_{i=1}^{n} \prod_{k=1}^{p} K\left(\frac{x_k - x_{ik}}{h_k}\right)$$

En la expresión anterior elegimos H como una matriz diagonal.

Ejemplo 3.1. *Vamos a utilizar el paquete "Boston"de* `R` *que contiene un conjunto de datos referido a la vivienda en 506 barrios de Boton en 1978. Utilizaremos como variables explicativas a la variable* `lstat` *que indica el porcentaje de población con estatus social en la categoría inferior, y a la variable* `age` *que mide el porcentaje de viviendas construidas antes de 1940. Dividiremos a los datos en dos clases según la variable* `rm` *(que indica el número de habitaciones promedio por vivienda). La mediana de la variable* `rm` *es* 6,2085*. Diremos que un barrio pertenecerá a la clase C_1 si* `rm` $<$ 6,2085*, de lo contrario, el barrio será de la clase C_2. A partir de esta división podemos estimar las probabilidades a priori por* 0,5 *para ambas clases. En la Figura 3.6 dibujamos las estimaciones de las densidades bidimensionales (*`lstat`,`age`*) hechas a partir de un núcleo bivariado normal y con ventana dada por la regla de referencia de la normal para cada una de las clases. Dejamos para la sección 3.3 la aplicación de la técnica de validación cruzada basada en la clasificación para la selección del parámetro de suavizado.*

En la Figura 3.7 representamos de forma gráfica la regla discriminante: tomamos la diferencia entre la estimación de la densidad de cada una de las clases ($\hat{f}_1 - \hat{f}_2$), cuando ésta es positiva asignaremos como clase C_1, de lo contrario asignaremos C_2. Con rojo representamos la zona fronteriza de desición (claramente no lineal).

3.2.2. Estimador Bayes Naive (BN)

Existe un inconveniente significativo en lo que respecta a la estimación no paramétrica de la densidad cuando el valor p es grande. Con esto nos referimos

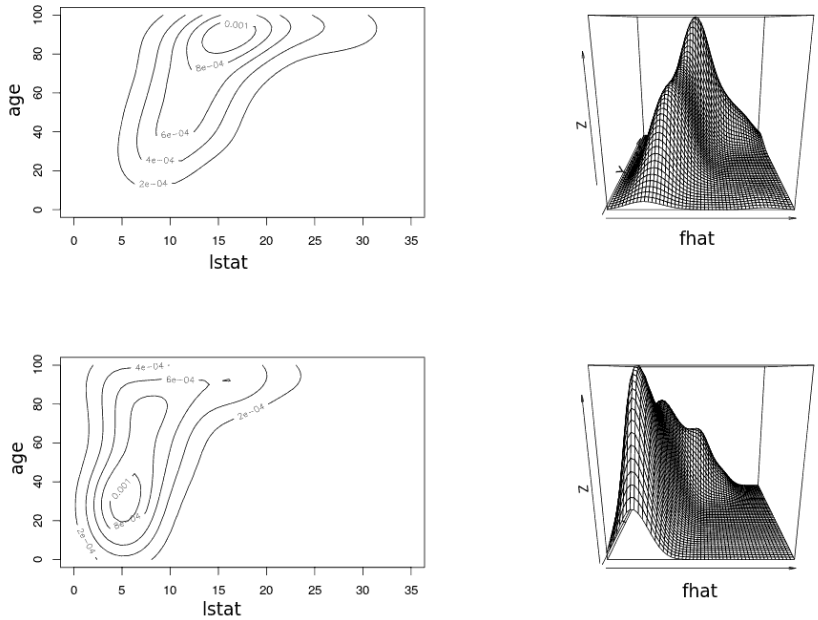

Figura 3.6: Estimacion de la densidad para cada clase. Primera fila clase C_1. Segunda fila clase C_2.

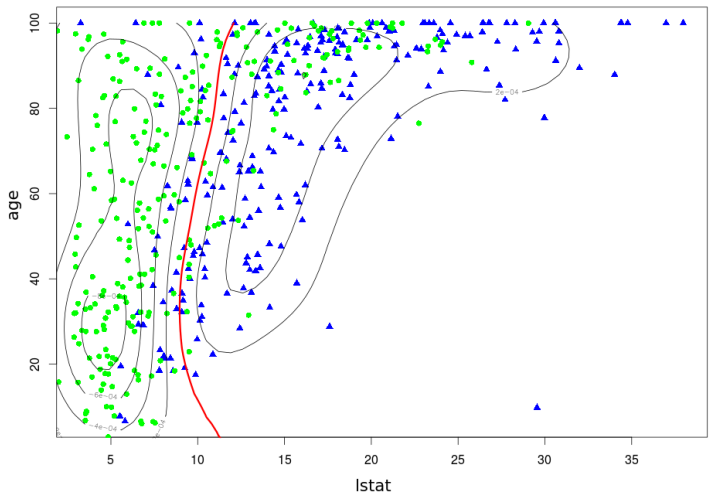

Figura 3.7: Diferencia de las densidades estimadas. En azul los puntos correspondientes a la clase C_1. En verde los puntos correspondientes a la clase C_2. En rojo: frontera de desición

que a valores mayores de p, la precisión de la estimación será menor. Esto se debe a que en dimensiones grandes los entornos de un punto se encuentran prácticamente vacíos de los puntos observados de la muestra. Este problema se conoce como la "Maldición de la dimensionalidad".

Como nuestro objetivo final está en la clasificación, conocer de forma precisa la densidad de los datos resulta inncesario, y si la dimensión de los datos es grande, ya vimos que puede ser problemático. Una forma de eludir este problema es construir un estimador bajo la hipótesis de que las p covariables son independientes. Luego estimamos:

$$\hat{f}(x) = \prod_{k=1}^{p} \hat{f}_k(x_k)$$

Este método se lo denomina **"Método de Bayes Naive"**. Si bien el mismo parte de una suposición que no suele ser correcta en la mayoría de los casos, no sólo simplifica de forma evidente el estimador sino que también en la práctica da muy buenos resultados. Hastie en su libro "The elements of statistical learning" [8], atribuye este comportamiento al hecho de que si bien las estimaciones de las densidades de cada una de las clases pueden ser sesgadas, este sesgo puede no interferir demasiado en la estimación de las probabilidades a posteriori, sobre todo en las zonas fronterizas de decisión.

3.2.3. K vecinos más cercanos

Este estimador fue también llamado "Densidad Globo"(Balloon density ó Balloonogram). Para deducir la construcción del mismo, partamos de la relación entre la función de densidad $f(x)$ y la función de distribución $F(x)$.

$$f(x) = \frac{d}{dx}F(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}$$

Si estimamos $F(x)$ con la función empírica $\hat{F}(x)$ y notamos $k(x,h) = \#\{x_i : x_i \in [x-h, x+h]\}$, tenemos:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} = \frac{k(x,h)/n}{2h}$$

Ahora bien, podríamos en lugar de dejar que el numerador sea el que dependa de x y h, fijar un valor entero k y hacer que sea h en el denominador el que varíe en x, es decir $h = h(x,k)$. Si consideramos $h(x,k)$ a la distancia entre x y el k-ésimo vecino más cercano, podemos escribir:

$$\hat{f}(x) = \frac{k/n}{2h(x,k)} \tag{3.13}$$

Resulta sencillo ver que la expresión de arriba puede escribirse como:

$$\hat{f}(x) = \frac{1}{nh(x,k)} \sum_{i=1}^{n} K_U\left(\frac{x - x_i}{h(x,k)}\right)$$

Donde por K_U nos referimos al núcleo uniforme en $[-1,1]$. De la misma manera que sucede con el estimador núcleo de la densidad, $\hat{f}$ hereda las propiedades del

núcleo. Análogamente a lo que hicimos en la sección anterior, es posible reemplazar el núcleo uniforme por algún otro núcleo K. De esta forma obtenemos la expresión general del estimador por vecinos mas cercanos:

$$\hat{f}(x) = \frac{1}{Nh(x,k)} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h(x,k)}\right)$$

Resulta evidente que entre este estimador y el estimador núcleo existe una relación. Para esclarecer este hecho consideramos el núcleo:

$$K(x) = \begin{cases} 1/2 & si|x| \leq 1 \\ 0 & sino \end{cases}$$

Calculemos ahora el estimador núcleo con K y con ventena h:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{2nh}\{\#x_i : |x - x_i| \leq h\}$$

Tenemos entonces que para algún valor de h la expresión dada por (3.13) resulta ser el estimador núcleo evaluado en x y con ventana $h(x,k)$.

Cuando se trata de estimar la densidad en un único punto, la estimación vía núcleo y la estimación por vecinos más cercanos no difieren. Con esto último nos referimos a que dado un entero k existirá un valor de h para el cual se obtenga el mismo resultado usando el estimador por vecinos más cercanos con parámetro k, que utilizando el estimador núcleo de la densidad con parámetro de suavizado h. Esto cambia cuando se trata de una estimación sobre una cierta cantidad de puntos o bien cuando lo que se busca es construir una estimación de la densidad de forma completa.

Es importante aclarar que el estimador dado a partir de los vecinos mas cercanos, no siempre es una función de densidad. Además puede resultar abrupto para algunos valores de x ya que éste es propenso al ruido local. Este hecho no mejora demasiado con la intervención de un núcleo suave en la estimación. Esto último se debe a que las irregularidades locales son causadas por la dependencia del estimador de la función h, la cual no es diferenciable.

En la Figura 3.8 graficamos los dos estimadores para el caso de una mezcla de normales con medias 0 y 4 , varianza 1 para ambas y con proporciones $0,3$ para la primera y $0,7$ para la segunda. Partiendo de una muestra de tamaño 500, en color negro graficamos la densidad verdadera. En color azul el estimador núcleo con parámetro de suavizado dado por la regla de referencia a la normal y con núcleo gaussiano. Para el caso del estimador por vecinos más cercanos, utilizamos también un núcleo gaussiano y $k = 45$, el cual graficamos con color rojo. Se puede ver que, aunque ambos fueron construidos a partir del núcleo gaussiano, la línea roja no resulta suave.

De la misma manera que planteamos para el caso del estimador núcleo, uno se podría preguntar cual sería una elección adecuada del parámetro de suavizado k. En este caso, al buscar el valor óptimo (en el sentido de minimizar el MISE), resulta que éste dependerá del valor de x en donde queremos estimar. Esto último le otorga mayor dificultad al problema en cuestión. Se puede probar que para un valor de x fijo el valor óptimo para k es del orden de $n^{4/5}$.

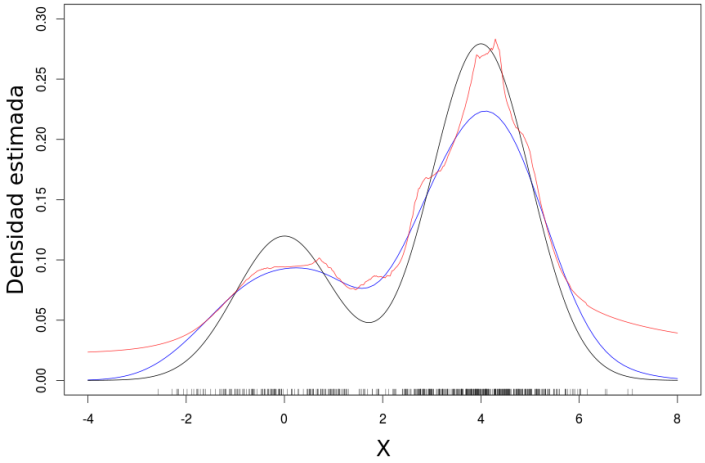

Figura 3.8: Estimaciones de la densidad. Partiendo de una mezcla de normales: En negro la densidad verdadera. En azul la densidad por estimación via núcleo con h por regla de referencia a la normal. En rojo el estimador por k-vecinos mas cercanos con k=45.

3.3. Métodos de validación cruzada

Como ya expusimos previamente, un buen clasificador, será aquel que posea Tasa de Error baja, es decir, será aquel para el cual $P(h(X) \neq C)$ sea lo menor posible. Uno podría evaluar este índice sobre los datos de entrenamiento, pero el valor obtenido no representaría realmente la bondad del modelo ya que se estaría evaluando con los mismos datos con los que fue construido el clasificador. Por esta razón, resulta más interesante computar este índice sobre datos nuevos. Hay situaciones en que no se dispone de nuevos datos, y en esos casos una solución puede ser recurrir a los llamados *"Métodos de validación cruzada"* o *"Cross validation methods"*. El mecanismo de estos métodos se basa esencialmente en excluír un subconjunto de la muestra de entrenamiento en el proceso de ajuste, y luego utilizar dicho subconjunto para testear el modelo resultante.

El asunto será entonces determinar que subconjunto extraer: de que manera, que tamaño y cuales seleccionar. Supongamos que contamos con una muestra de tamaño n. Como primera apoximación, podríamos dividir de manera aleatoria a la muestra en dos mitades: consideramos una mitad para ajustar el modelo, y la otra mitad para testearlo. Si repetimos el procedimiento anterior dividiendo de manera aleatoria la muestra por la mitad, obtendríamos diferentes valores para la tasa de error de clasificación. Si bien este procedimiento resulta sencillo de interpretar, cuenta con un poblema fundamental: no se utiliza la totalidad de los datos disponibles para el ajuste. En general, cuantos más sean los datos con lo que se dispone para ajustar algún modelo estadístico, mejor será el funcionamiento del mismo. De esta manera, al considerar unicamente la mitad de la muestra para el ajuste, se sobreestima la tasa error de clasificación.

Para solventar este inconveniente surge lo que se denimina *"Leave-one-out*

Cross validation" (LOOCV). En este caso, en vez de partir al conjunto de datos en dos partes iguales, se considera a una única observación para testar el modelo y todo el resto para el ajuste. Si esto lo hacemos con cada una de las observaciones disponibles, obtendremos n valores distintos para la tasa de error de clasificación: $L_1 ..., L_n$. De esta manera podemos estimar la tasa de error de clasificación como el promedio de todos ellos:

$$\hat{L}_{(n)} = \sum_{i=1}^{n} \frac{1}{n} L_i$$

Observemos que el estimandor LOOCV tiene varias ventajas sobre la aproximación anterior: en primer lugar, posee menos sesgo, ya que como bien dijimos utiliza más elementos para ajustar el modelo. En segundo lugar no hay nada aleatorio en esta técnica, es decir, si se aplicara más de una vez, el resultado no cambiaría ya que los subconjuntos a excluir en cada paso están pre-determinados. De todas maneras, la técnica LOOVC poseé un problema no trivial y que es necesario tener en cuenta sobre todo cuando la muestra tiene tamaño grande: implica alto costo computacional. Hay situaciones en donde no solo la muestra es grande sino que además el tiempo que lleva ajustar el modelo en cada paso es considerable, en estos casos resulta practicamente imposible aplicar LOOCV.

Otra alternativa posible es seleccionar un valor $2 \leqslant K \leqslant n$ y obtener una partición aleatoria de la muestra en K subconjuntos de manera que todos los subconjuntos tengan aproximadamente el mismo tamaño, sean disjuntos y cuya unión sea toda la muestra. Luego considerar uno de los subconjuntos, llamémoslo i y ajustar el modelo con todos los elementos de la muesta menos los que corresponden al grupo i. Posteriormente testear el modelo resultante en el subconjunto i. Así sobre cada uno de los K subconjuntos. Se obtienen entonces L_i, $i = 1, ..., K$ valores de la tasa de error. De allí es posible estimar la tasa de error de clasificación como el promedio de todos lo anteriores:

$$\hat{L}_{(K)} = \sum_{i=1}^{K} \frac{1}{K} L_i$$

A esta forma de validación cruzada se la conoce con el nombre de *"k-Fold Cross Validation"*. Es claro que en particular el caso $K = 2$ es la primera aproximación introducida al principio de esta sección y el caso $K = n$ es LOOCV.

Ahora bien, tratemos de entender cuál es la relación sesgo-varianza presente en los estimadores para los distintos valores de K. Cuanto menor sea el valor de K seleccionado, el estimador resultará menos sesgado pero tendrá más varianza. Esto se debe a que por un lado las muestras consideradas para los ajustes de los modelos serán más pequeñas pero a la vez estarán más correlacionadas. Esto último se explica simplemente por el hecho de que la muestras no diferirán mucho unas a otras. Entonces el caso $K = 2$ corresponde al estimador con menor sesgo pero con mayor varianza, mientras que el caso $K = n$ (LOOCV) equivale a considerar el estimador más sesgado pero con menor varianza. En genereal, se recomiendan usar $K = 5$ ó $K = 10$. De todas maneras, siempre que no sea muy costoso computacionalmente vale la pena aplicar la téncina LOOCV sobre los datos.

En este caso introdujimos las técnicas de validación cruzada con el objetivo de estimar la tasa de error de un cierto clasificador. Es posible, además utilizar

dichas técnicas para seleccionar el nivel de flixibilidad para algún modelo. Por ejemplo se podrían aplicar técnicas de validación cruzada para elegir la cantidad de vecinos a considerar en el método de k vecinos más cercanos, estimando las tasas de error de clasificación para ciertos valores de k (cantidad de vecinos a considerar en el modelo), para luego seleccionar aquel que tenga el valor más bajo.

Es importante aclarar que las técnicas de validación cruzada no solo pueden ser utilizadas en el contexto de métodos de clasificación, sino que también es posible aplicarlas para evaluar el comportamiento de cualquier modelo estadístico predictivo.

A modo de ejemplificar estas técnicas, a partir de los datos utilizados en el Ejemplo 3.1 de la sección 3.2.1, ajustamos un modelo discriminante lineal y estimamos la tasa de error de clasificación con técnicas de validación cruzada. Recordemos que los datos provenían de 506 barrios de la ciudad de Boston y los habiamos dividido en dos grupos según la variable `rm` (número de habitaciones promedio por vivienda). Nuevamente como variables explicativas utilizamos `lstat` (porcentaje de la población con estatus social en la categoría inferior) y `age` (porcentaje de viviendas construidas antes de 1940). En la Tabla 3.1 representamos los valores obtenidos de la tasa de error de clasificación para $K = 2, 5, 10, n$. Como vimos anteriormente, los casos con $K \neq n$ poseen una parte aleatoria en su procedimiento, lo que hace que al aplicar validación cruzada varias veces el valor del estiamador varie entre una y otra. Por ello, en estos casos repetimos el procedimiento 100 veces y lo que se indica en la Tabla es el promedio de la tasa de error obtenido a partir de estos (de todas maneras, el desvió del estimador en los tres casos fue bajo por lo que desde un punto de vista práctico podríamos restarle importancia: para $K = 2$,$sd = 0{,}006$; para $K = 5$ $sd = 0{,}0039$; para $K = 10$ $sd = 0{,}0064$).

(K=2)	**K=5**	**K=10**	**K=n(506)**
0,2807	0,2821	0,2125	0,2856

Tabla 3.1: Estimaciones de la tasa de error de clasificación para el modelo discriminante lineal sobre los datos del ejemplo 3.1 obtenidos a partir de técnicas de validación cruzada. Para los casos $K = 2, 5, 10$, el valor expresado indica el promedio obtenido a partir de 100 replicaciones de la técnica.

Se podría a partir de la estimación de la tasa de error de clasificación seleccionar algún parámetro del modelo. Por ejemplo el valor de la ventana h en la estimación no paramétrica de la densidad. Notemos que en la sección 3.2.1 expusimos técnicas de validación cruzada para la seleccion de éste parámetro, pero en ese caso basamos dicha selección en el valor del MISE. Como la meta final de la estimación no paramétrica de la densidad está, es este caso, en generar un buen clasificador, es razonable entonces la propuesta de seleccionar el parámetro h minimizando la tasa de error de clasificación. Con este objetivo y siguiendo con el Ejemplo 3.1 seleccionamos un conjunto de 20 valores equiespaciados en un intervalo de medida dos y centrados en los valores de h obtenidos en la sección anterior mediante el método de referencia a lo normal. Para cada uno de estos valores estimamos la tasa de clasificación por LOOCV y buscamos el mínimo. En la Tabla 3.2 exponemos los valores del error de clasificación y del parámetro de suavizado para ambas técnicas.

	Minimizando Tasa de error (por LOOOCV)	Regla de referencia a la normal
Tasa de error de Clasificación	0,23320	0,2411
$h = h(lstat, age)$ C_1	(2,2779; 10,2976)	(2,7306; 10,0330)
$h = h(lstat, age)$ C_2	(1,6418; 12,2599)	(2,2002; 11,8091)

Tabla 3.2: Valores de la Tasa de error de clasificación y del paramétro de suavizado para el problema del Ejemplo 3.1

3.4. Laboratorio R

A continuación mostraremos algunos comandos del `Software R` que nos permitirán utilizar las herramientas presentadas previamente. Para ello, utilizaremos los datos `Pima.tr` del paquete `MASS`. Estos datos corresponden a una población de 200 mujeres menores de 21 años provenientes de la comunidad Pima y que viven cerca de la capital del estado de Arizona (EEUU). Algunas de estas mujeres fueron diagnosticadas de diabetes según la Organización Mundial de la Salud (OMS). Para cada mujer analizada se midió el número de embarazos (`npre`), la concentración de glucosa (`glu`), la presión arterial (`bp`), el espesor cútaneo de los triceps (`skin`), el índice de masa corporal (`bmi`), la función diabética de pedigree (que mide de alguna forma la genética de la persona en cuanto a esta enfermedad) (`ped`) y la edad (`age`). Por último en los datos se encuentra el diganóstico de la OMS (`type`) para cada una de las mujeres.

```
> library('MASS')
> names(Pima.tr)
[1] "npreg" "glu" "bp"   "skin" "bmi"  "ped"  "age"  "type"
> summary(Pima.tr)
    npreg          glu             bp             skin
 Min.   : 0.00  Min.   : 56.0  Min.   : 38.00  Min.   : 7.00
 1st Qu.: 1.00  1st Qu.:100.0  1st Qu.: 64.00  1st Qu.:20.75
 Median : 2.00  Median :120.5  Median : 70.00  Median :29.00
 Mean   : 3.57  Mean   :124.0  Mean   : 71.26  Mean   :29.21
 3rd Qu.: 6.00  3rd Qu.:144.0  3rd Qu.: 78.00  3rd Qu.:36.00
```

```
 Max.   :14.00  Max.   :199.0  Max.   :110.00  Max.   :99.00
     bmi            ped             age        type
 Min.   :18.20  Min.   :0.0850  Min.   :21.00  No :132
 1st Qu.:27.57  1st Qu.:0.2535  1st Qu.:23.00  Yes: 68
 Median :32.80  Median :0.3725  Median :28.00
 Mean   :32.31  Mean   :0.4608  Mean   :32.11
 3rd Qu.:36.50  3rd Qu.:0.6160  3rd Qu.:39.25
 Max.   :47.90  Max.   :2.2880  Max.   :63.00
```

Para probar los modelos, divimos la muestra en dos. Una mitad la usaremos como muestra de entrenamiento, y la otra la usaremos para testear el modelo.

```
train.data = Pima.tr[1:100,]
test.data = Pima.tr[101:200,]
```

3.4.1. Modelo LDA

Para el modelo LDA, usaremos el comando `lda()` también del paquete `MASS`. Supongamos que queremos predecir `type` usando las variables `age`, `glu` y `bmi`. En `R` escribimos:

```
> ajuste.lda = lda(type~glu+age+bmi,data=train.data)
> ajuste.lda
Call:
lda(type ~ glu + age + bmi, data = train)

Prior probabilities of groups:
  No  Yes
0.67 0.33

Group means:
        glu     age     bmi
No  110.5075 27.92537 30.47761
Yes 145.0909 39.81818 34.52121

Coefficients of linear discriminants:
          LD1
glu 0.02481489
age 0.04884753
bmi 0.05845929

> plot(ajuste.lda)
```

Gruop means, nos da las estimaciones de las medias para cada variable. *Coefficients of linear discriminants* nos da una combinación lineal de las variables que es usada para la regla de decisión del modelo LDA (3.5). Es decir, si `glu`$\times 0.02481489+$`age`$\times 0.04884753+$`bmi`$\times 0.05845929$ es grande se clasifica la observación en "Yes", si es chica en "No".

Vemos que las probabilidades a priori de cada clase `R` las está estimando por $\pi_{No} = 0{,}67$ y $\pi_{Yes} = 0{,}33$. Es decir que en nuestra muestra el 67 % de los casos provienen de mujeres a las que no les fué diagnosticada diabetes y el 33 % restante a mujeres a las que si les fué diagnosticada la enfermedad. Es posible definir de antemano las probabilidades a priori. Supongamos que contamos con información adicional que nos permita suponer que las probabilidades a priori son por ejemplo: 0,5 para ambas clases. Para ingresar esta información en `R` utilizamos el parámetro `prior` y lo escribimos como:

```
>ajuste.lda = lda(type~glu+age+bmi,data=train.data,prior=c(0.5,0.5))
```

Supongamos ahora que queremos predecir la variable `type` sobre los datos test. Usamos la función `predict()`:

```
> pred.lda = predict(ajuste.lda,test.data)
> names(pred.lda)
[1] "class"   "posterior" "x"
```

Si ahora queremos ver como funciona el modelo:

```
> clase.lda=pred.lda$class
> table(clase.lda,test.data$type)

clase.lda No Yes
     No  47   9
     Yes 18  26
> mean(clase.lda==test.data$type)
[1] 0.73
```

Esto nos dice que el modelo predijo correctamente un 73 % de las veces. El comando `table()` nos permite ver en detalle los errores y aciertos.

3.4.2. Modelo QDA

El comando para el modelo QDA es `qda()`, y la forma de usarlo es análoga al anterior

```
> ajuste.qda=qda(type~glu+age+bmi,data=train.data)
> ajuste.qda
Call:
qda(type ~ glu + age + bmi, data = train.data)

Prior probabilities of groups:
  No  Yes
0.67 0.33

Group means:
        glu     age     bmi
No  110.5075 27.92537 30.47761
Yes 145.0909 39.81818 34.52121
```

Lo mismo para la función `predict()`

```
> pred.qda = predict(ajuste.qda,test.data)
> clase.qda=pred.qda$class
> table(test.data$type,clase.qda)
    clase.qda
     No Yes
  No  54  11
  Yes 15  20
> mean(clase.qda==test.data$type)
[1] 0.74
```

3.4.3. Estimador núcleo

Para la estimación no paramétrica de la densidad vía núcleo existen varios paquetes en `R` que nos permiten implementar de manera sencilla este método. Dependiendo de los datos y del objetivo final de la estimación convendrá usar uno ú otro. Nombramos algunos de ellos: `sm`, `KernSmooth`, `Locpol`, `locfit`. Estos

paquetes poseen herramientas que entre otras cosas permiten dibujar las estimaciones, tanto para una como para dos dimensiones, y seleccionar el parámetro de suavizado según las distintas técnicas existentes.

Un comando básico que se suele usar bastante es `density()` del paquete `stats` que computa un estimador núcleo de la densidad para datos univariados. Se puede seleccionar tanto el núcleo como el parámetro de suavizado. Por default usa el núcleo gaussiano y h dado por la regla de la referencia a la normal.

Supongamos que ahora queremos clasificar los datos mirando únicamente la variable `glu`.

```
> Glu.test=test.data$glu
> Glu.train=train.data$glu

> train.Yes=Glu.train[train.data$type=='Yes']
> train.No=Glu.train[train.data$type=='No']
```

`train.Yes` resulta entonces la variable `glu` del subconjunto de los datos de entrenamiento que son del tipo "Yes". De la misma forma `train.No` para el subconjunto del tipo "No". Pasemos ahora a estimar las densidades de estos dos conjuntos de datos.

```
> d.glu.y=density(train.Yes)
> d.glu.n=density(train.No)
> plot(d.glu.n)
> d.glu.n

Call:
       density.default(x = train.No)

Data: train.No (67 obs.);    Bandwidth 'bw' = 8.111

      x               y
 Min.   : 46.67  Min.   :8.267e-06
 1st Qu.: 89.33  1st Qu.:5.463e-04
 Median :132.00  Median :3.650e-03
 Mean   :132.00  Mean   :5.853e-03
 3rd Qu.:174.67  3rd Qu.:1.122e-02
 Max.   :217.33  Max.   :1.752e-02
```

Los valores anteriores corresponden a las coordenadas en x e y en las cuales fue calculada la función de densidad estimada.

Con la función `plot()` dibujamos la densidad estimada (Figura 3.9). Ahora supongamos que a partir de esta estimación, queremos estimar la clase de los datos test según lo visto en la sección 3.2.1. Vamos a usar la función `dkernel()` del paquete `klaR`. Esta última lo que hace es, a partir de la estimación de una cierta densidad dada por la función `density()`, calcular el valor de la densidad estimada en un determinado vector vector.

```
> library(klaR)
> pred_y = dkernel(Glu.test,kernel=d.glu.y)
> pred_n = dkernel(Glu.test,kernel=d.glu.n)
```

Si queremos dibujar la estimación del vector test:

```
> plot(Glu.test,pred_y)
```

Si ahora quisiéramos predecir la clase del vector test, simplemente lo que tendríamos que hacer es aplicar la regla de Bayes a partir de estas estimaciones. Recordemos que habiamos estimado las probabilidades a priori como $\pi_{No} = 0{,}67$ y $\pi_{Yes} = 0{,}33$. En `R`:

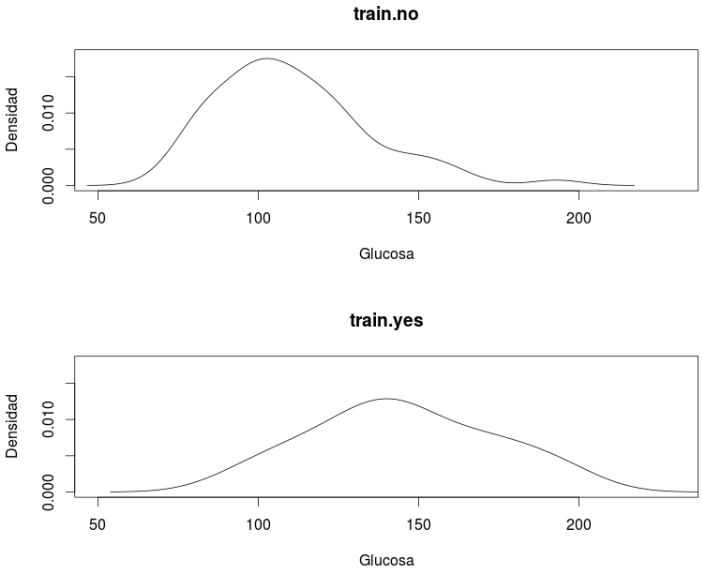

Figura 3.9: Estimación de la densidad de la variable glucosa para ambas clases. Estimación a partir de un núcleo gaussiano y selección del parámetro de suavizado por la regla de referencia a la normal.

```
> clase.pred=rep('No',length(Glu.test))
> clase.pred[0.33*pred_y>0.67*pred_n]='Yes'
> mean(clase.pred==test.data$type)
[1] 0.67
```

Veamos ahora una manera de implementar la técnica de Bayes Naive que vimos en la sección 3.2.1. Usamos la función `NaiveBayes()` del paquete `klaR`. Volvemos entonces al problema multivariado en que queremos predecir las clases de los datos test con las variables `glu`, `age` y `bmi`. Entonces en `R`:

```
> ajuste_bn = NaiveBayes(type~glu+age+bmi ,data = train.data ,usekernel=TRUE)
> pred.bn = predict(ajuste_bn,test.data,type='class')
> names(pred.bn)
[1] "class"    "posterior"
>
> mean(pred.bn$class==test.data$type)
[1] 0.71
```

Existe un comando similar dentro del paquete `e1071`, `naiveBayes()`. Este último no nos permite fijar la probabilidad a priori.

Capítulo 4

Modelos de Regresión

En este capítulo abordaremos el problema desde el segundo enfoque. Es decir, trataremos el problema de modelar $P(C = C_j|X)$, o lo que es lo mismo la esperanza condicional de la indicadora de $C = C_j$ dado X. Tal como vimos en la sección anterior, una vez hecho esto, la regla de clasificación consistirá sencillamente en asignar al nuevo individuo la clase que posea mayor probabilidad estimada.

4.1. Modelo Logit

Supongamos primero el caso $J = 2$, es decir sólo contamos con dos clases ($C = C_0 = 0$ ó $C = C_1 = 1$). Partiendo de un predictor X de dimensión p, consideramos $p(X) = P(C = 1|X)$. Una vez que tengamos un valor estimado de esta probabilidad, la regla propuesta de clasificación será:

$$\begin{cases} C = 1 \text{ si } \hat{p} \geq 0{,}5 \\ C = 0 \text{ sino} \end{cases}$$

Proponemos como primera medida considerar un modelo lineal:

$$\hat{p}(X) = \beta_0 + \beta_1' X$$

Este primer enfoque presenta un problema importante. Podemos estimar los parámetros del modelo lineal, y luego $\hat{p}$ representará la estimación de la probabilidad de que un individuo con características definidas por X pertenezca a la clase 1. No obstante, para que $\hat{p}$ resulte una probabilidad deberá tomar valores entre 0 y 1, lo cual podría no suceder para algunos valores de X. De todas formas es importante aclarar que ésto no es un inconveniente para clasificar una observación, pero si lo es a la hora de interpretar el resultado de la regla.

Para evitar este inconveniente, debemos transformar $p(X)$ usando alguna función que nos permita garantizar valores entre cero y uno:

$$p(X) = F(\beta_0 + \beta_1' X)$$

Proponemos entonces tomar como F la función de distribución logística. Obtenemos así:

$$p(X) = \frac{e^{\beta_0 + \beta_1' X}}{1 + e^{\beta_0 + \beta_1' X}} \tag{4.1}$$

Resulta entonces que:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1' X} \tag{4.2}$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1' X \tag{4.3}$$

A la expresión dada por (4.2) se la denomina *odd*. Entre p y odds existe una relación biunívoca y monótona: valores cercanos a cero de *odd* se corresponden con valores cercanos a cero de p. El lado izquierdo de la expresión (4.3) se suele denominar *logit(p(X))*.

Si nos reducimos al caso $p = 1$ (X de dimensión 1), esta variable *logit* representa en una escala logarítmica la diferencia de pertenecer a cada una de las clases, y al ser función lineal de X facilita su estimación e interpretación. Si bien la relación entre X y $p(X)$ no es lineal, se tiene que si β_1 es positivo, al aumentar X se aumentará $p(X)$, por el contrario si β_1 es negativo, incrementos en X supondrán disminuciones en $p(X)$.

4.1.1. Estimación de los parámetros

Para la estimación de lo parámetros β_0 y β_1 utilizaremos el método de *máxima verosimilitud*(MV). La idea esencial es la de estimar los parámetros β_0 y β_1 de forma tal que la probabilidad estimada de pertenecer a la verdadera clase de procedencia para cada elemento de la muestra sea, lo más posible, cercana a uno. Supondremos la muestra de entrenamiento de la forma: (x_i, c_i), $i = 1, ..., N$, en donde por c_i nos referimos a la verdadera clase de procedencia de cada elemento. Para x_i consideramos

$$p_i = P(C = 1|X = x_i) \tag{4.4}$$

Luego, dada cualquier respuesta c_i en C_i, la función de probabilidades para la misma será de la forma:

$$P(C_i = c_i) = p_i^{c_i}(1-p_i)^{1-c_i}$$

Entonces para la muestra:

$$P(C_1 = c_1, ..., C_N = c_N|X = x_1, ..., X = x_N) = \prod_{i=1}^{N} p_i^{c_i}(1-p_i)^{1-c_i}$$

Tomando logaritmo en la expresión anterior, obtenemos:

$$\log P(C|X) = \sum_{i=1}^{N} c_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{N} \log(1-p_i) \tag{4.5}$$

$$= \sum_{i=1}^{N} c_i \log(p_i) + (1-c_i)\log(1-p_i)$$

A partir de (4.4) y (4.1) expresamos los p_i en función de los parámetros de interés $\beta = (\beta_0, \beta_1)$. De esta forma obtenemos como función de verosimilitud:

$$L(\beta) = \sum_{i=1}^{N} c_i x_i' \beta - \sum_{i=1}^{N} \log(1 + e^{x_i'\beta}) \tag{4.6}$$

Buscamos entonces máximizar la expresión de arriba. Derivando (4.6) e igualando a cero obtenemos:

$$\sum_{i=1}^{N} c_i x_i = \sum_{i=1}^{N} x_i \left(\frac{1}{1 + e^{-x_i' \hat{\beta}}} \right) = \sum_{i=1}^{N} \hat{c}_i x_i$$

Esta ecuación nos dice por un lado que el producto de los valores observados por las variables explicativas debe ser el mismo que el de los valores previstos. Por otro lado, notemos que los residuos $(c_i - \hat{c}_i)$ deben ser ortogonales a la variable x. Obtener el valor $\hat{\beta}_{MV}$ que maximiza la verosimilitud no resulta ser una tarea trivial. De todas formas, métodos iterativos para aproximar el estimador han sido implentados en `R`, utilizando funciones que detallaremos en la sección 4.3.

Entonces dado x, un punto a clasificar, hemos estimado el valor de $p(C = 1|X = x)$, y lo hemos hecho a partir de un modelo paramétrico.

Ahora bien, puede suceder que este modelo no se adapte correctamente a todo el rango de valores de las variables explicativas. Para estos casos es posible ajustar el método de forma que resulte más apto para los datos con los que se trabaja. Si bien puede ocurrir que el modelo no sea bueno en forma global, sí lo puede ser de forma local, es decir en un entorno del punto a clasificar. El propósito es hacer que las observaciones más cercanas a x sean las que aporten más peso al modelo. Consideramos entonces para cada observación de la muestra de entrenamiento un peso $d_i = d(x_i, x)$, el cual sea decreciente en distancia a x. Siguiendo con lo expuesto en (4.5) tenemos que, en este caso, la contribución de cada observación a la función de verosimilitud será:

$$d_i \left(c_i \log \left(\frac{p_i}{1 - p_i} \right) + \log(1 - p_i) \right)$$

Entonces, sumando obtenemos:

$$\begin{aligned} L_x(\beta) &= \sum_{i=1}^{N} d_i \left(c_i \log \left(\frac{p_i}{1 - p_i} \right) + \log(1 - p_i) \right) \\ &= \sum_{i=1}^{N} d_i c_i x_i' \beta - \sum_{i=1}^{N} d_i \log(1 + e^{x_i' \beta}) \end{aligned}$$

Es claro que $L_x(\beta)$ es función del punto x, ya que es éste el que determina cuales son las observaciones a tener en cuenta y en que grado.

Nuevamente maximizando esta función obtenemos el valor $\hat{\beta}$ que permite obtener la estimación de $p(x)$.

En la práctica los pesos d_i suelen definirse a partir de funciones núcleo de dimensión p, del mismo modo en que lo haciamos para la estimación no paramétrica multivariante de la densidad en el capítulo anterior.

Ejemplo 4.1. *Vamos a trabajar con los datos de radar que presentamos en el capítulo 2. Partimos entonces de nuestra muestra de entrenamiento, en este caso analizaremos los datos de forma univariada. Es importante aclarar que si bien en el análisis final utilizaremos tres covariables para la clasificación, en esta instancia emplearemos dos de ellas de forma de intentar entender el comportamiento de cada una por separado a la hora de discriminar. Consideraremos*

por un lado el módulo de la velocidad `|V|` *y por el otro la variable* `Rho`*. Dado que la muestra con la que contábamos era de un tamaño muy grande, decidimos trabajar con un subconjunto de 1000000 elementos de cada clase.*

Tomamos una ventana fija h que nos dará el peso de cada uno de los elementos de la muestra en la estimación. Estimamos así la probabilidad de que un nuevo pixel pertenezca o no a la clase Eco meteorológico. En la Figura 4.1 graficamos los resultados: Del lado izquierdo tenemos la probabilidad de pertenecer a la clase No meteorológica dado el módulo de la velocidad. Vemos como a valores más cercanos a cero esta probabilidad es más alta, lo cual tiene sentido si pensamos que tanto los ecos de terreno (montañas, edificios) como el fenómeno de la Propagación anómala tienen velocidad cero (esto no ocurre para los pájaros o el polvo, por ejemplo). Del lado derecho graficamos la probabilidad de pertenecer a la clase meteorológica dado el valor del coeficiente copolar `Rho`*. Es claro aquí que para valores mas cercanos a uno, la probabilidad de estar en la clase meteorológica es mayor, lo cual también es coherente con el comportamiento de esta variable (un área de lluvia tiende a mostrar homogeneidad en relación a la distribución de tamaño y forma de las gotas ,y esto produce valores más cercanos a uno)*

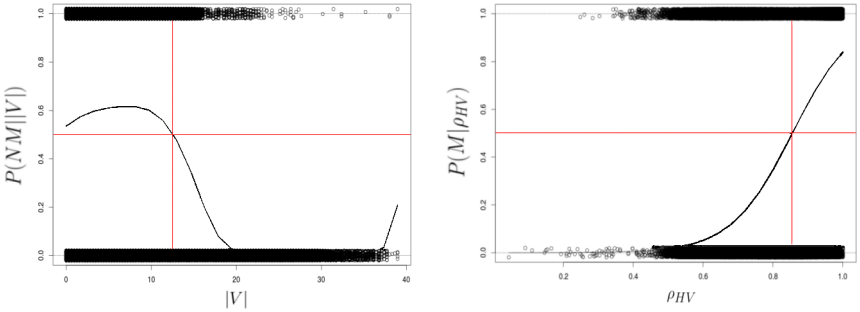

Figura 4.1: Izquierda: en negro la $P(NM||V|)$ estimada, en rojo el punto donde la probabilidad estimada es de 0,5 . Derecha: en negro la $P(M|\rho_{HV})$ estimada, en rojo el punto donde la probabilidad estimada es de 0,5.

Si bien ambas variables presentan una clara distinción entre las clases, el hecho de utilizar una sola de ellas puede ser problemático. Como ya dijimos, la velocidad detecta bien la presencia de propagación anómala pero no así los pájaros o el polvo. El coeficiente copolar, si bien es una variable con alto nivel discriminante, suele poseer valores más bajos para el caso de nieve o granizo, lo cual puede ser confuso. Es por eso que el simple análisis de las variables por separado puede llevar a clasificaciones erróneas.

4.1.2. Relación con modelo LDA

Existe una relación estrecha entre el modelo logístico y el modelo LDA presentado en el capítulo anterior. Partamos de la suposición de que las funciones de distribución de cada una de las clases son normales con distintas medias pe-

ro con la misma matriz de covarianzas. Veamos que bajo estos supuestos, $C|X$ verifica el modelo logístico. En efecto, a partir de la expresión (3.2) obtenemos:

$$\log\left(\frac{P(C=C_1|X)}{P(C=C_0|X)}\right) = \log\left(\frac{\pi_1 f_1}{\pi_0 f_0}\right) = \log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{f_1}{f_0}\right)$$

Como suponemos que las funciones f_0 y f_1 son densidades normales con las misma matriz de covarianzas (3.4), tenemos que la expresión anterior equivale a:

$$\log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_0+\mu_1)'\Sigma^{-1}(\mu_1-\mu_0) + x'\Sigma^{-1}(\mu_1-\mu_0) = \alpha_0 + \alpha_1 x$$

la cual resulta ser una función lineal en x.

Por otro lado, bajo esta misma suposición, para el caso del modelo logístico teniendo en cuenta la expresión (4.3) se cumple que:

$$\log\left(\frac{P(C=C_1|X)}{P(C=C_0|X)}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1' x$$

Observamos entonces que de las dos formas llegamos a modelos lineales en x. Ahora bien, la diferencia entre ambos métodos radica en como se estiman los parámetros desconocidos.

Analicemos un poco más. Para ello consideremos la densidad conjunta de una observación:

$$f(x,c) = f(x|c)f(c) = f(c|x)f(x)$$

Tal como vimos en la sección 3.1.1 en el modelo LDA estimamos $\hat{\Sigma}$, $\hat{\mu}_0$ y $\hat{\mu}_1$, es decir la distribución completa de X. En cambio en el modelo logístico sólo estimamos $p+1$ parámetros correspondientes a la distribución de C condicionada a X.

Entonces, el esquema para el caso LDA será:

$$\prod_i f(x_i,c_i) = \underbrace{\prod_i f(c_i)}_{\text{Bernoulli}}$$

Mientras que para el caso Logit:

$$\prod_i f(x_i,c_i) = \underbrace{\prod_i f(c_i|x_i)}_{\text{Logístico}} \underbrace{\prod_i f(x_i)}_{\text{Ignorado}}$$

Como para la clasificación sólo se requiere saber $f(C|X)$ en realidad no es necesario estimar la distribución conjunta. El modelo logit deja la distribución marginal $f(x)$ sin especificar, por lo que resulta ser menos paramétrica que el modelo LDA, lo cual puede pensarse como una ventaja. De todas maneras fue demostrado (Efron 1975) que cuando los datos son efectivamente normales con la misma matriz de covarianzas, el modelo LDA funciona mejor que la regresión logística. Es importante aclarar que el modelo LDA es más sencillo desde el punto de vista computacional. Con esto último nos referimos a lo que respecta a la estimación de los parámetros en el modelo: en LDA es posible obtener expresiones analíticas de los mismos (3.8 y 3.10), en cambio en el modelo Logit es necesario obtenerlos a partir de métodos numéricos.

4.1.3. Modelo Multilogic

Podemos ahora generalizar fácilmente éste método para el caso en que contamos con más de dos poblaciones ($J > 2$). Si bien el modelo multilogic no es de los más usados cuando se trata de clasificar en mas de dos poblaciones, creemos que vale la pena mencionarlo. Entonces, consideramos p_{ij} la probabilidad de que la observación i pertenezca a la clase j:

$$p_{ij} = \begin{cases} \frac{e^{\beta_{0j}+\beta'_{1j}x_i}}{1+\sum_{k=1}^{J-1} e^{-\beta_{0k}-\beta'_{1k}x_i}} & \text{si } j = 1, ..., J-1 \\ \frac{1}{1+\sum_{k=1}^{J-1} e^{-\beta_{0k}-\beta'_{1k}x_i}} & \text{si } j = J \end{cases} \tag{4.7}$$

Es claro que para cada observación se satisface $\sum_{j=1}^{J} p_{ij} = 1$. Luego para comparar las probabilidades entre dos clases, realizando el cociente:

$$\frac{p_{ij_1}}{p_{ij_2}} = \frac{e^{-\beta_{0j_1}-\beta'_{1j_1}x_i}}{e^{-\beta_{0j_2}-\beta'_{1j_2}x_i}} = e^{(\beta_{0j_1}-\beta_{0j_2})(\beta'_{1j_1}-\beta'_{1j_2})x_i}$$

Observemos que la ecuación anterior indica que las probabilidades relativas entre dos clases no dependen de las demás.

4.2. Estimación no paramétrica de regresión: K vecinos próximos (KNN)

El método que presentaremos a continuación, es realmente sencillo e intuitivo pero a la vez muy eficaz en ciertas situaciones. El procedimiento es el siguiente: primero se define una medida de distancia (generalmente la distancia de Mahalanobis o la distancia euclidea), y un número natural k. Luego dada una observación test x, se calcula la distancia de la misma hacia todos los elementos de la muestra de entrenamiento. Como próximo paso, se seleccionan los k individuos de la muestra más próximos al punto a clasificar, los cuales notaremos como $\Lambda_k(x)$. De esta forma, será posible estimar la probabilidad condicional para la clase C_j como la proporción de estos k puntos seleccionados que pertenecen a dicha clase. Es decir:

$$E(I_{\{C=C_j\}}|X) = P(C = C_j|X = x)$$

La cual es posible estimar como

$$\frac{1}{k} \sum_{i \in \Lambda_k(x)} I(c_i = C_j) \tag{4.8}$$

Por último se aplica la Regla de Bayes y se clasifica la nueva observación x en la clase que posea mayor probabilidad estimada.

Notemos que la expresión 4.8 coincide con el estimador no paramétrico de regresión usando núcleo uniforme y paramétro de suavizado $h = h(x, k)$ dado por los k vecinos más cercanos.

Como es de suponer, la elección del paramétro k implicará efectos drásticos en el resultado final de la clasificacion. Valores muy chicos de k llevarán a un estimador con bajo sesgo pero con alta varianza, mientras que si consideramos

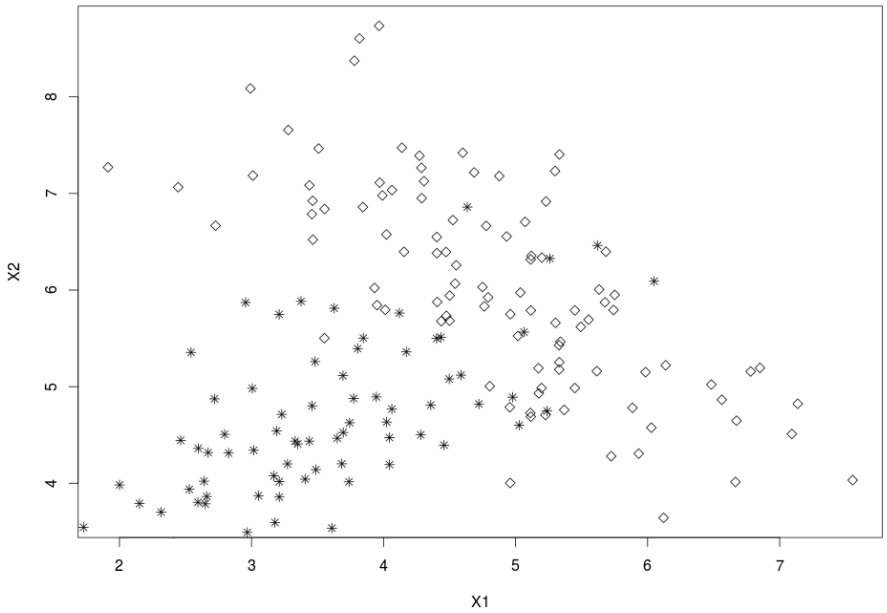

Figura 4.2: Distribución de los datos. Con rombos una de las clases, con asteríscos la otra

tomar valores cada vez mas grandes de k, el clasificador será más estable pero irá disminuyendo en flexibilidad (compromiso sesgo-varianza).

Ejemplo 4.2. *En este caso partimos de datos simulados. Contamos con una muestra de entrenamiento de tamaño* 300, 150 *elementos de cada clase. Por otro lado usamos una muestra test de* 200 *puntos, en los que la mitad de ellos son de la clase* C_1 *y la mitad de la clase* C_2*. En la Figura 4.2 mostramos la distribución de los datos test. En la Figura 4.3 graficamos los resultados. Con azul marcamos las observaciones que fueron clasificadas correctamente por el modelo, y en rojo las que no lo fueron.*

En la tabla 4.1 presentamos el porcentaje total de elementos bien clasificados para cada caso. Vemos como para valores extremos de k *el modelo da peores resultados.*

k	Bien clasificado
1	0.81 %
10	0.86 %
80	0.90 %
300	0.50 %

Tabla 4.1: Resultados

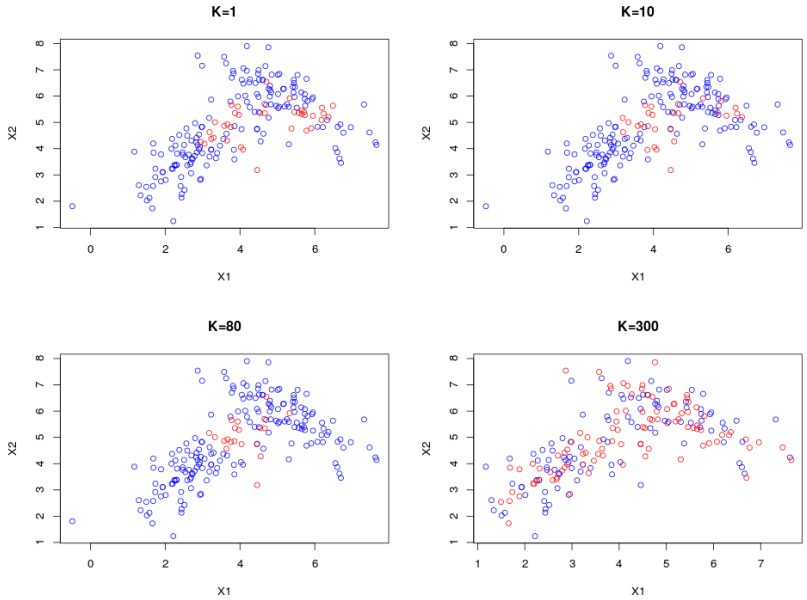

Figura 4.3: Resultados de la clasificación para cada uno de los valores de k. En azul: clasificados correctamente. En rojo: clasificados erroneamente

Si tomamos $k = 1$ el método se reduce simplemente a asignar x a la clase a la que pertence el elemento más próximo. Una práctica habitual es tomar $k = \sqrt{n_g}$ donde n_g es el tamaño promedio de las clases.

Algo que también puede ser de utilidad a la hora de seleccionar éste parámetro, es utlizar técnicas de *validación cruzada* para estimar el error de clasificación obtenido con distintos valores de k. Veamos brevemente como funciona ésta técnica en este caso:

- Se toma un valor cualquiera k.
- Se selecciona un elemento de la muestra de entrenamiento. Se lo saca de la muestra y se lo clasifica con el método KNN.
- Se repite el punto anterior sobre cada punto de la muestra.
- Se calcula el error de clasificación como la proporción de las clasificaciones erróneas sobre el total de la muestra

Se repite lo anterior para distintos valores de k, y luego se selecciona aquel que posea el error de clasificación mas bajo.

En los casos en los que se posee una muestra de entrenamiento grande, éste modelo no es aplicable ya que exige un gasto computacional extremo. Esto se debe simplemente al hecho de que para cada nueva asignación que se desee hacer es necesario comparar al nuevo individuo con cada uno de los elementos de la muestra, lo que puede tornase inviable si ésta es grande. De todas formas, cuando la muestra es chica y poco se sabe sobre la distribución de los datos KNN produce clasificadores muy buenos.

Un aspecto importante a tener en cuenta al momento de trabajar con este método de clasificación es, que como presisamente el mismo se basa en analizar las distancias de los puntos de la muestra a la observación a clasificar, resulta crucial tener en cuenta la escala en la que se encuentran las variables. Por ejemplo, supongamos que estamos ajustando algún modelo de clasificación con variables predictoras dadas por $X_1 = Salario$ (en pesos) y $X_2 = Edad$ (en años), para el método KNN la diferencia de 500$ es muy grande en comparación con 5 años. Esta diferencia en la magnitud de las variables hace que X_1 sea la que defina la clasificación del modelo, dajando X_2 casi sin efecto. Una forma de solucionar este problema puede ser estandarizar los datos, de forma que todas las variables tengan media cero y desvío uno. De esta manera las variables deberían poder ser comparables ya que todas estarían expresdas en magnitudes similares.

4.3. Laboratorio R

Para este segundo laboratorio seguiremos usando los datos `Pima.tr` del paquete `MASS` utilizados en la sección anterior. Siguiendo con el propósito de distinguir aquellas mujeres que fueron consideradas enfermas de diabetes por la OMS de aquellas que no, y utilizando las variables `age`, `glu` y `bmi`, nuevamente dividimos la muestra en dos: una parte para ser considerada como muestra de entrenamiento, y la otra para testear.

```
> train.data=Pima.tr[1:100,]
> test.data=Pima.tr[101:200,]
```

4.3.1. Modelo Logit

Para ajustar el modelo de regresión logística usaremos la función `glm()` del paquete `stats`. La sintaxis es muy similar a la de las funciones `lda()` y `qda()`

```
> ajuste_logit<-glm(type~glu+age+bmi,data=train.data,family=binomial)
```

En este caso es necesario agregar el argumento `family=binomial` para indicarle a `R` que estamos en un caso de regresión logística y no en otra generalización del caso lineal. Es posible también pedir un resumen del ajuste:

```
> summary(ajuste_logit)

Call:
glm(formula = type ~ glu + age + bmi, family = binomial, data = train.data)

Deviance Residuals:
    Min      1Q   Median      3Q     Max
-1.9601 -0.6284 -0.3369  0.4903  2.5007

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.63665  2.25323 -4.721 2.35e-06 ***
glu          0.03420   0.01056  3.240  0.00119 **
age          0.06670   0.02485  2.684  0.00727 **
bmi          0.10310   0.04946  2.085  0.03709 *
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1        1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 126.836 on 99 degrees of freedom
Residual deviance: 83.085 on 96 degrees of freedom
AIC: 91.085

Number of Fisher Scoring iterations: 5
```

Busquemos ahora predecir la clase de pertenencia de la muestra test a partir del modelo anterior. Para ello, utilizaremos la función `predict()`. Antes de proseguir, veamos a cual de las dos clases ("`Yes`" y "`No`") `R` considera "Clase 1" y a cual considera "Clase 0". Esto lo hacemos con el comando `constrasts()`

```
> contrasts(train.data$type)
    Yes
No    0
Yes   1
```

Entonces la Clase 1 será la de las mujeres que si fueron diagnosticadas de diabetes por la OMS, y la Clase 0 las que no. Entonces para obtener la $P(C = 1|X)$ para cada elemento de la muestra test, es decir para obtener la probabilidad estimada de que la mujer sea diagnosticada de diabetes dado los valores de `age`, `bmi` y `glu`, escribimos:

```
> prob=predict(ajuste_logit,test.data,type="response")
```

Al indicar `type="response"` lo que hacemos es precisamente pedir esta probabilidad condicional. Si quisieramos enumerar las probabilidades estimadas de las primeras 8 mujeres de la muestra test

```
> prob[1:8]
       101        102        103        104        105        106        107        108
0.12614514 0.17196343 0.03697961 0.07086664 0.38853255 0.60863736 0.15274689
     0.23698058
```

Entonces como ya tenemos las probabilidades estimadas, lo único que falta es dividir la muestra en aquellas con probabilidad mayor a 0.5 (serán clasificadas como "Yes"), y en aquellas con probabilidad estimadas menor a 0.5 (clasificadas como "No"). Creamos un vector de longitud el tamaño de la muestra test y allí guardamos la clasificación.

```
> pred <- rep("No",length(test.data[,1]))
> pred[prob>0.5] = "Yes"
```

Veamos que tan bien resulta esta clasificación

```
> table(test.data$type,pred)
    pred
     No Yes
  No  53  12
  Yes 16  19
> mean(test.data$type==pred)
[1] 0.72
```

Supongamos que quisieramos predecir diagnóstico de diabetes positivo para los casos en donde la probabilidad condicional estimada sea mayor a 0,4. Entonces:

```
> pred[prob>0.4]="Yes"
> mean(test.data$type==pred)
[1] 0.73
```

En este caso mejora un poco.

4.3.2. K vecinos próximos: KNN

En el caso del modelo knn, la implementación en `R` es un poco distinta de las anteriores. En vez de generar el modelo y luego hacer la predicción, ahora con el comando `knn()` se logra hacer la predicción directamente. La función `knn()` pertene a la librería `class`.

```
> library("class")
> ?knn
```

En `R` si escribimos `?algún comando`, se abre la ayuda del mismo. Entonces si leemos la ayuda de la función `knn()` resulta que para poder implementar este modelo deberíamos tener los datos organizados de una manera un poco diferente. Se tiene que `knn()` toma como parámetros de entrada:

- Una matriz con los valores de las variables predictoras de la muestra de entrenamiento
- Una matriz con los valores de las variables predictoras de los datos a clasificar
- Un vector con la clase de cada elemento de la muestra entrenamiento
- El parámetro k

Luego usando la función `cbind()` armamos las matrices con los valores de `age`, `bmi` y `glu` que necesitamos. Recordemos que para poder utilizar de manera correcta este método, será necesario estandarizar las variables previo a la clasificación. Para ello, utilizamos en comando `scale()`.

```
> Pima.stand = scale(cbind(Pima.tr$glu,Pima.tr$age,Pima.tr$bmi))
> train.X=Pima.stand[1:100,]
> test.X=Pima.stand[101:200,]
```

Guardamos la clase correcta de los elementos de la muestra de entrenamiento en una variable

```
> train.type=Pima.tr[1:100,]$type
```

Ahora si usamos la función `knn()`, pero previo a eso seteamos una semilla de forma que sea fácil reproducir estos resultados para el lector. Esto lo hacemos ya que hay casos en que hay empates en ciertas observaciones en la cantidad de vecinos entre las clases y `R` desempata de forma aleatoria.

```
> set.seed(111)
> pred.knn_1 = knn(train.X,test.X,train.type,k=1)
> mean(pred.knn_1==train.type)
[1] 0.57
```

Sólo en el 57 % de las veces la predicción resulta correcta. Veamos si cambiando el parámetro k se mejora esta proporción. Hacemos un ciclo para iterar en los valores de k entre 2 y 10, y guardamos la proporción de bien clasificados en el vector `correctos`.

```
> correctos=rep(NA,10)
> correctos[1]<-mean(pred.knn_1==train.type)
>
> for (k in 2:10)
+ {
+   pred.knn = knn(train.X,test.X,train.type,k=k)
```

```
+   correctos[k]=mean(pred.knn==train.type)
+
+ }
>  correctos
 [1] 0.57 0.51 0.53 0.56 0.53 0.51 0.50 0.49 0.51 0.50
```

Los resultados no mejoraron aumentado el parámetro k.

Como bien dijimos en el capítulo 3, las técnicas de validación cruzada son útiles a la hora de estimar la tasa de error de clasificación. En Tabla 4.2 exponemos los resultados de aplicar algunas de estas técnicas en los datos `Pima.tr`, para algunos de los métodos de clasificación trabajados. En todos los casos consideramos las variables predictoras `glu`, `age` y `bmi`.

Metodo CV	Lda	Qda	Logit	BN	KNN
LOOCV	0.250	0.245	0.240	0.240	0.275
K-fold, k=5	0.235	0.245	0.255	0.220	0.260

Tabla 4.2: Estimación de la tasa de error de clasificación para los distintos modelos. Primera fila: estimación a partir de la técnica LOOCV. Segunda fila: estimación a partir de la técnica k-fold Cross Validation con k=5. En el método KNN consideramos k=1

Si se tuviera que seleccionar uno de los modelos mirando la tabla anterior, sería fácil elegir ya que el modelo BN es el que para ambas técnicas posee la tasa de error de clasificación más baja.

Capítulo 5

Resultados en datos meteorológicos y conclusiones

En las secciones anteriores estudiamos varios métodos de clasificación. En este capítulo presentamos los resultados obtenidos al aplicar algunos de ellos para el caso del problema de clasificación en los datos de radar.

Implementamos cuatro métodos distintos: el modelo discriminante lineal (LDA), el modelo discriminante cuadrático (QDA), el modelo de Bayes Naive (BN), y el modelo de regresión logística (Logit). Tal como fué descripto en el Capítulo 2, luego de la clasificación inicial de cada uno de los modelos, realizamos un post-procesamiento de forma de incluir la estructura espacial de los datos. Recordemos que el post-procesamiento consitió en reasignar algunos puntos teniendo en cuenta la clase de sus vecinos. Más especificamente, sobre cada pixel miramos una ventana de 3×3, si la mayoría de sus vecinos (la mitad $+1$) fueron asignados a su misma clase, entonces no se hace ningún cambio, sino, se cambia la clase previamente asignada.

Todos los clasificadores se testean en un mismo conjunto de 12 imágenes, provenientes de 4 escenarios correspondientes a 4 eventos distintos (Tabla 5.1). Contamos con tres imágenes de cada uno de estos escenarios, resultando así el total de 12 imágenes. Como la clasificación utiliza a posteriori la estructura espacial de los puntos (postproceso), el testeo debió ser sobre imágenes completas. Al momento de la selección de estas imágenes procuramos que las mismas resulten representativas de los escenarios tipicamente encontrados en las imagenes existentes.

Para entrenar los modelos, seleccionamos 18 imagenes buscando variabilidad entre las mismas. El número total de pixeles con información a lo largo de las 18 imágenes es 1187967 (la cantidad de pixeles con información depende de la imagen). De ellos, el 0,477 % corresponde a señal meteorológica y un 0,5282 % a señal no meteorológica.

En las Figuras 5.1 y 5.2 mostramos a modo de ejemplo el resultado de los cuatro modelos para casos del 22 de Noviembre de 2009 y el 09 de Febrero de 2010. Si bien es posible observar que en ambos casos se dan situaciones en donde las dos clases (M y NM) se encuentran presentes, es claro que en el

Fecha	Ventana horaria	Situación
27 de Febrero de 2010	Desde 00:00 hasta 02:00 UTC[a]	Presencia de ecos no meteorológicos alrededor del sitio del radar aislados de ecos meteorológicos
20 de Junio 2009	Desde 04:00 hasta 06:00 UTC	Presencia de ecos meteorológicos extendidos espacialmente. Sobre el sitio del radar se observa eco no meteorológico.
09 de Febrero 2010	Desde 21:00 hasta 23:00 UTC	Mezcla de eco metorológico y no meteorológico. Eco meteorológico intenso
22 de Noviembre 2009	Desde 08:30 hasta 09:30 UTC	Presencia de eco meteorológico moderado. Alrededor del radar: eco no meteorológico

[a]UTC: Tiempo universal coordinado. Es el principal estándar de tiempo por el cual se regula el horario mundial

Tabla 5.1: Casos testeados

caso del 2009, los ecos están más mezclados que para el caso de la situación del 2010. Podemos observar como en general los resultados se corresponden bastante entre los modelos. De todas formas, es posible notar que las diferencias se encuentran en las zonas de frontera entre ambas clases. Esto último es de esperar ya que, tal como lo notan varios autores en la literatura, es el mayor desafío en lo que respecta al problema de clasificación de ecos en imágenes de radar. Es importante aclarar que los métodos de clasificación supervisada que trabajamos en esta tesis, analizan al *pixel* de la imagen como un individuo aislado sin tener en cuenta la estructura espacial de los mismos. Es por esto último que agregamos el post-procesamiento al modelo de forma de contemplar este punto.

Para poder cuantificar la performace de los diferentes modelos, utilizamos como índice de bondad al valor de la proporción de *pixeles* bien clasificados sobre el total de cada elevación. En la literatura se lo suele llamar *índice de éxito critico (CSI)*. Para el análisis que sigue supondremos que la clasificación del experto es perfecta.

Tal como indicamos previamente, para obtener las muestras para entrenar y testear los modeles, fue necesaria una clasificación manual de las imágenes, lo que hizo muy dificultoso poder contar con un volumen de datos considerable para trabajar. Frente a esta situación fue necesario plantear una manera de darle variabilidad a los datos de forma de poder hacer un análisis más profundo de los resultados. En esta instancia es importante notar lo siguiente: según la metodología utilizada, al momento de emplear la muestra de entrenamiento como componente de los modelos de clasificación, no se contempló la estructura espacial de los datos, consideración que si será necesaria al momento de evaluar los resultados sobre los datos clasificados. Es decir, al momento de entrenar

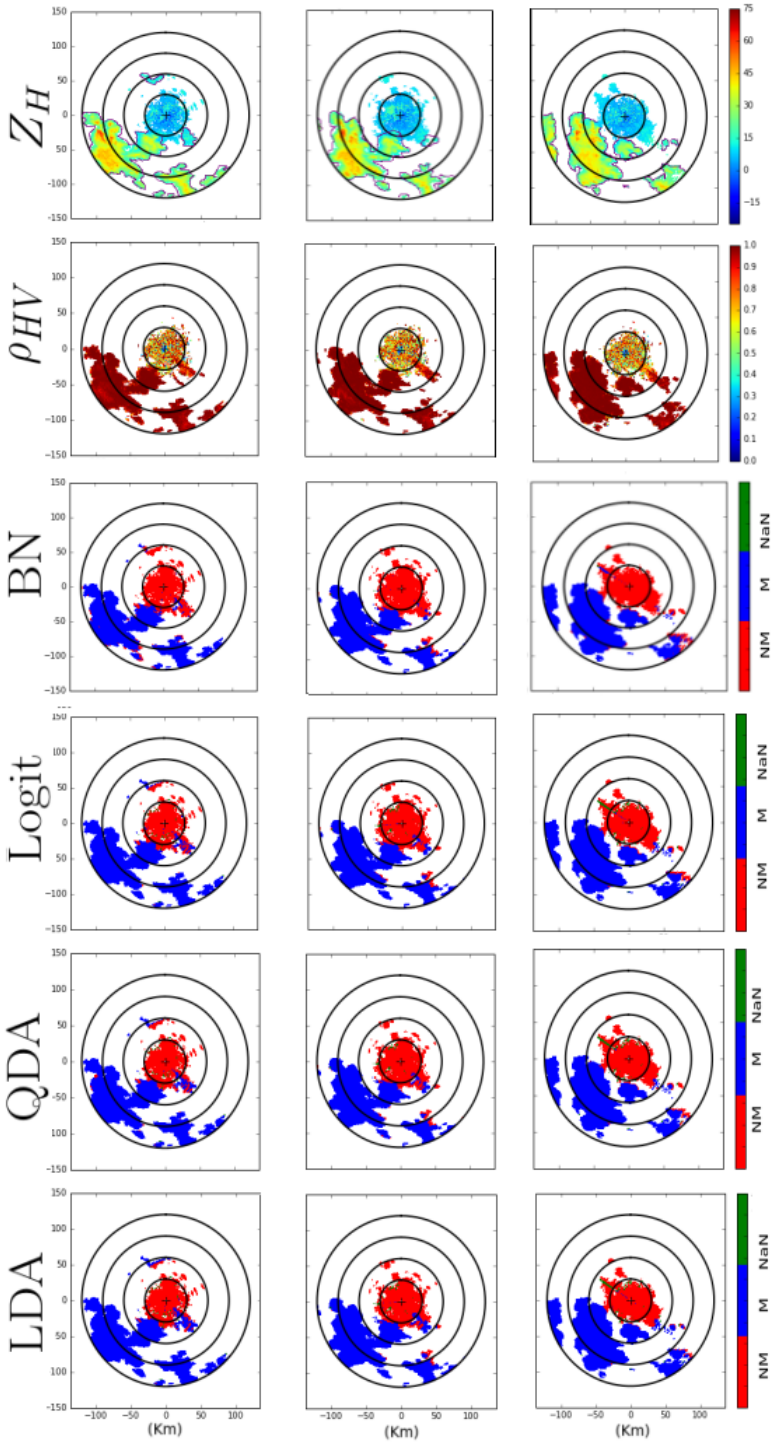

Figura 5.1: Resultados obtenidos para el 22 de Noviembre de 2009. Por fila y sobre la primera elevación: campo de Reflectividad, valor del coeficiente de correlación ρ_{HV}, modelo BN, modelo Logit, modelo QDA, modelo LDA. En azul los puntos clasificados como meteorológicos (M), en rojo los clasificados como no meteorológicos (NM), y en verde los considerados sin clase (NaN). En el campo de reflectividad se indica con contorno negro las áreas consideradas como eco meteorológico por el experto . Los anillos concéntricos poseen una separación de 30km entre sí.

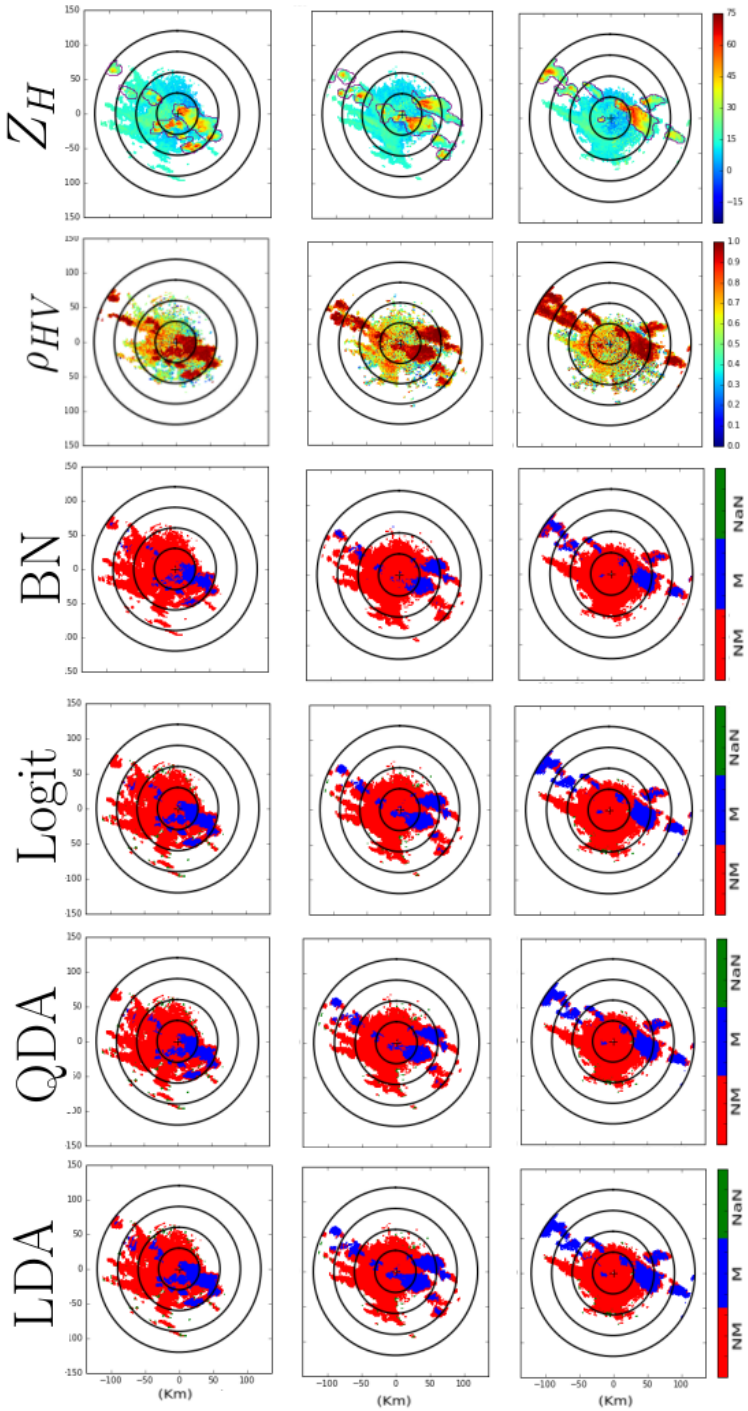

Figura 5.2: Resultados obtenidos para el 09 de Febrero de 2010. Por fila y sobre la primera elevación: campo de Reflectividad, valor del coeficiente de correlación ρ_{HV}, modelo BN, modelo Logit, modelo QDA, modelo LDA. En azul los puntos clasificados como meteorológicos (M), en rojo los clasificados como no meteorológicos (NM), y en verde los considerados sin clase (NaN). En el campo de reflectividad se indica con contorno negro las áreas consideradas como eco meteorológico por el experto . Los anillos concéntricos poseen una separación de 30km entre sí.

el modelo, consideramos los puntos de manera aislada, pero, al momento de la clasifacación la vecindad de cada punto es importante. Además es claro que para distintas configuraciones en las imágenes el nivel de complejidad puede diferir mucho. En efecto, el nivel de dificultad a la hora de clasificar situaciones en donde los ecos de tormenta estén alejados de ecos no meteorológicos no será el mismo que en situaciones en donde por ejemplo el fenómeno de la propagación anómala esté inmerso en eco de lluvia.

Por esta razón, luego de evaluar varias posibilidades decidimos aplicar téncinas de *simulaciones boostrap* remuestreando sobre los pixeles de las imágenes de entreamiento y aplicando los modelos sobre los datos test considerados en la Tabla 5.1.

5.1. Simulación 1

Llamemos N a la cantidad de pixeles totales de la muestra de entrenamiento. Entonces, la configuración a aplicar será la siguiente:

1. Se seleccionan de manera aleatoria y con reposición N puntos de la muestra de entrenamiento.

2. Tomando como muestra para entrenar la resultante del ítem anterior, se ajustan los cuatro modelos (BN, LDA, QDA, Logit)

3. Se clasifican las 12 imágenes de los casos test

4. Se aplica el post-procesamiento

5. Se calculan los valores de CSI para cada caso test y para cada modelo

6. Se vuelve al punto 1

Computamos la secuencia anterior 100 veces. En la Tabla 5.2 presentamos los resultados obtenidos para la media, mediana y desvío del índice CSI en cada uno de los modelos testeados.

Modelo	Media	Mediana	Desvío
LDA	0.9227	0.9428	0.0577
QDA	0.91771	0.9490	0.07378
Logit	0.9134	0.9494	0.0772
BN	0.9040	0.9470	0.0977

Tabla 5.2: Resultados obtenidos a partir de la simulación boostrap. Los valores de la tabla son referidos al índice de exito critico (CSI)

En las Figuras 5.4 y 5.3 graficamos los histogramas y los boxplots correspondientes a los resultados de los valores de los CSI en cada uno de los cuatro modelos correspondientes a la simulación boostrap.

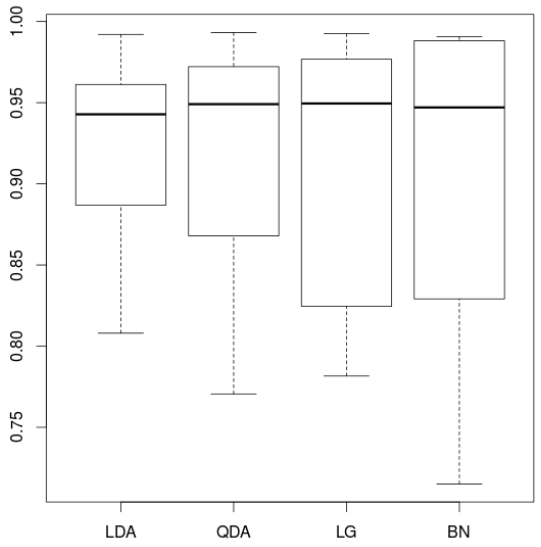

Figura 5.3: Boxplot de los valores de CSI obtenidos a partir de la simulación boostrap en los cuatro modelos. De izquierda a derecha: LDA, QDA, Logit, BN.

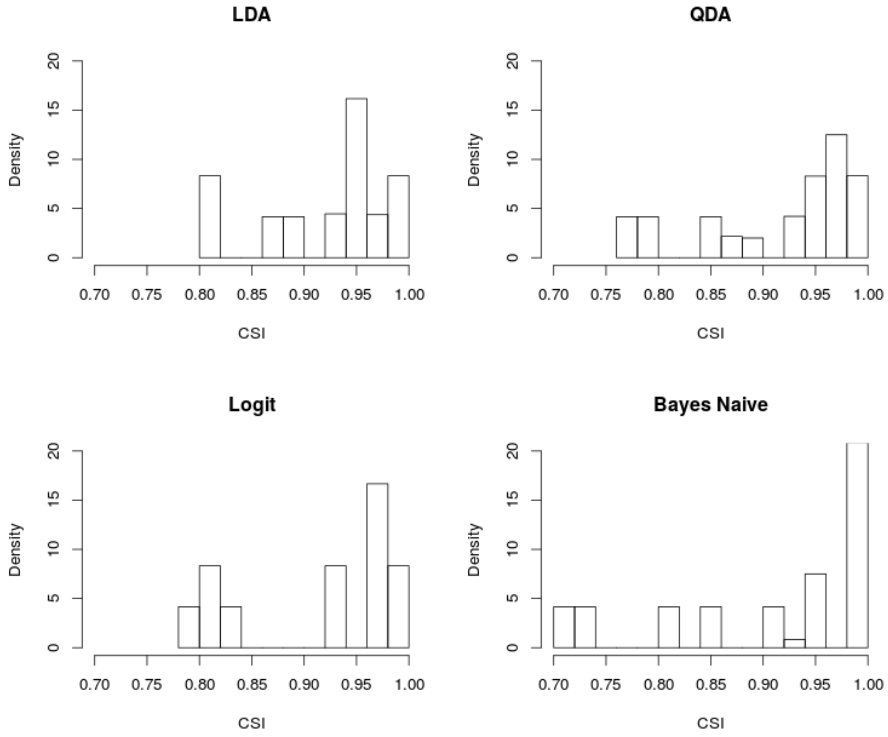

Figura 5.4: Histogramas de los valores de CSI obtenidos a partir de la simulación boostrap en los cuatro modelos. Primera fila, de izquierda a derecha: LDA, QDA. Segunda fila, de izquierda a derecha Logit, BN.

Como primera observación podemos notar que si bien los valores de las medianas en los cuatro modelos son muy semejantes entre sí, el modelo con menor varianza es el modelo lineal. Además el modelo de Bayes Naive, si bien tiene mayor nivel de varianza, es el que logra valores más altos de CSI. Es interesante notar que el modelo con menor varianza es justamente el modelo mas simple, es decir el modelo con menos parámetros que ajustar. Esto último es de esperarse si uno tiene en cuenta la relación sesgo-varianza presentes en estos modelos. De la misma forma, uno podría conjeturar que así como el modelo más simple es el que posee menor varianza también debería ser el más sesgado, lo cual no ocurre, o por lo menos no es evidente en este caso. Por consiguiente, en el caso de ser necesario elegir únicamente alguno de los cuatro modelos, en principio parecería que el lineal es el que mejor ajusta.

5.2. Simulación 2

De forma de completar el análisis anterior, testeamos el comportamiento de los cuatro modelos anteriores y del modelo "K vecinos próximos" tomando como covariables no solo `Rho`, `SZDR` y `V`, sino también sus cuadrados y sus productos de a dos. Notemos que un vector con coordenadas X y X^2, no tiene densidad. Sin embargo, quisimos explorar como se comportaban los modelos QDA y LDA

Modelo	**Covariables**	**Notación**
Modelo discriminante lineal	`Rho`, `V`, `SZDR`	LDA
Modelo discriminante cuadrático	`Rho`, `V`, `SZDR`	QDA
Modelo de regresión logística	`Rho`, `V`, `SZDR`	Logit
Modelo de Bayes Naive	`Rho`, `V`, `SZDR`	BN
Modelo K vecinos próximos ($k = 5$)	`Rho`, `V`, `SZDR`	KNN
Modelo discriminante lineal	`Rho`, `V`, `SZDR`, $\texttt{Rho}^2$, $\texttt{V}^2$, $\texttt{SZDR}^2$	LD1
Modelo discrimiante cuadrático	`Rho`, `V`, `SZDR`, $\texttt{Rho}^2$, $\texttt{V}^2$, $\texttt{SZDR}^2$	QD1
Modelo de regresión logística	`Rho`, `V`, `SZDR`, $\texttt{Rho}^2$, $\texttt{V}^2$, $\texttt{SZDR}^2$	LG1
Modelo K vecinos próximos ($k = 5$)	`Rho`, `V`, `SZDR`, $\texttt{Rho}^2$, $\texttt{V}^2$, $\texttt{SZDR}^2$	KNN1
Modelo discriminante lineal	`Rho`, `V`, `SZDR`, `Rho` $\times$ `V`, `Rho` $\times$ `SZDR`, `SZDR` $\times$ `V`	LD2
Modelo discriminante cuadrático	`Rho`, `V`, `SZDR`, `Rho` $\times$ `V`, `Rho` $\times$ `SZDR`, `SZDR` $\times$ `V`	QD2
Modelo de regresión logística	`Rho`, `V`, `SZDR`, `Rho` $\times$ `V`, `Rho` $\times$ `SZDR`, `SZDR` $\times$ `V`	LG2
Modelo K vecinos próximos ($k = 5$)	`Rho`, `V`, `SZDR`, `Rho` $\times$ `V`, `Rho` $\times$ `SZDR`, `SZDR` $\times$ `V`	KNN2

Tabla 5.3: Modelos analizados en la segunda simulación boostrap.

en estos escenarios. Para este caso, realizamos una segunda serie de simulaciones boostrap para las cuales consideramos como muestras de entrenamiento submuestras de 500 pixeles tomadas de manera aleatoria y con reposición de la muestra de entrenamiento total. El motivo de tomar muestras de entrenamiento mas pequeñas se explica por el costo computacional que implica ajustar cada uno de los modelos a evaluar. Para esta segunda simulación se testearon los trece modelos descriptos en las Tabla 5.3.

En las Figuras 5.5, 5.6, 5.7 y 5.8, se muestran los histogramas y los boxplots de los valores de CSI para cada uno de los modelos anteriores correspondientes a la segunda simulación boostrap. A partir de dichos gráficos, es claro que el modelo "K vecinos próximos", en sus tres versiones, parece ser el peor clasificador de los modelos analizados, ya que su mediana es la mas baja de todos los modelos, y para el primer y tercer caso, posee una varianza mayor en comparación del resto. En general es posible notar que agregar covariables a lo modelos no parece mejorar los resultados. En todos los casos, disminuye la performance del clasificador en cuanto a que, en mayor o menor medida, la varianza del estimador de CSI aumenta y disminuye el sesgo. Si bien modelos más grandes ajustan mejor en la muestra de entrenamiento (pues son más flexibles), el hecho de que al evaluar en la muestra de testeo no mejoren respecto de los más chicos (incluso empeoren) nos indica un sobreajuste por parte de estos modelos nuevos. Reafirmamos entonces a partir de este segundo análisis, que de los modelos estudiados, los modelos LDA y QDA parecen ser los que mejor ajustan

al problema de discriminación entre ecos meteorológicos y no meteoroloógicos en las imágenes del radar doppler banda C de doble polarización ubicado en la localidad de Anguil.

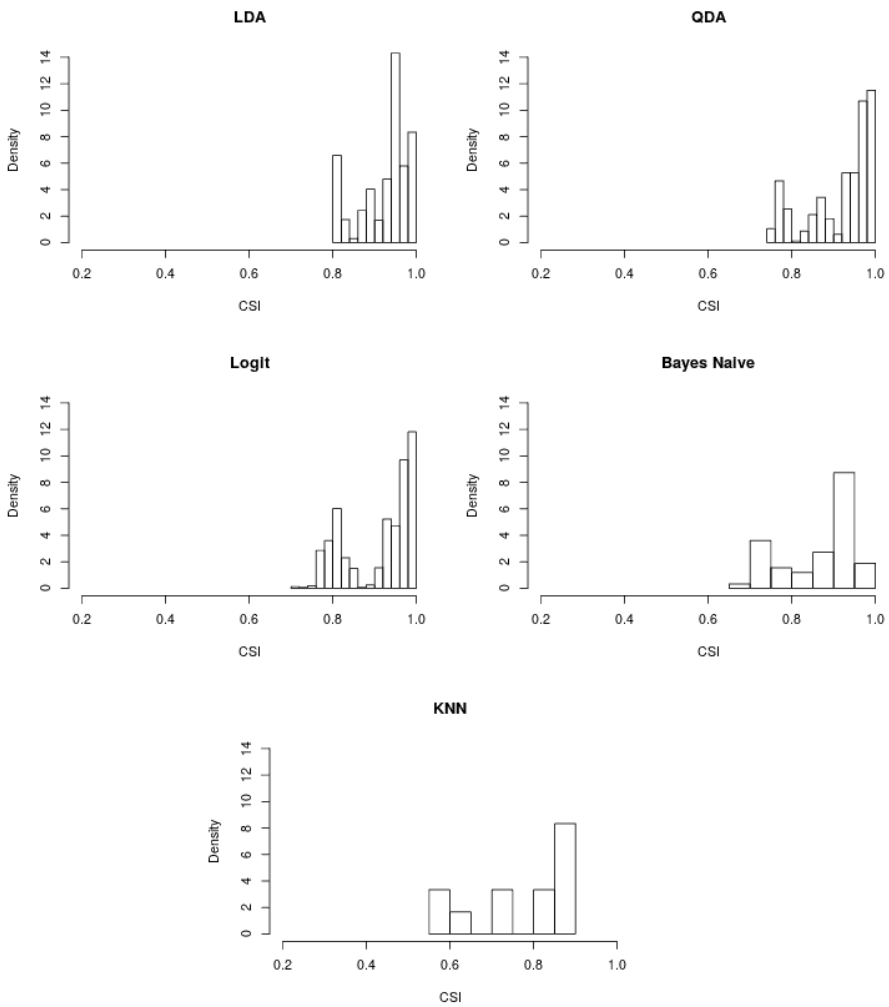

Figura 5.5: Histograma de los valores obtenidos de CSI a partir de simulaciones boostrap en los modelos: de izquierda a derecha, primera fila: LDA, QDA, segunda fila: Logit, BN, KNN.

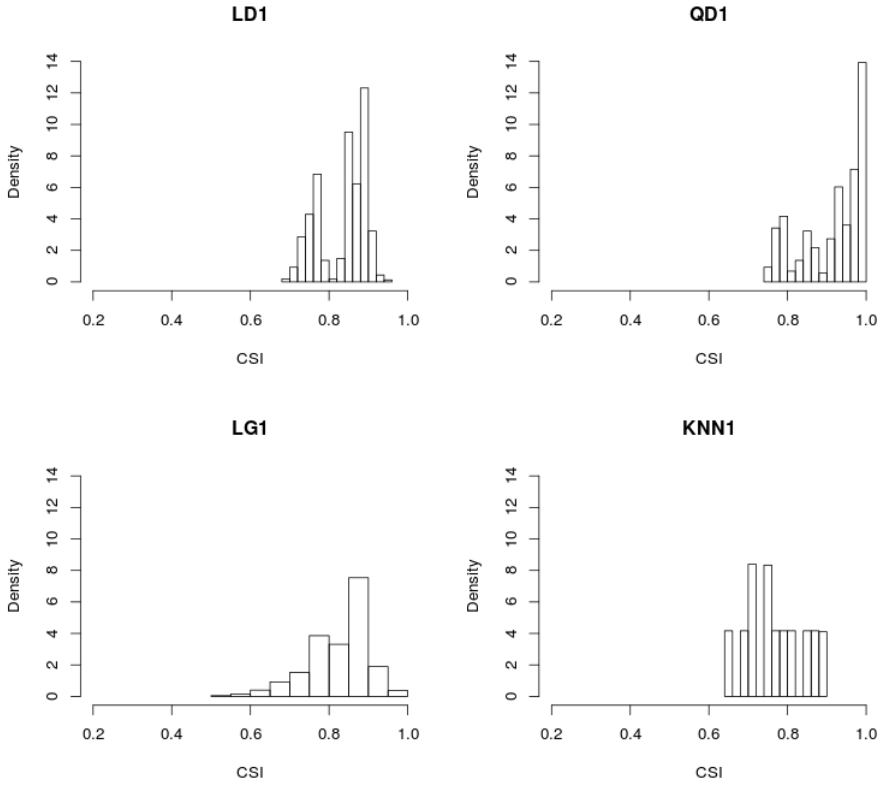

Figura 5.6: Histograma de los valores obtenidos de CSI a partir de simulaciones boostrap en los modelos: de izquierda a derecha, primera fila: LD1, QD1, segunda fila: LG1, KNN1.

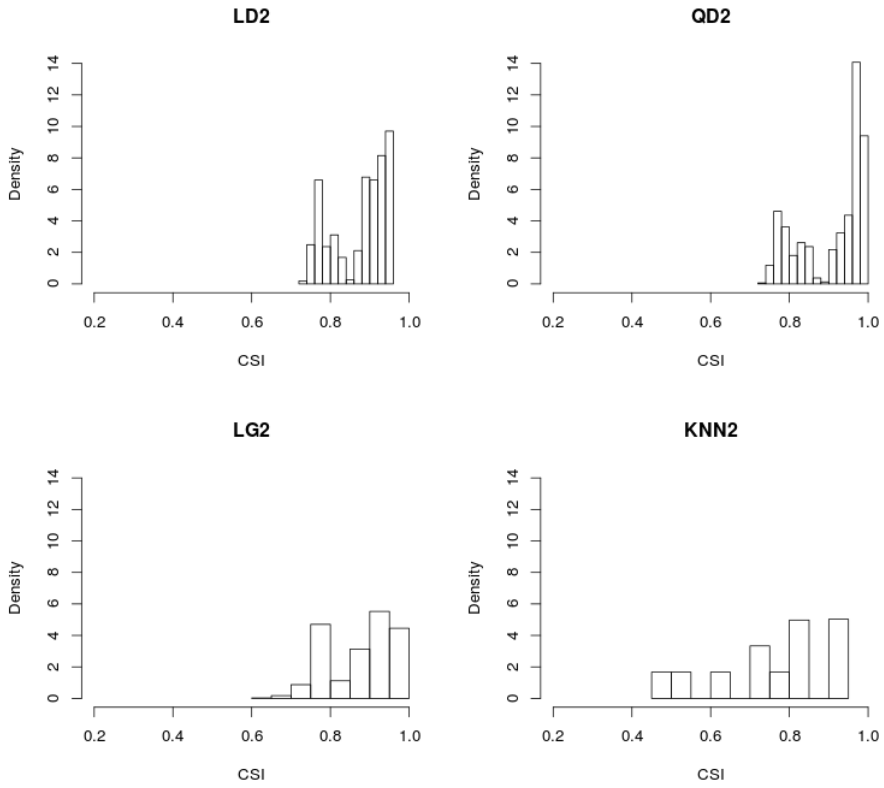

Figura 5.7: Histograma de los valores obtenidos de CSI a partir de simulaciones boostrap en los modelos: de izquierda a derecha, primera fila: LD2, QD2, segunda fila: LG2, KNN2.

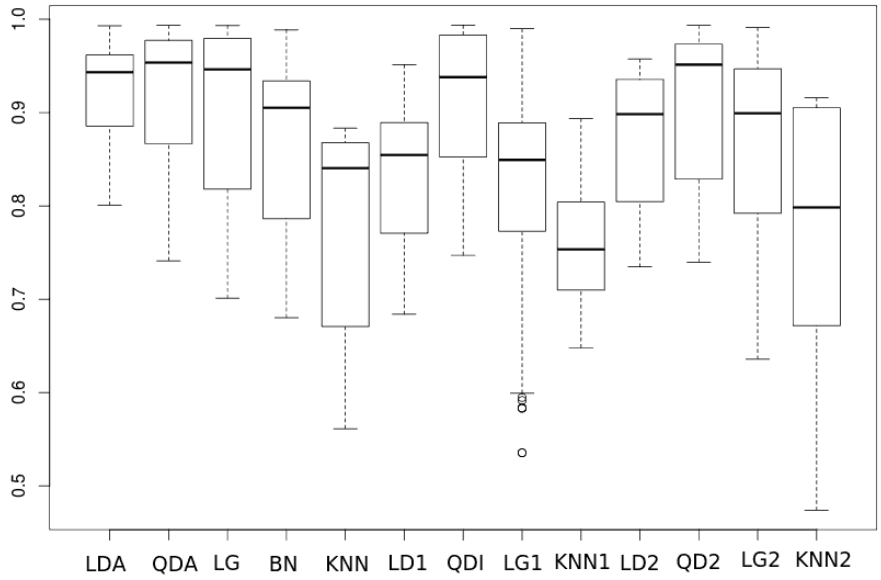

Figura 5.8: Boxplot de los valores de CSI obtenidos a partir simulacines boostrap los cuatro modelos: de izquierda a derecha: LDA, QDA, Logit, BN, KNN, LD1, QD1, LG1, KNN1, LD2, QD2, LG2, KNN2.

5.3. Conclusiones

En esta tesis se presentaron una serie de técnicas de clasificación supervisada, las cuales son aplicables a diversos problemas en distintos contextos. En particular, trabajamos con el problema de discriminación entre ecos meteorológicos y no meteorológicos en imágenes de radar. Si bien los resultados fueron buenos, consideramos necesario seguir avanzando en esta línea de investigación. Como primera medida, creemos que incorpar nuevas variables polarimétricas al análisis ayudaría a obtener mejores resultados. Para ello será necesario trabajar en la calibración de las mismas sobre los radares que se encuntran operativos. En segunda instancia tal como consideramos la variable `SZDR` para incorporar nociones de la estructura espacial en sentido horizontal, estimamos que incluir también variables que den nociones de la estructura espacial en el sentido vertical sería positivo. Por otra parte, como en esta tesis trabajamos con técnicas clásicas de clasificación, creemos que es necesario en un futuro poder explorar y evaluar algunas técnicas mas avanzadas, como árboles de desición, redes neuronales, técnicas de clusterización, entre otras. Por último esperamos extender estás técnicas a otros radares de doble polarización existentes en nuestro país.

Bibliografía

[1] Berenguer M, Sempere-Torres D, Corral C,Sánchez-Diezma R (2006): A fuzzy logic technique for identifying nonprecipitating echoes in radar scans.*J. Of Atmos. And Oceanic Tech*, vol. 23, p 1157-1180.

[2] Breiman.L, Friedman.J, Stone C.J, Olshen R.A (1984): Classification and regression trees. *Chapman and Hall/CRC.*

[3] Cho.Y.H, Lee.G, Kim K.E, Zawadski.I (2006): Identification and removal of ground echoes and anomalous propagation using the characteristics of radar echoes. *J.Of Atmos. And Oceanic Tech*, vol. 23, p 1206-1222.

[4] Delicado.P (2008): Curso de Modelos no Paramétricos. *Departamento de estádistica e incestigación operativa, Universidad de Cataluña.*

[5] Everitt.B, Hothorn.T (2001): An introduction to applied Multivariate Analysis with R *Springer Texts in Statistics.*

[6] Gourley.J, Tabary.P, Chatelet.J.P (2007): A Fuzzy Logic Algorithm for the Separation of Precipitating fron Nonprecipitanting Echoes Using Polarimetric Radar Obsservations, *J. Of Atmos. And Oceanic Tech*, vol. 24, p 1439-1451.

[7] Hastie.T, Witten.D, Tibshirani.R, James.G (2013): An introduction to Statical Learning with Applications in R. *Springer Texts in Statistics.*

[8] Hastie.L, Tibshirani.R, Friedman.J (2009): The Elements of statical learning: Data Mining, Inference, and Prediction *Springer Text in Statistics.*

[9] Moszkowicz.S, Cian G.J, Krajewski w.f (1993): Statistical Detection of Anamalous Propagation in Radar Reflectivity Patterns.*J. Of Atmos. And Oceanic Tech*, vol. 11, p 1026-1034.

[10] Peña.D (2002): Análisis de datos Multivariantes. *S.A. Mcgraw-hill / Interamericana de España.*

[11] Rico-Ramirez.M.A, Cluckie I.D (2008): Classification of ground clutter and anomalous propagation using dual-polarization weather radar. *IEEE Tran- sactions on geoscience and remote sensing*, vol. 46, no.7, p 1892-1904.

[12] Schuur T., Ryzhkov A., Heinselman P. (2003): Observations and classification of echoes with the polarimetric WSR-88D radar. *National Severe Storms Laboratory (NOAA) and Cooperative Institute for mesoscale Meteorological Studies (University of Oklahoma).*

[13] Silverman B.W (1986): Density estimation for statistics and data analysis. *Chapman and Hall/CRC.*

[14] Steiner M., Smith J.A (2001): Use of Three-Dimensional Reflectivity Structure for Automated Detection and Removal of Nonprecipitating Echoes in Radar Data. *J.Of Atmos. And Oceanic Tech*, vol. 19, p 673-685.

[15] Villarini.G, Krajewski.W.F (2010): Sensitivity Studies of the Models of Radar-Rainfall Uncertainties. *J. Appl. Meteor. Climatol*,vol. 49, p 288–307.

[16] Wasserman.L (2004): All of Statistics: A Concise Course in Statistical Inference. *Springer Texts in Statistics.*

[17] Wasserman.L (2007): All of Nonparametric Statistics. *Springer Texts in Statistics.*