Práctica: Introducción al Aprendizaje Estadístico

1 Conceptual

1. Determine en cada situación si se trata de un problema de regresión o de clasificación. Especifique la cantidad de datos (n) y la cantidad de predictores (p).

   (a) Se juntan los datos de 500 empresas. De cada empresa registramos el número de empleados, la facturación y la ganancia de la empresa, y el sueldo del CEO. Queremos determinar qué factores son los determinanes del sueldo del CEO.

   (b) Estamos por lanzar un producto y queremos saber si será un *éxito* o un *fracaso*. Para tal fin, juntamos los datos de otros 20 productos ya lanzados de los cuales contamos con información acerca de si fue o no un éxito, el precio del producto, el presupuesto de marketing, el precio del producto más similar lanzado por la competencia y 10 otras variables.

   (c) Queremos predecir el porcentaje de cambio del valor del dólar norteamericano en término de la variación porcentual de diversos mercados mundiales. Juntamos todos los datos de una semana particular del 2012 del cambio diario de la bolsa alemana, la inglesa y la norteamericana.

   (d) Queremos tener una idea acerca del precio de una casa en términos de sus metros cuadrados cubiertos, los no cubiertos y tres variables más vinculadas a aspectos de calidad de la construcción.

   (e) Ídem anterior pero ahora cada casa se clasifica en *Barata*, *Moderada* y *Cara*.

2. (a) Piense o busque tres problemas que encajen en un problema de *regresión*. Identifique las variables predictoras y de respuesta.

   (b) Piense o busque tres problemas que encajen en un problema de *clasificación*. Describa la respuesta así como los predictores.

3. Se tiene Y variable respuesta y X variable predictora y se obtienen n muestras (x_i, y_i). Obtenga, para los siguientes modelos, el estimador de cuadrados mínimos:

   (a) $Y \approx \beta X$. En qué situaciones se puede obtener el estimador de cuadrados mínimos?

   (b) $Y \approx \beta_0 + \beta_1 X^2$.

2 Aplicado: regresión lineal simple

1. Se tiene en el archivo **autos.txt** una serie de modelos de autos clasificados según su precio y una clasificación de calidad.

   (a) Levante los datos del archivo.

   (b) Grafique en un diagrama de dispersión precios vs calidad.

   (c) Bajo los siguientes modelos, obtenga el estimador de mínimos cuadrados:

      i. $Calidad \approx \beta_0$.
      ii. $Calidad \approx \beta_1 Precio$.
      iii. $Calidad \approx \beta_0 + \beta_1 Precio$.

   (d) Bajo los tres modelos anteriores, estime la calidad de un auto de valor $50000.

   (e) Repita lo de antes pero separando un 30% de datos al azar que conformarán un **testing set**. Es decir, esos datos no se usarán para construir la regresión, mientras que el 70% restante (el **training set**) son los que se usarán para estimar. Fijado ambos conjuntos, construya las tres regresiones anteriores y determine el error cuadrático de predicción sobre el **testing set**. Es decir, sobre los $N_testing$ datos que conforman el **testing set**, calcule:

$$RSE_{testing} = \frac{1}{N_{testing}} \sum_{i=1}^{N_{testing}} (Calidad - \widehat{Calidad}_i)^2.$$

   Determine así el "mejor" modelo.

2. En el siguiente ejercicio, ensayaremos con simulación y regresión simultáneamente. Para empezar, escoja un valor para la semilla de aleatoriedad. Luego:

   (a) Con el comando **rnorm**, genere 100 observaciones provenientes de una distribución normal estándar. Guarde esto en la variable x.

   (b) Con el mismo comando, genere 100 observaciones provenientes de una $N(0, \sigma^2 = 0.25)$, guarde esto en la variable *eps*.

   (c) Usando x y *eps*, construya la variable dependiente y según el modelo $y = -1+0.5x+eps$. Quiénes serían los valores de β_0 y β_1 en este modelo?

   (d) Efectúe un *scatter plot* entre las variables x e y.

   (e) Efectúe un ajuste de cuadrados mínimos para obtener $\hat{\beta}_0$ y $\hat{\beta}_1$, comparelos con los valores verdaderos β_0 y β_1.

   (f) Grafique en un mismo plot el *scatter plot*, la "verdadera" recta de y vs x y la recta obtenida por cuadrados mínimos, en distinto color.

   (g) Repita el procedimiento de antes, pero ahora habrá aproximadamente un 10% de los datos contaminados. Para eso, ahora el modelo que usaremos es el siguiente, $y = V(-1+0.5x+eps)+(1-V)W$ donde $V \sim Bi(1, 1/10)$ (es decir, vale 1 con probabilidad 1/10) y $W \sim N(\mu = 50, \sigma^2 = 1)$.

   (h) Experimento con distintos niveles de contaminación, más allá del 10%.

3. Se tiene en el archivo **girasol.txt** el rinde de diversas parcelas de girasol (en toneladas) según la cantidad de dinero invertida en fertilizantes (en miles de pesos).

   (a) Levante los datos del archivo.

   (b) Grafique en un diagrama de dispersión inversion vs rinde.

   (c) Bajo un modelo de regresión lineal simple obtenga el estimador de mínimos cuadrados:

   (d) Grafique la recta de regresión obtenida, detecta algo sospechoso?

   (e) Efectúe una "limpieza" de los datos y repita el procedimiento.

4. Considere el archivo **abalone.txt** que contiene información sobre distintas muestras de abalones. Los atributos están separados por coma, con los siguientes campos:

   - Sexo (categórica): M (masculino), F (femenino) o I (infante).
   - Longitud (continua), en milímetros.
   - Diámetro (continua), en milímetros.
   - Altura (continua), en milímetros.
   - Peso completo del abalone (continua), en gramos.
   - Peso de la carne (continua), en gramos.
   - Peso de las vísceras (continua), en gramos.
   - Peso del caparazón (continua), en gramos.
   - Anillos (entera).

   (a) Efectúe una regresión lineal simple por cuadrados mínimos para obtener el diámetro en función de la longitud usando todos los datos.

   (b) Para tener una idea de calidad, haga como en el ejercicio 1) separando un training y un testing set.