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

Tesis de Licenciatura

Clasificación no supervisada: El método de k–medias

Yanina Gimenez

Director: Ricardo Fraiman

Fecha de Presentación: 23 de marzo de 2010

Dedicatoria y agradecimientos

A mis padres por apoyarme en mis decisiones y alegrarse por cada uno de mis logros. A mi papá por compartir conmigo el gusto por las matemáticas. A mi mamá por todo el cariño que me brindó intentando siempre ayudarme de algún modo.

A Marisol, mi hermana, quien me acompañó en el día a día compartiendo todo conmigo dándome lo mejor de ella y haciendo que todo sea más sencillo.

A Maury por su amor y por su paciencia, porque siempre buscó y encontró como ayudarme dando hasta su aporte en mi tesis. Por estar siempre a mi lado.

A Ricardo Fraiman por ser mi director, por transmitirme sus conocimientos con tanta alegría y entusiasmo.

A Marce por responder todas mis preguntas siempre con la mejor predisposición.

A todos los profesores que contribuyeron a mi formación durante mis cursos de Licenciatura. En especial a Miguel por la paciencia que ha tenido frente a mis consultas frecuentes.

A las Monis por ser mis amigas de la vida, porque nada hubiese sido igual sin su continua amistad. Por saber que ellas siempre están.

A mis amigos de la facultad, entre los que quiero destacar a Quimey, Julián, Xime, Vero, Martín y Lucas. Por las horas de estudio y las largas charlas. Por escuchar mis dudas y mis preocupaciones y aconsejarme del mejor modo.

Índice

1. Introducción

En esta monografía estudiamos uno de los métodos de "cluster analysis" más utilizados en la práctica: el método de k–medias.

El término cluster analysis (usado por primera vez por Tryon, 1939) se refiere a diferentes métodos para agrupar objetos de tipo similar en categorías. Más recientemente se le suele llamar clasificación no supervisada.

Es una técnica de análisis exploratorio que intenta ordenar diferentes objetos en grupos, de forma que el grado de asociación entre dos objetos sea máximo si pertenecen al mismo grupo.

Esta formulación es claramente vaga, y dependerá de la forma con la que optemos por considerar que un determinado procedimiento cumple con estos requerimientos. ¿Entonces, como sabremos si el procedimento considerado cumple con los objetivos planteados? En algún sentido, cada método tiene una función objetivo poblacional que será en definitiva la que responda a esta pregunta. En este marco, no encontraremos un procedimiento óptimo para todo tipo de datos, sino que cada método de cluster esta diseñado de alguna forma (a través de su función objetivo) para cierto tipo de datos.

Cluster analysis se usa para encontrar estructuras en los datos sin proveer una explicación/interpretación. Cluster analysis simplemente descubre estructuras en los datos sin explicar en principio por qué existen. Será el analista quien a posteriori interprete los posibles motivos de esa estructura.

Hemos organizado esta monografía de la siguiente forma. En la Sección 2 describimos algunos métodos de cluster, introducimos el método de k-medias y damos un ejemplo de aplicación a datos de chips de ADN de tumores humanos. En la Sección 3 estudiamos el problema de la consistencia casi segura del método de k-medias. Esta demostración hace uso de técnicas de procesos empíricos que han demostrado ser muy útiles para la teoría asintótica en estadística. Eso motiva considerar en las Secciones 4 y 5 algunos resultados generales de convergencia uniforme para procesos empíricos. Finalmente en el Apéndice incluímos la demostración de algunos resultados que hemos utilizado en la demostración de los resultados de las secciones previas.

2. Métodos de Análisis de cluster

Los métodos de cluster, también llamados segmentación de datos, tienen una gran variedad de objetivos. Todos relacionados con agrupar o segmentar una colección de objetos en subconjuntos o "clusters", de manera que dentro de cada grupo están más estrechamente relacionados entre sí que con los que están en diferente grupo.

Un objeto puede ser descripto por un conjunto de medidas, o por su relación con otro objeto. A veces el objetivo es organizar a los conjuntos en una jerarquía (métodos jerárquicos). Ésto implica agrupar sucesivos clusters de modo tal que en cada nivel de la jerarquía los clusters que están dentro de un mismo grupo son más similares entre ellos que con los que pertenecen a un grupo diferente.

El análisis de cluster también es usado en la estadística descriptiva para determinar si los datos pertenecen o no a distintos subgrupos, donde cada grupo representa objetos con propiedades sustancialmente distintas. Este objetivo final requiere de información respecto al grado de diferencia entre objetos asignados a distintos clusters.

Como mencionamos en la introducción, un aspecto central de cada procedimiento de cluster es la noción de grado de semejanza (o diferencia) entre los objetos que van a ser agrupados. Los métodos de cluster intentan agrupar los objetos basándose en la definición de semejanza que hay entre ellos. Esta definición se construirá en base a los aspectos propios del problema, y debería ser independiente del método de clustering en sí.

2.1. Las matrices de proximidad

A veces los datos están representados directamente en términos de proximidad (semejanza o afinidad) entre pares de objetos. Esto puede ser, o bien por similitudes o disimilitudes (la diferencia o la falta de afinidad). Por ejemplo, en experimentos de ciencias sociales, los participantes tienen que juzgar en cuanto ciertos objetos difieren entre si. La desemejanza puede ser entonces computada como el promedio de la colección de tales juicios. Este tipo de datos puede ser representado con una matriz D de $N \times N$, donde N es el número de objetos, y cada elemento $d_{ii'}$ representa la proximidad entre el objeto i y el objeto i'. Esta matriz es entonces el input del algoritmo de cluster.

La mayoría de los algoritmos presumen una matriz de desemejanzas con enteros no negativos y ceros en los elementos de la diagonal: $d_{ii} = 0$, $i = 1, 2, \ldots, N$. Si los datos originales son tomados como semejantes una adecuada función monótona decreciente puede convertirlos en desemejantes. Por tanto, la mayoría de los algoritmos asumen matrices simétricas desemejantes, y si la matriz original D es no simétrica va a ser reemplazada por $(D+D^T)/2$. Subjetivamente, las diferentes opiniones rara vez están tan distanciadas como las distancias en el sentido estricto, ya que la desigualdad triangular $d_{ii'} \leq d_{ik} + d_{i'k}$, $\forall k \in \{1, \ldots, N\}$ no se sostiene. Por eso algunos algoritmos asumen que la distancia no puede ser usada como un dato.

2.2. Desemejanzas basadas en atributos

Más a menudo tenemos medidas x_{ij} para $i = 1, 2, \ldots, N$, con variables $j = 1, 2, \ldots p$ (también llamadas atributos). Dado que la mayoría de los más populares algoritmos de cluster toman una matriz de desemejanzas como su input, tenemos que en primer lugar construir pares de diferencias entre las observaciones. En los casos más comunes definimos la desemejanza $d_j(x_{ij}, x_{i'j})$ entre valores del atributo j y luego definimos

$$D\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} d_j\left(x_{ij}, x_{i'j}\right), \tag{1}$$

como la desemejanza entre el objeto i e i'. Por lejos la más común de las elecciones es la distancia cuadrática

$$d_j\left(x_{ij}, x_{i'j}\right) = \left(x_{ij} - x_{i'j}\right)^2.$$

Sin embargo, otras elecciones son posibles y pueden llevar a potencialmente diferentes resultados. Para atributos no cuantitativos (ej. datos categóricos), la distancia cuadrática no es la apropiada. Además, a veces se prefiere darle diferente pesos a distintos atributos antes de darle el mismo peso a todos como en (1).

Primero discutamos diferentes alternativas en términos de los diferentes tipos de atributos:

• *Variables cuantitativas.* Mediciones de este tipo de variables o atributos son representados con valores reales continuos. Es natural definir el "error" entre ellos como una función monótona creciente de la diferencia de sus valores absolutos

$$d\left(x_i, x_{i'}\right) = l\left(\left|x_i - x_{i'}\right|\right).$$

Además de la función cuadrática $l(u) = u^2$ (que da origen a la distancia Euclideana), una elección frecuente es $l(u) = u$ que da origen a la distancia L^1. Una alternativa es que la agrupación esté basada en la correlación

$$\rho\left(x_i - x_{i'}\right) = \frac{\sum_j \left(x_{ij} - \overline{x}_i\right)\left(x_{i'j} - \overline{x}_{i'}\right)}{\sqrt{\sum_j \left(x_{ij} - \overline{x}_i\right)^2 \sum_j \left(x_{i'j} - \overline{x}_{i'}\right)^2}},$$

con $\overline{x}_i = \sum_j x_{ij}/p$. Notemos que este promedio es sobre variables, no sobre observaciones. Si al input primero lo estandarizamos, entonces

$$\sum_j \left(x_{ij} - x_{i'j}\right)^2 \propto 2\left(1 - \rho\left(x_{i,} x_{i'}\right)\right).$$

Por lo tanto la agrupación basada en correlación (de semejanza) es equivalente a la basada en distancia cuadrática (de desemejanzas).

• *Variables ordinales.* El valor de este tipo de variables esta mayormente representado con enteros contiguos, y los valores realizados son considerados un conjunto ordenado. Ejemplo de esto son los grados académicos (A, B, C, D, E, F) y grados de preferencias (no puedo soportar, me disgusta, está bien, me gusta, me encanta). El rango de datos es un tipo especial de datos ordinales. Errores de medición de variables ordinales son generalmente definidos reemplazando su valor original M con

$$\frac{i-1/2}{M},\ i=1,\ldots,M$$

en los órdenes prescriptos de sus valores originales. Son entonces tratados como variables cuantitativas en esta escala.

• *Variables categóricas.* Con un desorden categórico (también llamado nominal) de las variables, el grado de diferencia entre pares de valores tiene que ser definido explícitamente. Si la variable toma M valores distintos, entonces puede ser organizado con una matriz simétrica de $M \times M$ con elementos $L_{rr'} = L_{r'r}$, $L_{rr} = 0$, $L_{rr'} \geq 0$. La elección más común es $L_{rr'} = 1$ para todo $r \neq r'$, porque la pérdida de la igualdad puede ser usada para enfatizar más un error que otro.

2.3. Desemejanzas entre objetos

Ahora definamos un procedimiento que combina los p-individuales atributos de desemejanzas $d_j\left(x_{ij}, x_{i'j}\right), j = 1, 2, \ldots, p$ con una medida global de desemejanzas $D\left(x_i, x_{i'}\right)$ entre dos objetos u observaciones $(x_i, x_{i'})$ que poseen los valores de atributos respectivos. Esto es casi siempre hecho por medio de una media ponderada

$$D\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} w_j \cdot d_j\left(x_{ij}, x_{i'j}\right);\ \sum_{j=1}^{p} w_j = 1. \tag{2}$$

Acá w_j es el peso asignado a la atribución j regulando la influencia relativa de esa variable para determinar la diferencia total entre objetos. Esta elección debe ser basada en la importancia que se le dé a cada variable.

Es importante darse cuenta que asignar el mismo peso w_j a todos los valores de cada variable (es decir, $w_j = 1\ \forall j$), no necesariamente le da a todos los atributos igual influencia. La influencia del atributo j, en el objeto de desemejanzas $D\left(x_i, x_{i'}\right)$ (2), depende, su relativa contribución, de la medida promedio del objeto de desemejanzas sobre todos los pares de observaciones del conjunto de datos

$$\overline{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} D\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} w_j \cdot \overline{d}_j,$$

con

$$\overline{d}_j = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_j\left(x_{ij}, x_{i'j}\right), \tag{3}$$

el promedio de desemejanzas en el atributo j. Así, la influencia relativa de la variable j es $w_j \cdot \overline{d}_j$, y poniendo $w_j \sim \overline{d}_j$ daría a todos los atributos la misma influencia en la caracterización general entre objetos. Por ejemplo, con p variables cuantitativas y usando la distancia del error cuadrático para cada coordenada, entonces (2) es la cuadrática distancia Euclidea

$$D_I\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} w_j \cdot \left(x_{ij} - x_{i'j}\right)^2,$$

entre pares de puntos en $\mathbb{R}^p$, con las variables cuantitativas como ejes. En este caso (3) sería

$$\overline{d}_j = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(x_{ij} - x_{i'j}\right)^2 = 2 \cdot var_j,$$

donde var_j es la estimación de la muestra de $Var\left(X_j\right)$. Así, la importancia relativa de cada una de dichas variables es proporcional a la variación en el conjunto de datos. En general poner peso $w_j = 1/\overline{d}_j$ a todos los atributos, independiente del tipo, va a provocar que cada uno de ellos tenga la misma influencia sobre la diferencia total entre los pares de objetos $(x_i, x_{i'})$. A pesar de que parezca ser razonable, y es en muchos casos recomendado, puede ser altamente contraproducente. Si el objetivo es segmentar los datos en grupos de objetos similares, todos los atributos no pueden tener la misma influencia. Algunas diferencias en los valores de los atributos pueden reflejar una mayor desigualdad en el contexto del objeto real del problema.

Si el objetivo es descubrir grupos naturales en los datos, algunos atributos deben demostrar una mayor tendencia a agrupar que otros. A las variables que son más relevantes para separar los grupos, se les debería de asignar un mayor peso en la definición de desemejanzas entre objetos. Darle a todos los atributos el mismo peso en este caso produciría una tendencia a ocultar los grupos de puntos donde los algoritmos de cluster no pueden acceder. Aunque las elecciones individuales del atributo para las desemejanzas $d_j\left(x_{ij}, x_{i'j}\right)$ y de sus pesos w_j pueden ser una herramienta adecuada, no hay un sustituto para el pensamiento cuidadoso que se debe tener en el contexto de cada problema. Especialmente una apropiada medida de desemejanzas es a menudo mucho más importante para obtener un resultado exitoso con cluster que la elección del algoritmo. Finalmente, muchas observaciones tienen valores faltantes en uno o más atributos. El método más común para incorporar estos valores faltantes en el cálculo de desemejanzas (2) es omitir de cada par de observaciones al menos un valor $x_{ij}, x_{i'j}$, cuando se computa la diferencia entre las observaciones x_i e $x_{i'}$. Este método puede fallar si se dan las circunstancias de que ambas observaciones no tienen valores medibles en común. En este caso ambas observaciones pueden ser eliminadas del análisis. Otra alternativa es incorporar los valores faltantes como input usando la media o la mediana de cada atributo sobre los datos no faltantes. Para variables categóricas uno puede considerar el valor "faltante" como otro valor categórico, si fuera razonable considerar que dos objetos son semejantes si ambos tienen valores faltantes en la misma variable.

K-Medias

El algoritmo de las k-medias es uno de los más populares iterativos descendentes métodos de cluster. Está destinado a situaciones en las cuales todas las variables son del tipo cuantitativo, y la distancia cuadrática Euclidea

$$d\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \left\|x_i - x_{i'}\right\|^2,$$

es elegida como medida de diferencia. Notemos que los pesos en la distancia Euclidea pueden ser usados redefiniendo los valores x_{ij}.

Los puntos de dispersión pueden ser escritos como

$$\begin{aligned} W\left(C\right) &= \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \left\|x_i - x_{i'}\right\|^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k} \left\|x_i - \overline{x}_k\right\|^2, \end{aligned}$$

donde $\overline{x}_k = \left(\overline{x}_{1k}, \ldots, \overline{x}_{pk}\right)$, es el vector de medias asociado con el k-ésimo cluster, y $N_k = \sum_{i=1}^{N} I\left(C\left(i\right) = k\right)$. Así, el criterio es asignar las N observaciones a los K clusters de modo que dentro de cada cluster el promedio de las diferencias de cada observación a la media del cluster, definido por los puntos del cluster, sea mínima.

2.4. Ejemplo: Chips de ADN y los tumores humanos

Apliquemos k-medias a los datos de chips de ADN de tumores humanos.

Describamos de que se trata:

La célula es la unidad básica de vida. Un componente escencial de la misma son las proteínas, las cuales tienen funciones de lo más diversas, tales como formar el esqueleto propio de la célula o servir de receptor para señales hormonales (entre muchas otras).

La estructura particular de cada proteína es la que le da su función característica. Cada proteína está compuesta por una cadena de aminoácidos (moléculas biológicas de relativa simplicidad) y la estructura depende de los aminoácidos que la componen.

La información de la secuencia de aminoácidos que compone cada proteína está contenida en el ADN. La cadena ADN es una molécula de gran complejidad formada por 4 nucleótidos (abreviados con las letras A, C, G y T) que se alternan uno tras otro formando un verdadero código (el genoma). En este código está comprendida la información de la secuencia de aminoácidos que compone cada proteína del cuerpo.

Para que el código del genoma se transforme en una proteína, son necesarios varios pasos (ver Figura 1). Primero una molécula llamada transcriptasa lee el código del ADN y forma otra molécula más corta y manejable, llamada ARN mensajero (ARNm) conteniendo el fragmento del código del ADN necesario para formar la proteína en cuestión. A la información necesaria para crear una proteína se la denomina gen. Luego este ARNm es usado por otras moléculas llamadas ribosomas que leen el código del ARNm y lo traducen a aminoácidos que se van pegando uno tras otro. Cuando una célula fabrica cierta proteína, se dice que el gen de esa proteína se expresa en esa célula.

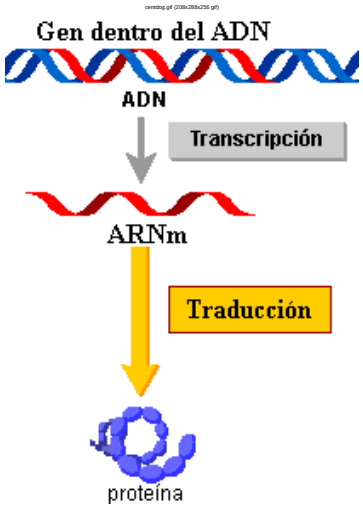

Figura 1: Etapas para la transformación del código del genoma en proteínas.

Los chips de ADN son diminutas placas que sirven para verificar la presencia de

determinadas proteínas en una célula contando los ARN mensajeros presentes en la misma (ver Figura 2).

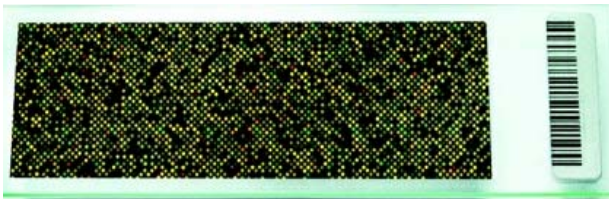

Figura 2: Imagen de un chip de ADN

Los chips de ADN contienen una microgrilla en la cual cada cuadrado tiene adosado distintas secuencias de ADN ya conocidas por el fabricante. Cada una de estas secuencias es un fragmento correspondiente únicamente a un gen.

Además de la muestra celular a ser estudiada (muestra objetivo) se usa otra muestra celular de referencia. Las células de ambas muestras son destruídas liberando todo el contenido celular (proteínas, ADN, ARNm, y muchos otras moléculas) en sus respectivos recipientes. La muestra de referencia tiene un contenido de proteínas (y sus respectivos ARNm) ya conocido por el biólogo. Cada una de las muestras es teñida con un colorante diferente (rojo y verde por ejemplo). Luego el chip de ADN se expone a cada una de las muestras. Si algún ARNm contenido en las muestras corresponde al ADN de alguno de los cuadrados del chip, estos se acoplarán fuertemente (un proceso llamado hibridación). Luego se remueven todos los excedentes celulares del chip, dejando únicamente los ARNm que se hayan hibridizado a algún cuadrado.

Después se registra la intensidad de rojo/verde en cada cuadrado del chip midiendo así cuánto ARNm se ha hibridizado en cada sitio.

La fluoroscopía da como resultado números, que normalmente están entre -6 y 6, midiendo la cantidad de cada ARNm en la muestra objetivo en comparación con la muestra de referencia. Los valores positivos indican una mayor expresión en la muestra objetivo en comparación con la referencia, y valores negativos lo contrario.

En un conjunto de datos de estudios con chips de ADN, cada columna representa una muestra celular estudiada y en las filas se colocan los genes de las proteínas de mayor relevancia. En el ejemplo particular de la Figura 3 hay 6830 genes (filas) y 64 muestras (columnas), aunque por mayor claridad sólo una muestra aleatoria de 100 filas es mostrada. La figura expone el conjunto de datos como un mapa de calor, con una gama de colores que va del verde (para los valores negativos, es decir con menos de esa proteína que en la muestra de referencia) al rojo (para los valores positivos, con más de esa proteína que en la muestra de referencia). Las muestras son de 64 tumores de cáncer de diferentes pacientes.

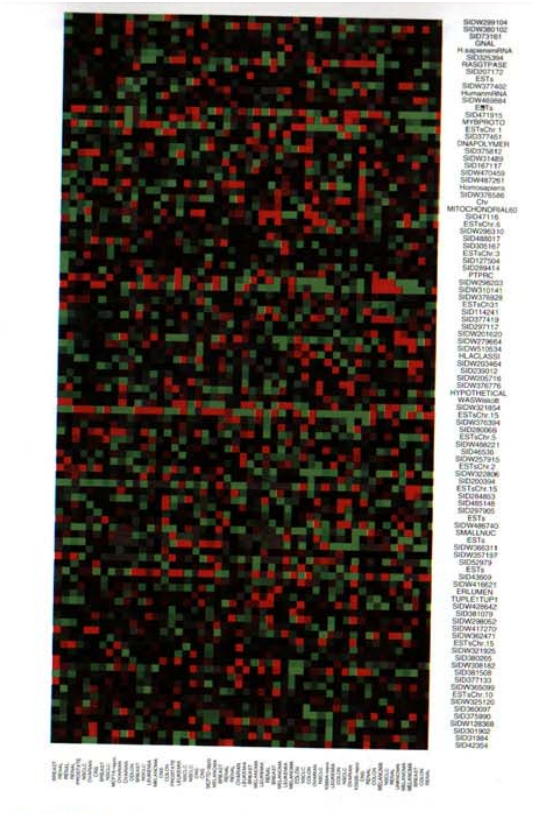

Figura 3: Vista parcial de datos de un chip (microarray) de DNA. Es una matriz con 6830 filas (genes) y 64 columnas (muestras) para datos de tumores humanos. Se muestran 100 columnas elegidas al azar.

El desafio es entender cómo los genes y las muestras se organizan. Entre las preguntas típicas están incluidas:

(a) ¿qué tumores son más similares a cuáles, en términos de la expresión a traves de sus genes?

(b) ¿qué genes son más similares a cuáles, según su aparición en las distintas muestras?

(c) ¿qué genes aparecen mucho o poco en cierto tipo de muestras de cáncer?

Podríamos ver estas preguntas como un problema de regresión, con dos categóricas variables predictoras (genes y muestras), siendo la variable de respuesta el nivel de expresión. Sin embargo, probablemente sea más útil verlo como un problema de aprendizaje no supervisado. Por ejemplo para la pregunta (a), podemos pensar a las muestras como 6830 puntos en el espacio que queremos agrupar de cierto modo.

Esto es un ejemplo de un cluster de alta dimensión. Los datos son una matriz real de tamaño 6830×64, cada número representa cuánto se expresa un gen (fila)

en determinada muestra (columna). Agrupamos las muestras de manera que cada una es un vector de tamaño 6830, correspondiendo al valor de la expresión de cada uno de los 6830 genes. Cada muestra está etiquetada: **mama** (para el cancer de mama), **melanoma**, y así sucesivamente; no usamos estas etiquetas para el cluster, pero vamos a examinar después cual etiqueta cae en que cluster.

Aplicamos k-medias con $k = 3$, obteniendo los siguientes 3 clusters:

Cluster	Mamas	CNS	Colon	K562	Leucemia	MCF7
1	3	5	0	0	0	0
2	2	0	0	2	6	2
3	2	0	7	0	0	0

Cluster	Melanoma	NSCLC	Ovarios	Próstata	Renal	Desconocido
1	1	7	6	2	9	1
2	7	2	0	0	0	0
3	0	0	0	0	0	0

Podemos observar que el procedimiento es exitoso agrupando muestras de un mismo cáncer. De hecho, los dos cáncer de mamas que fueron ubicados en el cluster 2, luego se descubrió que habían sido diagnosticados mal y que en realidad eran melanomas que tuvieron metástasis.

3. Consistencia de k–medias

Expliquemos lo que vamos a hacer en este capítulo. El análisis de cluster por k-medias prescribe un criterio de cómo partir un conjunto de puntos en k-grupos. Para dividir los puntos $X_1, \ldots, X_n$ de $\mathbb{R}^s$ acordes a este criterio, primero tenemos que elegir los centros de los cluster de modo que minimicen

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|X_i - a_j\|^2 ,$$

donde $\|\cdot\|$ denota la usual norma Euclidea, luego asignamos cada X_i al cluster cuyo centro esté más cercano. De esta manera, cada centro a_j adquiere un subconjunto C_j de los $X's$ en su cluster. La media de los puntos en C_j tiene que ser a_j, sino podemos achicar a W_n reemplazando a a_j por la media de este cluster, si es necesario reasignamos algunos $X's$ a nuevos centros. El crieterio es, por lo tanto, equivalente a minimizar la suma de cuadrados dentro de cada cluster.

Vamos a asumir que $\{X_1, \ldots, X_n\}$ es una muestra de observaciones independientes e idénticamente distribuidas con cierta distribución $\mathcal{P}$. Vamos a pedir ciertas condiciones que aseguren la convergencia casi segura a los centros de los clusters cuando el tamaño de la muestra crece. Esto generaliza uno de los resultados de Hartigan (1978), quien hace un análisis detallado de la fragmentación de observciones en una dimención en dos clusters, él prueba convergencia en probabilidad de los puntos definiendo la óptima partición.

MacQueen (1967) obtenía resultados débiles de consistencia para algoritmos de k-medias que distribuían los puntos $X_1, X_2, \ldots$ sequencialmente en k-clusters. Con este algoritmo los centros no son elegidos para minimizar W_n; sin embargo, cada X_n es asignado al cluster con el centro más cercano, luego este centro es movido a la media del cluster. MacQueen prueba que el correspondiente W_n converge casi seguramente; él no prueba convergencia de los centros de los clusters.

Por las dificultades que pueden surgir por ambigüedades en la asignación de los puntos $X_1, \ldots, X_n$ a los centros $a_1, \ldots, a_k$, es ventajoso considerar W_n como una función de los conjuntos de centros de clusters y de la medida empírica P_n obtenida de la muestra al colocarle peso n^{-1} a cada $X_1, \ldots, X_n$. Esto es el problema de minimizar

$$W(A, P_n) := \int \min_{a \in A} \|x - a\|^2 P_n(dx),$$

sobre todas las posibles elecciones del conjunto A que contenga k (o menos) puntos. Para cada A la ley fuerte de los grandes números dice que

$$W(A, P_n) \to W(A, \mathcal{P}) := \int \min_{a \in A} \|x - a\|^2 \mathcal{P}(dx), \text{ cs.}$$

Se espera por lo tanto que A_n, el conjunto de los óptimos centros de la muestra, esté cerca de $\bar{A}$, el conjunto de centros que minimiza $W(\cdot, \mathcal{P})$, siempre que $\bar{A}$ esté unívocamente determinado. Con una apropiada definición de cercano, éste es el resultado que vamos a probar. Esto implica que hay una asignación $a_{n1}, a_{n2}, \ldots, a_{nk}$ de los puntos de A_n, y una asignación $\bar{a}_1, \ldots, \bar{a}_k$ de los puntos de $\bar{A}$, tal que $a_{nl} \to \bar{a}_l$ cs. Este enfoque también evita problemas con la posible coincidencia de dos de los centros de los clusters; una posibilidad que le ocasionaba mucho problema a MacQueen(1967).

En la práctica, encontrar un A para el cual $W(\cdot, P_n)$ alcanza su mínimo global involucra una cantidad prohibida de cálculos, excepto en el caso de dimención 1 ($s = 1$). Este problema ha sido discutido por Hartigan. Sin embargo, existen algoritmos eficientes que encuentran localmente particiones óptimas para la muestra de puntos en k clusters.

Nuestro método de prueba está basado en repetir aplicaciones de la ley de los grandes números; los argumentos se aplican a casi toda la muestra de puntos ω. Primero mostraremos que la muestra óptima de centros del cluster A_n está en una región compacta de $\mathbb{R}^s$. La prueba de ésto es inductiva empezando por el simple caso de 1-media. Una vez que está establecida la consistencia para $(k-1)$-medias , el caso de k-medias es tratado primero mostrando que existe una bola compacta $B(M)$ que contiene al menos un punto del óptimo A_n, para n suficientemente grande. Si ésto no fuera así, $W(\cdot, P_n)$ podría decrecer (cuando n es suficientemente grande) moviendo todos los centros de los clusters a un solo punto. Luego mostraremos que la bola compacta $B(5M)$ contiene todos los puntos de A_n. Es más, $B(5M)$ puede ser agrandada tanto que cualquier punto de A_n que se encuentre fuera de la $B(5M)$ puede ser borrado sin que $W(A_n, P_n)$ aumente. Esto daría un conjunto de como mucho $k-1$ puntos para el cual si n es lo suficientemente grande, $W(\cdot, P_n)$ es menor que el mínimo valor para $k-1$ medias; y ésto es una contradicción.

El segundo paso en la prueba involucra mostrar la convergencia uniforme casi segura de $W(A, P_n) - W(A, \mathcal{P})$ sobre los subconjuntos de la $B(5M)$ que contienen k o menos puntos.

Funciones objetivo poblacional y empírica. Dada una medida de probabilidad $\mathcal{Q}$ en $\mathbb{R}^s$ para cada conjunto finito A de $\mathbb{R}^s$ definimos:

$$\Phi(A, \mathcal{Q}) := \int \min_{a \in A} \|x - a\|^2 \mathcal{Q}(dx), \tag{4}$$

y

$$m_k(\mathcal{Q}) := inf\{\Phi(A, \mathcal{Q}) : A \text{ contiene } k \text{ o menos puntos}\}.$$

La función objetivo poblacional corresponderá a tomar $\mathcal{Q} = \mathcal{P}$ en (4), mientras que la empírica lo hará para $\mathcal{Q} = P_n$.

Para un dado k, el conjunto $A_n = A_n(k)$ de los óptimos centros de la muestra tiene que ser elegido tal que $\Phi(A_n, P_n) = m_k(P_n)$; el conjunto poblacional de los

centros $\bar{A} = \bar{A}(k)$ satisface $\Phi(\bar{A}, \mathcal{P}) = m_k(\mathcal{P})$. El objetivo es mostrar que $A_n \to \bar{A}$ casi seguramente.

La convergencia de conjuntos va a ser tomada de manera tal que sea determinada por la distancia de Hausdorff $H(\cdot, \cdot)$, que está definida para subconjuntos compactos A, B de $\mathbb{R}^s$ del siguiente modo: $H(A, B) < \delta$ si y sólo si todo punto de A está a una distancia menor a δ de al menos un punto de B y viceversa.

Más precisamente

$$H(A, B) = \{\text{ínf}_\epsilon : A \subset B^\epsilon \ \text{ y } \ B \subset A^\epsilon\},$$

siendo $A^\epsilon = \{x : d(x, A) < \epsilon\}$.

Supongamos que A contiene exactamente k puntos distintos y que δ es elegido menor que la mitad de la mínima distancia entre puntos de A. Entonces si B es un conjunto de k o menos puntos del cual $H(A, B) < \delta$ tiene que contener exactamente k puntos distintos los cuales están a una distancia menor a δ de algún punto de A.

Observación 1 *Si* $\int \|x\|^2 \mathcal{P}(dx) < \infty$ *entonces* $\Phi(A, \mathcal{P}) < \infty$ *para cada* A.

En efecto, como $\Phi(A, \mathcal{P}) := \int \min_{a \in A} \|x - a\|^2 \mathcal{P}(dx)$ y para cada $a \in \mathbb{R}^s$:

$$\begin{aligned}
\int \|x - a\|^2 \mathcal{P}(dx) &\leq \int (\|x\| + \|a\|)^2 \mathcal{P}(dx) \\
&\leq \int_{\|x\| \leq \|a\|} (\|x\| + \|a\|)^2 \mathcal{P}(dx) + \int_{\|x\| \geq \|a\|} (\|x\| + \|a\|)^2 \mathcal{P}(dx) \\
&\leq (2\|a\|)^2 + \int_{\|x\| \geq \|a\|} (2\|x\|)^2 \mathcal{P}(dx) \\
&\leq (2\|a\|)^2 + 4 \int (\|x\|)^2 \mathcal{P}(dx).
\end{aligned}$$

El siguiente Teorema muestra la consistencia fuerte de los centros de la versión empírica a los poblacionales así como la convergencia de las funciones objetivo empíricas a la poblacional. La demostración se hace en dos etapas. En la primera se muestra que los centros óptimos muestrales $A_n = A_n(k)$ están últimamente en un compacto. La segunda etapa utiliza una ley fuerte de grandes números uniforme sobre conjuntos de k o menos puntos contenidos en el compacto hallado en la primera etapa. Esto requerirá del estudio de algunos resultados de procesos empíricos.

Teorema 1 (Consistencia) *Supongamos que* $\int \|x\|^2 \mathcal{P}(dx) < \infty$ *y que para cada* $j = 1, 2, ..., k$ *hay un único conjunto* $\bar{A}(j)$ *para el cual* $\Phi(\bar{A}(j), \mathcal{P}) = m_j(\mathcal{P})$.

Entonces $A_n \to \bar{A}(k)$ *c.s. y* $\Phi(A_n, P_n) \to m_k(\mathcal{P})$ *c.s.*

Observación 2 *La unicidad de los $\bar{A}(j)$'s implica que $m_1(\mathcal{P}) > m_2(\mathcal{P}) > \ldots > m_k(\mathcal{P})$. Pues si $m_{j-1}(\mathcal{P}) = m_j(\mathcal{P})$ para algún j, entonces $\bar{A}(j-1)$ puede ser aumentado por un punto arbitrario para dar un NO único conjunto A, de no más de j puntos distintos, para el cual $\Phi(A, \mathcal{P}) = m_j(\mathcal{P})$. Similares condiciones implican que $\bar{A}(j)$ contiene exactamente j puntos distintos. Pues si $\bar{A}(j)$ contiene $j-1$ o menos puntos distintos entonces $\bar{A}(j-1)$ (que contiene los $j-1$ o menos puntos distintos de $\bar{A}(j)$) puede ser aumentado por un punto arbitrario para dar un NO único conjunto $\bar{A}(j)$.*
Como las conclusiones del teorema están en términos de una convergencia casi segura, va a haber conjuntos ω's de medida cero en los cuales no va a converger.

Demostración. Primer etapa. Queremos encontrar una bola suficientemente grande (que no dependa de ω) de modo que si n es suficientemente grande los centros empíricos óptimos estén contenidos en ella. Este primer paso lo haremos en dos pasos.

En el primer paso encontraremos M (que no dependa de ω) de modo que si n es suficientemente grande, al menos uno de los centros empíricos óptimos esté contenido en la bola centrada en 0 de radio M, $B(M)$ (alguno de los puntos de A_n esté en $B(M)$). En el segundo paso probaremos que para n suficientemente grande, la bola $B(5M)$, centrada en el origen de radio $5M$ contiene a todos los centros empíricos óptimos ($A_n \subset B(5M)$).

Primer paso. Primer etapa.

Sea r tal que la bola K de radio r centrada en el origen tenga medida positiva $\mathcal{P}$.

Va a ser suficiente para ello que $(M-r)^2\mathcal{P}(K) > \int \|x\|^2\mathcal{P}(dx)$ para garantizarlo, lo que siempre es posible tomando M suficientemente grande, pues $\int \|x\|^2\mathcal{P}(dx) < \infty$ y la función cuadrática es estrictamente creciente para $\mathbb{R}_{>0}$.

Para los siguientes pasos requeriremos más condiciones a M.

Busquemos M tal que para n suficientemente grande al menos un punto de A_n está contenido en la bola cerrada $B(M)$.

Asumimos $\Phi(A_n, P_n) \leq \Phi(A_0, P_n)$ para cualquier conjunto A_0 que contenga como mucho k puntos (pues A_n es el conjunto de los óptimos centros de la muestra). Elegimos A_0 que contenga un único punto en el origen.

Entonces, $\Phi(A_0, P_n) = \int \|x\|^2 P_n(dx) \to \int \|x\|^2\mathcal{P}(dx)$ c.s. (Por la Ley de los Grandes Números).

Si para infinitos valores de n, ningun punto de A_n, está contenido en $B(M)$, entonces

$$\lim_{n\to\infty} \sup \Phi(A_n, P_n) \;\geq\; \lim_{n\to\infty} (M-r)^2 P_n(K) = (M-r)^2\mathcal{P}(K) \;\; c.s.,$$

pues,

$$\begin{aligned}\Phi(A_n, P_n) &:= \int \min_{a\in A_n} \|x-a\|^2 P_n(dx) \\ &\geq \int_K \min_{a\in A_n} \|x-a\|^2 P_n(dx) \\ &\geq \inf_{a\in A_n,\ x\in K} \|x-a\|^2 P_n(K) \\ &\geq (M-r)^2 P_n(K) \qquad \forall n;\end{aligned}$$

y $(M-r)^2\mathcal{P}(K) > \int \|x\|^2\mathcal{P}(dx)$, $\lim_{n\to\infty} \Phi(A_0, P_n) = \int \|x\|^2\mathcal{P}(dx)$ c.s.

Entonces $\Phi(A_n, P_n) > \Phi(A_0, P_n)$ para infinitos valores. Esto indica que llegamos a una contradicción.

Luego sin pérdida de generalidad vamos a asumir que A_n siempre contiene como mínimo un punto en $B(M)$.

Segundo paso. Primer etapa.

- Si $k = 1$ como A_n contiene un solo punto entonces por lo anteriormente demostrado la bola cerrada $B(M)$ contiene todos los puntos.
- Si $k > 1$ veamos que para n suficientemente grande la bola cerrada $B(5M)$ contiene todos los puntos de A_n.

Para demostrarlo vamos a usar un método inductivo. Supongamos que el teorema es válido para $1, 2, \ldots, k-1$ centros de clusters.

Si A_n no está contenido en $B(5M)$, los centros de los clusters que están fuera de la $B(5M)$ pueden ser eliminados, obteniendo $k-1$ (o menos) centros, para los cuales $\Phi(\cdot, P_n)$ estaría por debajo de su mínimo sobre todos los conjuntos de $k-1$ puntos. Lo cual sería una contradicción.

En efecto, bastaría con elegir M tal que

$$4\int_{\|x\|\geq 2M} \|x\|^2\mathcal{P}(dx) < \varepsilon,$$

donde $\varepsilon > 0$ es elegido para satisfacer $\varepsilon + m_k(\mathcal{P}) < m_{k-1}(\mathcal{P})$. Este es el segundo requisito pedido de M.

Supongamos que A_n contiene al menos un punto fuera de $B(5M)$. ¿Cual sería el efecto sobre $\Phi(A_n, P_n)$ de eliminar esos puntos como centros de clusters? En el peor de los casos el centro a_1 que era conocido por pertenecer a la $B(M)$ va a tener que aceptar a su cluster todos los puntos que fueron asignados a clusters cuyos centros se

encontraban fuera de la $B(5M)$. Estos puntos tienen que haber estado como mínimo a una distancia $2M$ del origen, sino hubiesen estado más cerca de a_1 que de cualquier otro centro ubicado fuera de la $B(5M)$. La contribución extra de $\Phi(\cdot, P_n)$ al borrar los centros que se encuentran fuera de la $B(5M)$ es como mucho

$$\begin{aligned}\int_{\|x\|\geq 2M} \|x - a_1\| P_n(dx) &\leq \int_{\|x\|\geq 2M} (\|x\| + \|a_1\|)^2 P_n(dx) \\ &\leq \int_{\|x\|\geq 2M} (2\|x\|)^2 P_n(dx) \\ &\leq 4\int_{\|x\|\geq 2M} \|x\|^2 P_n(dx).\end{aligned}$$

El conjunto A_n^* obtenido al eliminar de A_n todos los centros que se encontraban fuera de la $B(5M)$ es un candidato para minimizar $\Phi(\cdot, P_n)$ en los conjuntos de $k-1$ o menos centros. Sea B_n el óptimo con $k-1$ centros.

Entonces $\Phi(A_n^*, P_n) \geq \Phi(B_n, P_n) \to m_{k-1}(\mathcal{P})$ c.s. (por la hipótesis inductiva).

Si $A_n \nsubseteq B(5M)$ en una subsecuencia $\{n_l\}$ de valores de n, entonces

$$\begin{aligned}m_{k-1}(\mathcal{P}) &\leq \lim_{i\to\infty} \text{ínf}\, \Phi(A_{n_i}^*, P_n) \text{ c.s.} \\ &\leq \lim_{n\to\infty} \sup[\Phi(A_n, P_n) + 4\int_{\|x\|\geq 2M} \|x\|^2 P_n(dx)],\end{aligned}$$

como $\Phi(A_n, P_n) = m_k(P_n) := \text{ínf}\{\Phi(A, P_n) : A \text{ contiene } k \text{ o menos puntos}\}$, luego, $\Phi(A_n, P_n) \leq \Phi(A, P_n)$ para cualquier conjunto con k o menos puntos y, por la LGN, tenemos que $4\int_{\|x\|\geq 2M} \|x\|^2 P_n(dx) \to 4\int_{\|x\|\geq 2M} \|x\|^2 \mathcal{P}(dx)$ c.s. Por lo tanto,

$$\begin{aligned}m_{k-1}(\mathcal{P}) &\leq \lim_{n\to\infty} \sup[\Phi(A_n, P_n) + 4\int_{\|x\|\geq 2M} \|x\|^2 P_n(dx)] \\ &\leq \lim_{n\to\infty} \sup \Phi(A, P_n) + 4\int_{\|x\|\geq 2M} \|x\|^2 \mathcal{P}(dx) c.s.,\end{aligned}$$

para cualquier conjunto A con k o menos puntos.

Elegimos $A = \bar{A}(k)$, el conjunto óptimo de k centros de $\Phi(\cdot, P_n)$, entonces,

$$\begin{aligned}m_{k-1}(\mathcal{P}) &\leq \lim_{n\to\infty} \sup \Phi(A, P_n) + 4\int_{\|x\|\geq 2M} \|x\|^2 \mathcal{P}(dx) \\ &< \Phi(\bar{A}(k), \mathcal{P}) + \varepsilon = m_k(\mathcal{P}) + \varepsilon < m_{k-1}(\mathcal{P}),\end{aligned}$$

entonces: $m_{k-1}(\mathcal{P}) < m_{k-1}(\mathcal{P})$ lo cual es una contradicción.

Segunda etapa.

Ahora sabemos que para n suficientemente grande es suficiente chequear A_n en la clase de los conjuntos

$$\mathfrak{E}_k := \{A \subseteq B(5M) : A \text{ contiene } k \text{ o menos puntos}\}.$$

Como requisito final de M, asumimos que M es suficientemente grande como para asegurar que $\mathfrak{E}_k$ contiene a $\bar{A}(k)$.

Luego la función $\Phi(\cdot, \mathcal{P})$ alcanza su único mínimo en $\mathfrak{E}_k$ en $\bar{A}(k)$. Con la métrica de Hausdorff $\mathfrak{E}_k$ es un compacto (esto se sigue de la compacidad de la $B(5M)$) y la asignación $A \to \Phi(A, \mathcal{P})$ es continua en $\mathfrak{E}_k$ (lo probaremos más adelante). Por tanto, se cumple que para cualquier entorno $\mathfrak{N}$ de $\bar{A}(k)$ $\exists$ un $\eta > 0$ que depende de $\mathfrak{N}$, tal que,

$$\Phi(A, \mathcal{P}) \geq \Phi(\bar{A}(k), \mathcal{P}) + \eta \qquad \forall A \in \mathfrak{E}_k \backslash \mathfrak{N}.$$

La prueba puede ahora ser completada probando una ley de los grandes números uniforme

$$\sup_{A \in \mathfrak{E}_k} |\Phi(A, P_n) - \Phi(A, \mathcal{P})| \to 0 \quad c.s.,$$

(este resultado lo probaremos más adelante).

Necesitamos mostrar que A_n está últimamente en el entorno $\mathfrak{N}$. Para ello, es suficiente chequear que $\Phi(A_n, \mathcal{P}) < \Phi(\bar{A}(k), \mathcal{P}) + \eta$ últimamente.

Sabemos que

$$\Phi(A_n, P_n) \leq \Phi(\bar{A}(k), P_n).$$

Y por la ley de los grandes números uniforme tenemos que

$$\sup_{A \in \mathfrak{E}_k} |\Phi(A, P_n) - \Phi(A, \mathcal{P})| \to 0 \quad c.s.,$$

y $A_n \in \mathfrak{E}_k$, entonces tomando $A = A_n$,

$$\Phi(A_n, P_n) - \Phi(A_n, \mathcal{P}) \to 0 \text{ c.s.}$$

Y

$$\Phi(\bar{A}(k), P_n) - \Phi(\bar{A}(k), \mathcal{P}) \to 0 \; c.s.$$

por la ley fuerte de los grandes números.

Luego,

$$\Phi(A_n, \mathcal{P}) < \Phi(\bar{A}(k), \mathcal{P}) + \eta,$$

que es lo mismo que $\Phi(A_n, \mathcal{P}) < m_k(\mathcal{P}) + \eta$.

Por lo tanto $A_n \to \bar{A}(k)$ c.s. con la métrica de Hausdorff.

Veamos ahora que,

$$\Phi(A_n, P_n) = \text{ínf}\{\Phi(A, P_n) : A \in \mathfrak{E}_k\} \to \text{ínf}\{\Phi(A, \mathcal{P}) : A \in \mathfrak{E}_k\} = m_k(\mathcal{P}) c.s.$$

Sabemos que

$$\Phi(A_n, P_n) \leq \Phi(\bar{A}(k), P_n) \to \Phi(\bar{A}(k), \mathcal{P}) = m_k(\mathcal{P}) c.s.$$

por la ley fuerte de los grandes números.

Y por la ley de los grandes números uniforme tenemos que

$$\sup_{A \in \mathfrak{E}_k} |\Phi(A, P_n) - \Phi(A, \mathcal{P})| \to 0 \quad c.s.,$$

y $A_n \in \mathfrak{E}_k$, entonces tomando $A = A_n$,

$$|\Phi(A_n, P_n) - \Phi(A_n, \mathcal{P})| \to 0 \text{ c.s.}$$

Entonces, $\forall \varepsilon > 0$ $\Phi(A_n, P_n) + \varepsilon \geq \Phi(A_n, \mathcal{P})$ para n suficientemente grande c.s.

Y $\Phi(A_n, \mathcal{P}) \geq \Phi(\bar{A}(k), \mathcal{P}) = m_k(\mathcal{P})$.

Luego, $\forall \varepsilon > 0$ $m_k(\mathcal{P}) \leq \Phi(A_n, P_n) + \varepsilon \leq m_k(\mathcal{P}) + 2\varepsilon$ para n suficientemente grande c.s.

Por lo tanto queda demostrado que

$$\Phi(A_n, P_n) = \text{ínf}\{\Phi(A, P_n) : A \in \mathfrak{E}_k\} \to \text{ínf}\{\Phi(A, \mathcal{P}) : A \in \mathfrak{E}_k\} = m_k(\mathcal{P}) c.s.$$

■

Resta probar la ley de los grandes números uniformes y la continuidad de $\Phi(\cdot, \mathcal{P})$.

Demostración.

Veamos que

$$\sup_{A \in \mathfrak{E}_k} |\Phi(A, P_n) - \Phi(A, \mathcal{P})| \to 0 \quad c.s.$$

Sea G la familia de funciones $\mathcal{P}$-integrables en $\mathbb{R}^s$ de la forma $g_A(x) := \min_{a \in A} \phi(\|x - a\|)$, donde A es cualquier subconjunto de $\mathfrak{E}_k$ que contiene k o menos puntos. Y $\phi(r) := r^2$ si $r \geq 0$ y $\phi(r) := 0$ si $r < 0$.

Necesitamos probar una ley de los grandes números uniforme:

$$\sup_{g \in G} \left| \int g dP_n - \int g d\mathcal{P} \right| \to 0 \quad c.s.$$

Una condición suficiente es: para cada $\varepsilon > 0$ existe una clase finita G_ε de funciones tales que para cada $g \in G$, existen funciones $\overset{\circ}{g}$, $\bar{g} \in G_\varepsilon$ tales que

$$\overset{\circ}{g} \leq g \leq \bar{g} \quad \text{y} \quad \int (\bar{g} - \overset{\circ}{g}) d\mathcal{P} < \varepsilon.$$

La demostración de este resultado se encuentra en el Apéndice.

Para encontrar un apropiado G_ε, sea D_δ un subconjunto finito de $B(5M)$ tal que cada punto de $B(5M)$ tiene al menos un punto en D_δ a distancia $\leq \delta$.

Sea $\mathfrak{E}_{k,\delta} := \{A \in \mathfrak{E}_k,\ A \subseteq D_\delta\}$.

Sea G_ε la clase de funciones de la forma

$$\min_{a \in A'} \phi(\|x - a\| + \delta)$$

o

$$\min_{a \in A'} \phi(\|x - a\| - \delta),$$

donde $A' \in \mathfrak{E}_{k,\delta}$y $\phi(r) := r^2$ si $r \geq 0$, $\phi(r) := 0$ si $r < 0$.

Dado $A = \{a_1, a_2, ..., a_k\} \in \mathfrak{E}_k$, existe $A' = \{a'_1, a'_2, ..., a'_k\} \in \mathfrak{E}_{k,\delta}$ con $H(A, A') < \delta$ (simplemente eligiendo $a'_i \in D_\delta$ con $\|a_i - a'_i\| < \delta\ \forall i$).

Para $g_A \in G$, tomamos:

$$\bar{g}_A := \min_{a \in A'} \phi(\|x - a\| + \delta)$$

y

$$\overset{\circ}{g}_A := \min_{a \in A'} \phi(\|x - a\| - \delta),$$

donde $\phi(r) := r^2$ si $r \geq 0$ y $\phi(r) := 0$ si $r < 0$.

Como ϕ es monótona creciente y

$$\|x - a'_i\| - \delta \leq \|x - a_i\| \leq \|x - a'_i\| + \delta \qquad \forall i, \forall x \in \mathbb{R}^s,$$

entonces

$$\overset{\circ}{g}_A \leq g_A \leq \bar{g}_A.$$

Luego si R es más grande que $5M + \delta$

$$\begin{aligned}
\int \bar{g}_A(x) - \mathring{g}_A(x)\mathcal{P}(dx) &\leq \int \sum_{i=1}^{k} [\phi(\|x - a_i'\| + \delta) - \phi(\|x - a_i'\| - \delta)]\mathcal{P}(dx) \\
&= \int_{\|x\|\leq R} \sum_{i=1}^{k} [\phi(\|x - a_i'\| + \delta) - \phi(\|x - a_i'\| - \delta)]\mathcal{P}(dx) \\
&\quad + \int_{\|x\|\geq R} \sum_{i=1}^{k} [\phi(\|x - a_i'\| + \delta) - \phi(\|x - a_i'\| - \delta)]\mathcal{P}(dx) \\
&\leq k \sup_{\|x\|\leq R} \sup_{a\in B(5M)} [\phi(\|x - a\| + \delta) - \phi(\|x - a\| - \delta)] \\
&\quad + \int_{\|x\|\geq R} \sum_{i=1}^{k} [\phi(\|x\| + \|a_i'\| + \delta) - \phi(\|x\| + \|a_i'\| - \delta)]\mathcal{P}(dx) \\
&\leq k \sup_{\|x\|\leq R} \sup_{a\in B(5M)} [\phi(\|x - a\| + \delta) - \phi(\|x - a\| - \delta)] \\
&\quad + k \int_{\|x\|\geq R} 8\phi(\|x\|)\mathcal{P}(dx),
\end{aligned}$$

donde en la última desigualdad usamos que $\|a_i'\| + \delta \leq R \leq \|x\|$ pues $\|a_i'\| \leq 5M \leq R - \delta$ entonces $\phi(\|x\| + \|a_i'\| + \delta) \leq \phi(2\|x\|) \leq 4\phi(\|x\|)$ y $\|a_i'\| - \delta \leq R \leq \|x\|$, por lo tanto, $\phi(\|x\| + \|a_i'\| - \delta) \leq 4\phi(\|x\|)$.

El segundo término es menor a $\varepsilon/2$ si elegimos R suficientemente grande. Luego por la continuidad uniforme de ϕ sobre conjuntos acotados, podemos encontrar $\delta > 0$ lo suficientemente chico para que el primer término sea menor a $\varepsilon/2$.

Un argumento similar puede ser utilizado para probar la continuidad de la asignación $A \to \Phi(A, \mathcal{P})$ en $\mathfrak{E}_k$.

Si A, $B \in \mathfrak{E}_k$ y $H(A, B) < \delta$ entonces para todo $b \in B$ existe un punto $a(b) \in A$ tal que $\|b - a(b)\| < \delta$.

Entonces

$$\begin{aligned}
\Phi(A, \mathcal{P}) - \Phi(B, \mathcal{P}) &= \int \min_{a\in A} \|x - a\|^2 - \min_{b\in B} \|x - b\|^2 \mathcal{P}(dx) \\
&\leq \int \max_{b\in B} [\ \|x - a(b)\|^2 - \|x - b\|^2\ ]\mathcal{P}(dx) \\
&\leq \int \sum_{b\in B} [(\|x - b\| + \delta)^2 - \|x - b\|^2]\mathcal{P}(dx),
\end{aligned}$$

que es menor que ε si δ es elegido como antes.

La otra desigualdad necesaria para probar la continuidad se obtiene intercambiando los roles de A y B. ■

4. El método combinatorio

Teorema 2 (Glivenko-Cantelli) *El Teorema de Glivenko-Cantelli (o Teorema Fundamental de la Estadística) afirma que para toda distribución $\mathcal{P}$ en $\mathbb{R}$,*

$$\sup_t |P_n(-\infty, t] - \mathcal{P}(-\infty, t]| \to 0 \text{ casi seguramente,}$$

cuando la medida empírica P_n proviene de una muestra independiente de $\mathcal{P}$.

En lo que sigue denotaremos por $\|\cdot\|$ al supremo sobre la clase $\mathfrak{F}$ de intervalos $(-\infty, t]$, para $-\infty < t < \infty$. Vamos a poder restringir al supremo sobre los racionales para poder asegurar medibilidad.

Para facilitar la lectura dividiremos la demostración del Teorema de Glivenko-Cantelli en 5 etapas.

Demostración.

Primera etapa: Primera simetrización

En vez de comparar a P_n con su distribución $\mathcal{P}$, miremos la diferencia entre P_n y una copia independiente de él mismo que llamaremos P'_n. La diferencia $P_n - P'_n$ está determinada por un conjunto de $2n$ puntos (tomados al azar) de la linea real.

Lema 3 (simetrización) *Sean $\{Z(t) : t \in T\}$ y $\{Z'(t) : t \in T\}$ procesos estocásticos independientes que comparten un conjunto T. Supongamos que existen constantes $\alpha > 0$ y $\beta > 0$ tales que $P\{|Z'(t)| \leq \alpha\} \geq \beta$ para todo t en T. Entonces:*

$$P\left\{\sup_t |Z(t)| > \varepsilon\right\} \leq \beta^{-1} P\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}.$$

Demostración. Sea τ una variable aleatoria para la cual $|Z(\tau)| > \varepsilon$ en el conjunto $\{\sup |Z(t)| > \varepsilon\}$. Como τ es determinado por Z, es independiente de Z'. Se comporta como un índice de valores fijos si condicionamos con Z:

$$P\{|Z'(\tau)| \leq \alpha | Z\} \geq \beta.$$

Entonces:

$$\begin{aligned}
\beta P\left\{\sup_t |Z(t)| > \varepsilon\right\} &\leq P\{|Z'(\tau)| \leq \alpha | Z\} P\left\{\sup_t |Z(t)| > \varepsilon\right\} \\
&\leq P\{|Z'(\tau)| \leq \alpha | Z\} P\{|Z(\tau)| > \varepsilon\} \\
&\leq P\{|Z'(\tau)| \leq \alpha, |Z(\tau)| > \varepsilon\} \\
&\leq P\{|Z(\tau) - Z'(\tau)| > \varepsilon - \alpha\} \\
&\leq P\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}.
\end{aligned}$$

■

Si se observa con detalle la demostración se puede ver que ciertas sutilezas teóricas fueron ignoradas. Pero, para nuestro objetivo es suficiente asumir que T es numerable y para esta clase de conjuntos la demostración está bien hecha. Vamos a reemplazar supremo sobre todos los intervalos $(-\infty, t)$ por supremo sobre los intervalos con t racional.

Para un t fijo, $P_n(-\infty, t]$ es un promedio de n variables aleatorias independientes de $I_{\{\xi_i \leq t\}}$, cada una con valor esperado de $\mathcal{P}(-\infty, t]$, y varianza $\mathcal{P}(-\infty, t] - (\mathcal{P}(-\infty, t])^2$, que es menor a uno. Utilizando la desigualdad de Tchebychev:

$$\begin{aligned} P\left\{|P'_n(-\infty, t] - \mathcal{P}(-\infty, t]| \leq \frac{1}{2}\varepsilon\right\} &\geq 1 - \frac{(\mathcal{P}(-\infty, t] - (\mathcal{P}(-\infty, t])^2)/n}{\left(\frac{1}{2}\varepsilon\right)^2} \\ &= 1 - \frac{4}{\varepsilon^2}\frac{(\mathcal{P}(-\infty, t] - (\mathcal{P}(-\infty, t])^2)}{n} \\ &\geq \frac{1}{2} \text{ si } n \geq 8\varepsilon^{-2}. \end{aligned}$$

Aplicando el lema de simetrización para $Z = P_n - \mathcal{P}$ y $Z' = P'_n - \mathcal{P}$, con $\mathfrak{F}$ un conjunto fijo, $\alpha = \frac{1}{2}\varepsilon$, y $\beta = \frac{1}{2}$.

$$\begin{aligned} P\left\{\|P_n - \mathcal{P}\| > \varepsilon\right\} &\leq 2P\left\{\|P_n - \mathcal{P} - (P'_n - \mathcal{P})\| > \varepsilon - \frac{1}{2}\varepsilon\right\} \\ &= 2P\left\{\|P_n - P'_n\| > \frac{1}{2}\varepsilon\right\} \text{ si } n \geq 8\varepsilon^{-2}. \end{aligned}$$

Segunda etapa: Segunda simetrización

La diferencia $P_n - P'_n$ depende de $2n$ observaciones. El tamaño doble de la muestra crea una molestia **menor** al menos en la notación. Puede ser evitada con un truco de un segunda simetrización, con el costo de una disminución de ε. Independientemente de las observaciones $\xi_1, \ldots, \xi_n, \xi'_1, \ldots, \xi'_n$ para las cuales la medida empírica fue construída, generamos variables aleatorias independientes $\sigma_1, \ldots, \sigma_n$ para las cuales $P\{\sigma_i = 1\} = P\{\sigma_i = -1\} = 1/2$. Las variables aleatorias simétricas $I_{\{\xi_i \leq t\}} - I_{\left\{\xi'_i \leq t\right\}}$, para $i = 1, \ldots, n$ y $-\infty < t < \infty$, tienen la misma distribución conjunta que las variables aleatorias $\sigma_i\left[I_{\{\xi_i \leq t\}} - I_{\left\{\xi'_i \leq t\right\}}\right]$.

Entonces:

$$\begin{aligned} P\left\{\|P_n - P_n'\| > \frac{1}{2}\varepsilon\right\} &= P\left\{\sup_t \left| n^{-1}\sum_{i=1}^n \sigma_i \left[I_{\{\xi_i \le t\}} - I_{\left\{\xi_i' \le t\right\}}\right]\right| > \frac{1}{2}\varepsilon\right\} \\ &\le P\left\{\sup_t \left| n^{-1}\sum_{i=1}^n \sigma_i I_{\{\xi_i \le t\}}\right| > \frac{1}{4}\varepsilon\right\} \\ &+ P\left\{\sup_t \left| n^{-1}\sum_{i=1}^n \sigma_i I_{\left\{\xi_i' \le t\right\}}\right| > \frac{1}{4}\varepsilon\right\}. \end{aligned}$$

Escribamos P_n^0 para la medida que le da peso $n^{-1}\sigma_i$ a $I_{\{\xi_i \le t\}}$.
Entonces

$$\begin{aligned} P\left\{\|P_n - \mathcal{P}\| > \varepsilon\right\} &\le 2P\left\{\|P_n - P_n'\| > \frac{1}{2}\varepsilon\right\} \\ &\le 4P\left\{\|P_n^0\| > \frac{1}{4}\varepsilon\right\}. \end{aligned} \tag{5}$$

Para controlar el término de la derecha, le vamos a agregar condiciones a las observaciones.

Tercera etapa: Desigualdad Maximal

Una vez que la localización de las obervaciones ξ está fija, el supremo de $\|P_n^0\|$ se reduce al máximo tomado sobre un conjunto estratégico de intervalos $I_j = (-\infty, t_j]$, para $j = 0, 1, \ldots, n$. Por supuesto que la elección de los intervalos depende de las observaciones ξ; necesitamos un t_j entre cada par de observaciones adyacentes (t_0 y t_n no son realmente necesarios). Con el número de intervalos reducidos tan drásticamente podemos plantear una cota drástica para el supremo.

$$\begin{aligned} P\left\{\|P_n^0\| > \frac{1}{4}\varepsilon|\xi\right\} &\le \sum_{j=0}^n P\left\{\left|P_n^0 I_j\right| > \frac{1}{4}\varepsilon|\xi\right\} \\ &\le (n+1)\max_j P\left\{\left|P_n^0 I_j\right| > \frac{1}{4}\varepsilon|\xi\right\}. \end{aligned} \tag{6}$$

Esta cota va a ser adecuada para el presente caso, porque las probabilidades condicionales disminuyen exponencialmente con n; gracias a la desigualdad de Hoeffding para sumas de variables aleatorias independientes y acotadas.

Cuarta etapa: Cota exponencial

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes, cada una con media cero y un rango acotado: $a_i \leq Y_i \leq b_i$. Para cada $\eta > 0$, la desigualdad de Hoeffding dice:

$$P\left\{|Y_1 + \ldots + Y_n| \geq \eta\right\} \leq 2\exp\left[-2\eta^2 / \sum_{i=1}^{n}(b_i - a_i)^2\right].$$

Esta desigualdad la demostraremos en el apéndice.

Apliquemos la desigualdad para $Y_i = \sigma_i I_{\{\xi_i \leq t\}}$. Dado ξ la variable aleatoria Y_i toma solamente los valores $\pm I_{\{\xi_i \leq t\}}$, cada uno con probabilidad $\frac{1}{2}$. Luego $-1 \leq Y_i \leq 1$

$$\begin{aligned} P\left\{\left|P_n^0(-\infty, t]\right| \geq \frac{1}{4}\varepsilon|\xi\right\} &= P\left\{\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{\xi_i \leq t\}}\right| \geq \frac{1}{4}\varepsilon|\xi\right\} \\ &= P\left\{\left|\sum_{i=1}^{n}\sigma_i I_{\{\xi_i \leq t\}}\right| \geq \frac{n\varepsilon}{4}|\xi\right\} \\ &\leq 2\exp\left[-2\left(\frac{n\varepsilon}{4}\right)^2 / \sum_{i=1}^{n}(1-(-1))^2\right] \\ &\leq 2\exp(-n\varepsilon^2/32), \end{aligned}$$

Usemos esto para cada I_j en la desigualdad (6).

$$P\left\{\|P_n^0\| > \frac{1}{4}\varepsilon|\xi\right\} \leq 2(n+1)\exp(-n\varepsilon^2/32). \tag{7}$$

Notemos que el lado derecho ahora no depende de ξ.

Quinta etapa: Integración

Luego tomando esperanza respecto a ξ y usando las dos desigualdades (5) y (7), obtenemos:

$$P\left\{\|P_n - \mathcal{P}\| > \varepsilon\right\} \leq 8(n+1)\exp(-n\varepsilon^2/32).$$

Esto da una muy rápida convergencia en probabilidad, tan rápida que

$$\sum_{n=1}^{\infty} P\left\{\|P_n - \mathcal{P}\| > \varepsilon\right\} < \infty,$$

para cada $\varepsilon > 0$. Luego por el lema de Borel-Cantelli la convergencia es casi segura, quedando demostrado el teorema de Glivenko-Cantelli.

■

5. Clases de conjuntos con discriminación polinomial

Para probar el Teorema de Glivenko-Cantelli, hemos utilizado muy pocas propiedades de los intervalos. El principal requerimiento para la ecuación (6) era que se puede elegir como mucho $n+1$ subconjuntos de cualquier conjunto de n puntos. El resto de la demostración vale para otras familias más generales de conjuntos con ligeras modificaciones. Esto motiva cosiderar otras familias de conjuntos para los cuales se pueda obtener una cota polinomial en (6) que nos llevaría a cambiar en (7) $(n+1)$ por un polinomio para el caul aún vale el argumento que sigue a (7). Por ejemplo, cuadrantes de la forma $(-\infty, t] \times (-\infty, s]$ en $\mathbb{R}^2$ pueden elegir como mucho $(n+1)^2$ diferentes subconjuntos de un conjunto de n puntos en el plano (hay como mucho $n+1$ lugares donde ubicar el límite vertical y $n+1$ lugares donde ubicar el límite horizontal). Reemplazando $n+1$ por $(n+1)^2$ podemos repetir los argumentos que utilizamos al demostrar el teorema de Glivenko-Cantelli. La cota exponencial controla a $(n+1)^2$ así como controlaba a $n+1$. Es más, controlaría a cualquier polinomio. Los argumentos sirven para intervalos, cuadrantes y cualquier otra clase de conjuntos para los cuales se pueda elegir una cantidad polinomial de subconjuntos.

Definición 4 *Sea $\mathfrak{D}$ la clase de subconjuntos de cierto espacio S. Se dice que tiene discriminación polinomial (de grado υ) si existe un polinomio $\rho(\cdot)$ (de grado υ) tal que, para cualquier conjunto de N puntos en S, se puede elegir como mucho $\rho(N)$ subconjuntos distintos. Formalmente, si S_0 contiene N puntos, entonces hay como mucho $\rho(N)$ conjuntos distintos de la forma $S_0 \cap D$ con D en $\mathfrak{D}$. LLamemos $\rho(\cdot)$ a la discriminación polinomial de $\mathfrak{D}$.*

Para acortar el nombre llamemos a "la clase que tiene discriminación polinomial", "clase polinomial" y adoptemos la misma terminología para polinomios de menor grado. Por ejemplo, los intervalos de la linea real tienen discriminación lineal (los llamaremos clase lineal) y los cuadrantes en el plano tienen discriminación cuadrática (los llamaremos clase cuadrática). Por supuesto hay clases que no tienen discriminación polinomial. Por ejemplo, para toda colección de N puntos ubicados en la circunferencia de un círculo de $\mathbb{R}^2$ la clase de conjuntos cerrados convexos puede elegir 2^N subconjuntos, y 2^N crece mucho más rápido que cualquier polinomio.

El método de la demostración anterior se puede aplicar a cualquier clase polinomial de conjuntos, pero pueden surgir problemas de medibilidad. Las clases que no tiene estos problemas son llamadas permisibles. Cualquier clase específica que encontremos va a ser permisible. Asique, ignoremos estos problemas a partir de ahora, pero utilicemos el término permisible para recordar que cierta condición de regularidad es necesaria.

Teorema 5 *Sea $\mathcal{P}$ una medida de probabilidad de un espacio S. Para toda clase permisible $\mathfrak{D}$ de subconjuntos de S con discriminación polinomial,*

$$\sup_{\mathfrak{D}} |P_n D - \mathcal{P} D| \to 0 \ cs.$$

Demostración. La demostración es igual a la de Glivenko-Cantelli, cambiando $\mathfrak{F}$ por $\mathfrak{D}$ y cambiando $n+1$ por el polinomio que corresponda a $\mathfrak{D}$, y tachando las referencias a intervalos y a la linea real. ■

¿Qué clases tiene discriminación polinomial? Ya sabemos de intervalos y cuadrantes; equivalentes con dimensiones mayores tienen las mismas propiedades. Otras clases pueden ser construídas a partir de estas.

Lema 6 *Si $\mathfrak{C}$ y $\mathfrak{D}$ tienen discriminación polinomial, entonces también:*
(i) $\{D^c : D \in \mathfrak{D}\}$;
(ii) $\{C \cup D : C \in \mathfrak{C} \ y \ D \in \mathfrak{D}\}$;
(iii) $\{C \cap D : C \in \mathfrak{C} \ y \ D \in \mathfrak{D}\}$.

Demostración. Sean $c(\cdot)$ y $d(\cdot)$ las discriminaciones polinomiales. Podemos suponer que ambas son funciones crecientes de N.

De un conjunto S_0 que contiene N puntos, supongamos que $\mathfrak{D}$ elige los subconjuntos $S'_1, \ldots, S'_l$ con $l \leq d(N)$. Entonces la clase de (i) elige los subconjuntos $S'^c_1, \ldots, S'^c_l$. Como $l \leq d(N)$ esto da la cota $d(N)$ para el tamaño de la clase de (i). Esto prueba (i).

De un conjunto S_0 que contiene N puntos, supongamos que $\mathfrak{C}$ elige los subconjuntos $S_1, \ldots, S_k$ con $k \leq c(N)$. Y $\mathfrak{D}$ elige los subconjuntos $S'_1, \ldots, S'_l$ con $l \leq d(N)$. Entonces la clase de (ii) elige los subconjuntos $S_1 \cup S'_1, \ldots, S_1 \cup S'_l, \ldots, S_k \cup S'_1, \ldots, S_k \cup S'_l$, esto da la cota $c(N)d(N)$. Esto prueba (ii) .

De un conjunto S_0 que contiene N puntos, supongamos que $\mathfrak{C}$ elige los subconjuntos $S_1, \ldots, S_k$ con $k \leq c(N)$. Supongamos que S_i contiene N_i puntos. La clase $\mathfrak{D}$ elige como mucho $d(N_i)$ distintos subconjuntos de S_i. Esto da la cota $d(N_1) + \ldots + d(N_k)$ para el tamaño de la clase de (iii). La suma es menor que $c(N)d(N)$. Esto prueba (iii). ■

El lema puede ser aplicado repetidamente para generar cada vez más clases polinomiales. Igual deberíamos de ponerle un límite al número de operaciones permitidas. Pues, por ejemplo, la clase de todos los conjuntos de un solo elemento tiene discriminación lineal, pero con una cantidad arbitraria de uniones finitas podemos elegir cualquier conjunto finito.

Usando este lema muy rápidamente construímos clases de poco interés. Afortunadamente hay otros sistemáticos métodos para encontrar clases polinomiales.

Los polinomios crecen mucho más lento que las funciones exponenciales. Para N suficientemente grande, una clase polinomial tiene que no contener como mínimo uno de los 2^N subconjuntos de cada colección de N puntos. Sorprendentemente esto caracteriza a la dicriminación polinomial. Una clase $\mathfrak{D}$ se dice que fragmenta un conjunto de puntos F si puede seleccionar cualquier subconjunto posible (incluyendo el conjunto vacio y el

conjunto F) ; esto quiere decir, $\mathfrak{D}$ fragmenta F si cualquier subconjunto de F tiene la forma $D \cap F$ para algún D en $\mathfrak{D}$.

Por ejemplo, la clase de los discos cerrados de $\mathbb{R}^2$ puede romper cualquier conjunto de 3 puntos siempre que no estén alineados. Pero para los conjuntos de 4 puntos, no importa cual sea su configuración sólo va a poder elegir 15 de los posibles 16 subconjuntos. El disco fragmenta algunos conjuntos de tres puntos pero no fragmenta de 4 puntos.

Teorema 7 *Sea S_0 un conjunto de N puntos en S. Supongamos que existe un entero $V \leq N$ tal que $\mathfrak{D}$ no fragmenta ningun conjunto de V puntos de S_0. Entonces $\mathfrak{D}$ no puede elegir más de $\binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{V-1}$ subconjuntos de S_0.*

Demostración. Sean $F_1, \ldots, F_k$ la colección de todos los subconjuntos de V elementos de S_0. Es claro que $k = \binom{N}{V}$. Por supuesto, cada F_i tiene un subconjunto "escondido" H_i que $\mathfrak{D}$ pasa por alto: $D \cap F_i \neq H_i$ para todo D en $\mathfrak{D}$. Esto quiere decir que todos los conjuntos de la forma $D \cap S_0$, con D en $\mathfrak{D}$, pertenecen a

$$\mathcal{C}_0 = \{C \subseteq S_0 : C \cap F_i \neq H_i \text{ para cada } i\}.$$

Va a ser suficiente encontrar una cota superior para el tamaño de $\mathcal{C}_0$.

En un caso especial va a ser posible contar el número de conjuntos de $\mathcal{C}_0$ directamente. Si $H_i = F_i$ para cada i, entonces ningun C en $\mathcal{C}_0$ puede contener un F_i; es decir, ningún C puede contener un conjunto de V puntos. En otras palabras, miembros de $\mathcal{C}_0$ consisten en $0, 1, \ldots$,o $V-1$ puntos. Entonces la suma de los coeficientes binomiales da el número de conjuntos de esta forma.

Jugando con los conjuntos escondidos podemos reducir el caso general al caso especial recién tratado. Etiquetemos los puntos de S_0 como $1, \ldots, N$. Para cada i definamos $H_i' = (H_i \cup \{1\}) \cap F_i$; esto quiere decir, aumentamos H_i con el punto 1, siempre y cuando esto pueda ser hecho sin violar la restricción de que el conjunto escondido este contenido en F_i. Definimos la correspondiente clase:

$$\mathcal{C}_1 = \{C \subseteq S_0 : C \cap F_i \neq H_i' \text{ para cada } i\}.$$

La clase $\mathcal{C}_1$ no tiene nada más que ver con la clase $\mathcal{C}_0$. La única conección es que todos los conjuntos ocultos, que el conjunto pasa por alto, son más grandes. Veamos que esto implica que $\mathcal{C}_1$ tiene un cardinal más grande que $\mathcal{C}_0$. (Observamos que no estamos diciendo que $\mathcal{C}_0 \subset \mathcal{C}_1$).

Chequeemos que la asignación $C \longrightarrow C \setminus \{1\}$ es una a una de $\mathcal{C}_0 \setminus \mathcal{C}_1$ a $\mathcal{C}_1 \setminus \mathcal{C}_0$. Comencemos con algún C en $\mathcal{C}_0 \setminus \mathcal{C}_1$. Por definición, $C \cap F_i \neq H_i$ para todo i, pero $C \cap F_j = H_j'$ para al menos un j. Luego $H_j \neq H_j'$, por lo que 1 pertenece a C, a F_j y a H_j', pero no a H_j. ¿Por qué debería $C \setminus \{1\}$ pertenecer a $\mathcal{C}_1 \setminus \mathcal{C}_0$? Observemos que

$$(C \setminus \{1\}) \cap F_j = H_j' \setminus \{1\} = H_j,$$

por lo tanto $C\setminus\{1\}$ no pertenece a $\mathcal{C}_0$. Luego, si F_i contiene al 1, H_i' también lo tiene que contener, pero $C\setminus\{1\}$ lógicamente no puede, por lo tanto si F_i no contiene al 1 entonces

$$(C\setminus\{1\})\cap F_i = C\cap F_i \neq H_i = H_i'.$$

En ambos casos $(C\setminus\{1\})\cap F_i \neq H_i'$, luego, $C\setminus\{1\}$ pertenece a $\mathcal{C}_1$. Luego el cardinal de $\mathcal{C}_1$ es más grande que el cardinal de $\mathcal{C}_0$.

Repitiendo el proceso, empezando con los nuevos conjuntos escondidos y con 2 en lugar de 1, definimos $H_i'' = (H_i'\cup\{2\})\cap F_i$ y $\mathcal{C}_2 = \{C\subseteq S_0 : C\cap F_i \neq H_i''$ para cada $i\}$. El cardinal de $\mathcal{C}_2$ es más grande que el cardinal de $\mathcal{C}_1$. Otras $N-2$ repeticiones generarían las clases $\mathcal{C}_3, \mathcal{C}_4, \ldots, \mathcal{C}_N$ cuyos cardinales van creciendo. Y además los conjuntos escondidos de $\mathcal{C}_N$ serían todos los F_i; y estaríamos en el caso especial. ■

Corolario 8 *Si una clase no fragmenta ningún conjunto de V puntos, entonces tiene discriminación polinomial de grado menor o igual a $V-1$.*

Todo lo que nos falta ahora es un buen método para identificar las clases que tienen problemas para elegir subconjuntos de conjuntos de puntos suficientemente grandes.

Lema 9 *Sea $\mathcal{G}$ un espacio vectorial de dimensión finita de funciones reales de S. La clase de conjuntos de la forma $\{g\geq 0\}$, con g en $\mathcal{G}$, tiene discriminación polinomial de grado menor o igual a la dimensión de $\mathcal{G}$.*

Demostración. Sea $V-1$ la dimensión de $\mathcal{G}$. Elijamos una colección $\{s_1,\ldots,s_V\}$ de puntos distintos de S. (Todo es trivial si S contiene V o menos puntos.) Definamos una asignación lineal L de $\mathcal{G}$ en $\mathbb{R}^V$ por

$$L(g) = (g(s_1),\ldots,g(s_V)).$$

Como $L\mathcal{G}$ es un subespacio lineal de $\mathbb{R}^V$ de dimensión como mucho $V-1$, existe en $\mathbb{R}^V$ un vector no nulo γ ortogonal a $L\mathcal{G}$. Esto quiere decir que

$$\sum_i \gamma_i g(s_i) = 0 \text{ para cada } g \text{ en } \mathcal{G},$$

o

$$\sum_{\{+\}} \gamma_i g(s_i) = \sum_{\{-\}} \gamma_i g(s_i) \text{ para cada } g. \tag{8}$$

Aca $\{+\}$ se refiere a los conjuntos de los i para los cuales $\gamma_i\geq 0$, y $\{-\}$ para los cuales $\gamma_i\leq 0$. Reemplazando γ por $-\gamma$ si es necesario, podemos asumir que $\{-\}$ no es vacio.

Supongamos que existe una g para la cual $\{g\geq 0\}$ elige precisamente aquellos puntos s_i con i en $\{+\}$. Para estos g, la parte izquierda de (8) va a ser ≥ 0 pero la parte derecha es <0. Luego encontramos un conjunto que no puede ser elegido. ■

Muchas clases familiares de objetos geométricos caen dentro del lema. Por ejemplo, la clase de subconjuntos del plano generados por la forma cuadrática $ax^2 + bxy + cy^2 + dx + ey + f$ incluye todos los discos cerrados, elipsoides, y (como caso degenerado) los semiespacios. Regiones más complicadas pueden ser construídas usando el lema. Esto se puede aplicar en los teoremas, obteniendo así generalizaciones del teorema clásico de Glivenko-Cantelli.

6. Apéndice

6.1. Demostración de una ley de los grandes números uniforme

En esta sección $\mathfrak{F}$ va a ser la clase de funciones medibles en un conjunto S con una $\sigma-$álgebra que tiene medida de probabilidad $\mathcal{P}$. La medida empírica P_n es construída a partir de un muestreo de $\mathcal{P}$. Asumimos $\mathcal{P}|f| < \infty$ para cada f en $\mathfrak{F}$. Si $\mathfrak{F}$ fuese finito, la convergencia de $P_n f$ a $\mathcal{P}f$ asegurada por la ley fuerte de los grandes números sería uniforme en f. Si $\mathfrak{F}$ puede ser aproximado por una clase finita (no necesariamente una subclase de $\mathfrak{F}$) de modo tal que los errores por la aproximación sean uniformemente pequeños, la uniformidad se trasladaría a $\mathfrak{F}$. El método directo logra esto requiriendo que cada elemento de $\mathfrak{F}$ esté rodeado por un par de funciones tomadas de la clase finita.

Teorema 10 *Si para cada $\varepsilon > 0$ existe una clase finita $\mathfrak{F}_\varepsilon$ que contenga una inferior y una superior aproximación para cada f en $\mathfrak{F}$, para el cual*

$$f_{\varepsilon,L} \leq f \leq f_{\varepsilon,U} \text{ y } \mathcal{P}(f_{\varepsilon,U} - f_{\varepsilon,L}) < \varepsilon$$

Entonces $\sup_{\mathfrak{F}} |P_n f - \mathcal{P}f| \to 0$ *casi seguramente.*

Demostración. Para probar la convergencia, basta demostrar :

$$\lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n f - \mathcal{P}f) \geq 0$$

y

$$\lim \sup \sup_{\mathfrak{F}} (P_n f - \mathcal{P}f) \leq 0,$$

o, equivalentemente,

$$\lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n(-f) - \mathcal{P}(-f)) \geq 0.$$

Primero veamos que

$$\lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n f - \mathcal{P}f) \geq 0.$$

Para todo $\varepsilon > 0$,

$$\begin{aligned}
\lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n f - \mathcal{P}f) &\geq \lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n f_{\varepsilon,L} - \mathcal{P}f) \text{ pues } f_{\varepsilon,L} \leq f \\
&\geq \lim \text{ínf} \left(\text{ínf}_{\mathfrak{F}} (P_n f_{\varepsilon,L} - \mathcal{P}f_{\varepsilon,L}) + \text{ínf}_{\mathfrak{F}} (\mathcal{P}f_{\varepsilon,L} - \mathcal{P}f) \right) \\
&\geq 0 + (-\varepsilon) \text{ pues } \mathcal{P}f_{\varepsilon,L} > \mathcal{P}f_{\varepsilon,U} - \varepsilon \geq \mathcal{P}f - \varepsilon
\end{aligned}$$

casi seguramente cuando $\mathfrak{F}_\varepsilon$ es finito.
Y

$$\lim \text{ínf} \, \text{ínf}_{\mathfrak{F}} (P_n(-f) - \mathcal{P}(-f)) \geq 0,$$

pues para todo $\varepsilon > 0$,

$$\begin{aligned}
\lim\inf \inf_{\mathfrak{F}}(P_n(-f) - \mathcal{P}(-f)) &\geq \lim\inf \inf_{\mathfrak{F}}(P_n(-f_{\varepsilon,U}) - \mathcal{P}(-f)) \\
&\geq \lim\inf \inf_{\mathfrak{F}}(P_n(-f_{\varepsilon,U}) - \mathcal{P}(-f_{\varepsilon,U})) \\
&\quad + \inf_{\mathfrak{F}}(\mathcal{P}(-f_{\varepsilon,U}) - \mathcal{P}(-f)) \\
&\geq 0 + (-\varepsilon), \quad \text{c.s. cuando } \mathfrak{F}_\varepsilon \text{ es finito.}
\end{aligned}$$

Donde en la primer desigualdad hemos usado que $-f_{\varepsilon,U} \leq -f$ y en la última desigualdad hemos usado que $\mathcal{P}(-f_{\varepsilon,U}) > \mathcal{P}(-f_{\varepsilon,L}) - \varepsilon \geq \mathcal{P}(-f) - \varepsilon$. ■

Se puede notar que la independencia es sólo utilizada para garantizar la convergencia casi segura de $P_n f_\varepsilon$ a $\mathcal{P} f_\varepsilon$ para cada aproximación f_ε. Hipótesis más débiles como estacionariedad y ergodicidad podrian remplazar la independencia.

6.2. Demostración de una desigualdad exponencial

Probemos la siguinte desigualdad:

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes, cada una con media cero y un rango acotado: $a_i \leq Y_i \leq b_i$. Para cada $\eta > 0$:

$$P\left\{|Y_1 + \ldots + Y_n| \geq \eta\right\} \leq 2\exp\left[-2\eta^2 / \sum_{i=1}^{n}(b_i - a_i)^2\right]. \tag{9}$$

Para eso primero probemos el siguiente Lema:

Lema 11 *Desigualdad de Hoeffding:*

Sea Y una variable aleatoria con media cero, $a \leq Y \leq b$. Entonces para $s > 0$,

$$E\left[e^{sY}\right] \leq e^{s^2(b-a)^2/8}. \tag{10}$$

Demostración. Por convexidad

$$e^{tY} \leq e^{ta}\frac{(b-Y)}{(b-a)} + e^{tb}\frac{(Y-a)}{(b-a)},$$

pues $\frac{(b-Y)}{(b-a)} + \frac{(Y-a)}{(b-a)} = 1$.

Como $E(Y) = 0$,

$$E(e^{tY}) \leq e^{ta}\frac{b}{(b-a)} - e^{tb}\frac{a}{(b-a)},$$

sea $\alpha = 1 - \beta = \frac{-a}{(b-a)}$ y $u = t(b-a)$, entonces $\beta = 1 + \frac{a}{(b-a)} = \frac{b}{(b-a)}$, y

$$ta = \frac{u}{(b-a)}(-\alpha)(b-a) = -\alpha u, \quad tb = \frac{u}{(b-a)}\beta(b-a) = \beta u.$$

Notemos que $\alpha > 0$ porque $a < 0 < b$ pues media de Y es cero.

$$\begin{aligned} \log E(e^{tY}) &\leq \log\left(\beta e^{-\alpha u} + \alpha e^{\beta u}\right) \\ &= \log\left(e^{-\alpha u}\left(\beta + \alpha e^{(\beta+\alpha)u}\right)\right) \text{ (obs: } \beta + \alpha = 1) \\ &= -\alpha u + \log\left(\beta + \alpha e^{u}\right). \end{aligned}$$

Sea $L(u) = -\alpha u + \log\left(\beta + \alpha e^{u}\right)$. Luego,

$$L'(u) = -\alpha + \frac{1}{\beta + \alpha e^{u}}\alpha e^{-u} = -\alpha + \frac{\alpha}{\alpha + \beta e^{-u}}.$$

$$L''(u) = \frac{-\alpha}{\left(\alpha + \beta e^{-u}\right)^2}\beta e^{-u}(-1) = \left[\frac{\alpha}{\alpha + \beta e^{-u}}\right]\left[\frac{\beta e^{-u}}{\alpha + \beta e^{-u}}\right].$$

Como $x(1-x) \leq \frac{1}{4}$ para todo $0 \leq x \leq 1$, sea $x = \frac{\alpha}{\alpha + \beta e^{-u}}$, entonces

$$1 - x = \frac{\alpha + \beta e^{-u} - \alpha}{\alpha + \beta e^{-u}} = \frac{\beta e^{-u}}{\alpha + \beta e^{-u}} = \beta\frac{e^{-u}}{\alpha + \beta e^{-u}}, \ 0 \leq x \leq 1.$$

Por lo tanto, $L''(u) \leq \frac{1}{4}$. Por el teorema de Taylor

$$\begin{aligned} L(u) &= L(0) + uL'(0) + \frac{1}{2}u^2 L''(u^*) \\ &\leq \log(\beta + \alpha) + u\left(-\alpha + \frac{\alpha}{\alpha + \beta}\right) + \frac{1}{2}u^2\frac{1}{4} \\ &= \frac{1}{8}u^2 \\ &= \frac{1}{8}t^2(b-a)^2. \end{aligned}$$

Entonces

$$\log E(e^{tY}) \leq \frac{1}{8}t^2(b-a)^2,$$

luego

$$E(e^{tY}) \leq e^{t^2(b-a)^2/8}.$$

■

Demostración. de (9). La desigualdad de Markov dice que para toda variable aleatoria Y no negativa, y $\eta > 0$,

$$P(Y \geq \eta) \leq \frac{E(Y)}{\eta}.$$

Se sigue de la desigualdad de Markov que si ϕ es estrictamente monótona creciente para los valores no negativos, entonces para cualquier variable aleatoria Y y cualquier número real η,

$$P(Y \geq \eta) = P(\phi(Y) \geq \phi(\eta)) \leq \frac{E(\phi(Y))}{\phi(\eta)}.$$

Una aplicación de esto con $\phi(x) = x^2$ es la desigualdad de Chebyshev's: si Y es una variable aleatoria arbitraria y $t > 0$, entonces

$$\begin{aligned} P\left\{|Y - E(Y)| \geq \eta\right\} &= P\left\{|Y - E(Y)|^2 \geq \eta^2\right\} \\ &\leq \frac{E\left[|Y - E(Y)|^2\right]}{\eta^2} \\ &= \frac{Var(Y)}{\eta^2}. \end{aligned}$$

Una generalización es tomando $\phi(x) = x^q$ $(x \geq 0)$, para cualquier $q > 0$ tenemos

$$P\left\{|Y - E(Y)| \geq \eta\right\} \leq \frac{E\left[|Y - E(Y)|^q\right]}{\eta^q}.$$

En casos específico uno elige el valor de q para minimizar la cota. Una idea relacionada es la base del método de la cota de Chernoff. Tomando $\phi(x) = e^{sY}$ donde s es un número positivo arbitrario, para cualquier variable aleatoria Y, para cualquier $\eta > 0$, tenemos

$$P\left\{Y \geq \eta\right\} = P\left\{e^{sY} \geq e^{s\eta}\right\} \leq \frac{E\left[e^{sY}\right]}{e^{s\eta}}.$$

En el método de Chernoff's, encontramos un $s > 0$ que minimiza la cota superior o hace la cota superior pequeña.

Recordemos algunas simples desigualdades de variables aleatorias independientes. Estamos principalmente interesados en cotas superiores para las probabilidades de las desviaciones de la media, esto es, para obtener desigualdades de $P\left\{S_n - E\left(S_n\right) \geq \eta\right\}$, con $S_n = \sum_{i=1}^{n} Y_i$, donde $Y_1, \ldots, Y_n$ son variables aleatorias independientes evaluadas en los reales.

La desigualdad de Chebychev e independencia implican

$$P\left\{|S_n - E(S_n)| \geq \eta\right\} \leq \frac{Var\left\{S_n\right\}}{\eta^2} = \frac{\sum_{i=1}^{n} Var\left\{Y_i\right\}}{\eta^2}.$$

El método de la cota de Chernoff's es especialmente conveniente para cotas de probabilidad de sumas de variables aleatorias independientes. Esto se debe a que los valores esperados de un producto de variables aleatorias independientes es igual al producto de los valores esperados, los límites Chernoff serían

$$\begin{aligned} P\left\{S_n - E(S_n) \geq \eta\right\} &\leq e^{-s\eta} E\left[\exp\left(s\sum_{i=1}^{n}\left(Y_i - E(Y_i)\right)\right)\right] \\ &= e^{-s\eta}\prod_{i=1}^{n} E\left[e^{s(Y_i - E(Y_i))}\right] \text{ (bajo independencia).} \end{aligned}$$

Aplicando (10) y usando que en nuetro caso $E(Y_i) = 0$ (luego también la $E(S_n) = 0$) tenemos

$$\begin{aligned} P\left\{S_n \geq \eta\right\} &\leq e^{-s\eta}\prod e^{s^2(b_i-a_i)^2/8} \\ &= \exp\left[-s\eta + \frac{1}{8}s^2\sum_{i=1}^{n}(b_i - a_i)^2\right]. \end{aligned}$$

Sea $s = 4\eta/\sum_{i=1}^{n}\left(b_i - a_i\right)^2$ para minimizar la función cuadrática entonces

$$\begin{aligned} P\left\{S_n \geq \eta\right\} &\leq \exp\left[-\frac{4\eta^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2} + \frac{2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right] \\ &= \exp\left[-\frac{2\eta^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right]. \end{aligned}$$

Aplicando el mismo argumento a $-Y_1, \ldots, -Y_n$ obtenemos

$$P\left\{S_n \leq -\eta\right\} \leq \exp\left[-\frac{2\eta^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right].$$

Luego

$$P\left\{|Y_1 + \ldots + Y_n| \geq \eta\right\} \leq 2\exp\left[-2\eta^2/\sum_{i=1}^{n}(b_i - a_i)^2\right],$$

que es lo que queríamos probar. ■

Referencias

[1] Boucheron, S., Lugosi, G. and Bousquet O. (2004), "Concentration Inequalities", in O. Bousquet, U.v. Luxburg, and G. Rätsch (editors), *Advanced Lectures in Machine Learning*, Springer, pp. 208–240. **Vol 9 No 1**, 135–140.

[2] Hastie, T., Tibshirani, R., Friedman, J.(2001), " The Elements of Statistical Learning". *Springer Series in Statistics*.

[3] Pollard, D. (1981), "Strong Consistency of K-Means Clustering". *The Annals of Statistics*, **Vol 9 No 1**, 135–140.

[4] Pollard, D.(1984), "Convergence of Stochastic Processes". *Springer–Verlag, New York*.