Hoja de Ejercicios 1

Aprendizaje Estadístico y Toma de Decisiones

Juan José Martín, Marina Moreno, Christian Strasser, Maria del Mar Bibiloni.

MADM, Curs 2016-17

1. **Busca un ejemplo de regresión donde la variable dependiente es cuantitativa. Elabora. ¿Qué es Y? ¿Que es X? ¿Que es $f(X)$? ¿Como estimaras $f(X)$ por K nearest neighbours? ¿Como se interpretara esta estimación $f(X)$ en este caso? ¿Como lo estimaras con un modelo paramétrico? ¿Como se interpretara esta estimación $\tilde{f}(X)$ en este caso?**

   Queremos encontrar la relación existente entre la altura de un individuo y la longitud de sus pies. Es decir, suponemos que la altura de un individuo depende del tamaño del pie, por tanto, el problema es encontrar un modelo que describa esta dependencia. Así, denominamos la variable dependiente como Y="La altura del individuo (en cms)" y la variable independiente como X="La longitud del pie (en cms)".

   Supongamos que existe un modelo que describle la dependencia de Y respecto de X, es decir, para cada valor de X, podemos saber que altura debería tener el individuo. La función del modelo es la que denominamos por $f(X)$. Así, $f(X)$ es la función que nos da información del **valor verdadero** de la altura que le corresponde a un valor X de la longitud del pie. Debemos añadir que, aunque conociéramos exactamente cuál es la función del modelo, cometeríamos errores de predicción, pues para cada valor de la variable X hay una distribución de posibles valores de Y. En definitiva podemos decir que $Y = f(X)+\varepsilon$.

   Para estimar $f(x)$ por K-Nearest Neighbours utilizaríamos la siguiente función:

$$\widehat{f}(x) = \frac{1}{K} \sum_{x_i \in N_K(x)} y_{x_i},$$

   dónde $N_K(x)$ es el conjunto de los K vecinos más cercanos a x y y_{x_i} es el valor de la altura que corresponde al valor de la longitud de pie x_i (dónde el par (x_i, y_{x_i}) pertenece al conjunto de entrenamiento).

Tal y como hemos construido la estimación, podemos interpretar que $\widehat{f}(x)$ es la media local de las alturas cercanas al valor x dado.

Por otro lado, supongamos que la realidad se puede explicar con un modelo lineal, en este caso podríamos estimar $f(x)$ con el siguiente modelo paramétrico:

$$\widehat{f}(x) = \beta_0 + \beta_1 x.$$

Con el método de mínimos cuadrados estimaríamos el valor de β_0 y β_1.

Podríamos interpretar que la relación existente entre X y la estimación de la altura correspondiente es tal que, si la lontitud del pie aumenta en un centímetro, la altura del individuo aumenta en β_1, ya que: 3

$$\widehat{f}(x+1) - \widehat{f}(x) = \beta_0 + \beta_1(x+1) - \beta_0 + \beta_1 x = \beta_1.$$

2. **¿Puedes construir datos de Y y X tal que el error irreducible = 0 y otros tal que es muy alto? ¿Cómo lo representarías en un gráfico? ¿Qué implicaría para la estimación?**

El error irreducible es el error que nunca variará a pesar de utilizar modelos diferentes. Esto se debe a que en la vida real siempre hay cierto error que viene dado por los propios datos, y no por el modelo.

Por lo tanto, dada la función

$$Y = f(X) + \epsilon$$

$f(X)$ es la función por la que, para cada $X = x$, hay una serie de posible valores de Y, y ϵ es el error irreducible.

Para poder construir datos de Y y X donde el error irreducible es igual a 0, sería necesario que, para cada valor de X, Y variara de igual forma siempre, es decir, que ningún factor aleatorio afecte al valor de Y. Por tanto, $Y = f(X)$.

Por ejemplo, para

$$Y = \beta_0 + \beta_1 X = 0 + 1X$$

donde $\beta_0 = 0$ y $\beta_1 = 1$, tendríamos datos tales que

X	1	2	3	4	5
Y	1	2	3	4	5

En la Figura 1 tenemos los datos representados en un gráfico. Notemos que los puntos están alineados, por lo que un modelo lineal sería un modelo perfecto, sin error irreducible.

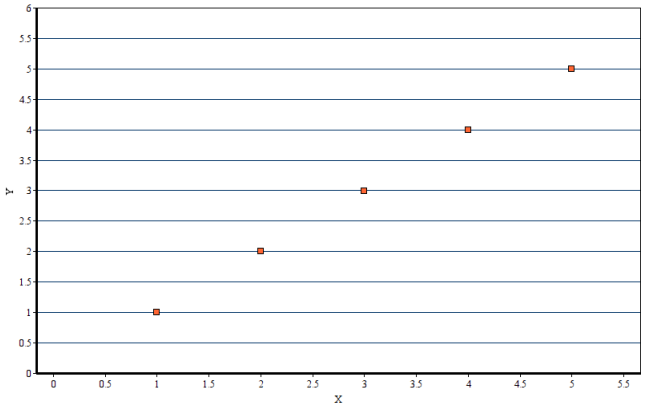

Figura 1: Valores de X e Y con error irreducible igual a 0.

Para representar datos donde el error irreducible es muy alto, sería todo lo contrario al anterior caso. Los valores que podría tener Y deberían tener una varianza significativa, lo que implicaría que, por muy ajustado que sea nuestro modelo (error reducible bajo), siempre tendríamos un error alto en nuestra predicción, ya que la alta variabilidad en los dados dificultaria una buena aproximación para una nueva predicción de y.

Por ejemplo,

X	1	2	1	4	5	1	4	3	1	5	2	4	3	5	2
Y	8	2	7	9	1	5	3	2	2	5	1	4	1	1	6

En los datos anteriores se puede observar que hay una gran variabilidad en los datos, ya que, por ejemplo, para un mismo valor de $x = 1$ se corresponden diferentes valores de $y = 8, 7, 2$, en consecuencia la varianza del error irreducible será muy grande. En este caso, los datos estan representados en la Figura 2.

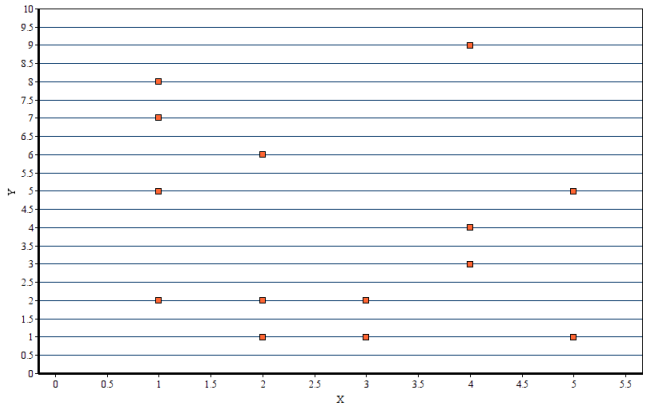

Figura 2: Valores de X e Y con error irreducible alto.

3. **Busca un ejemplo de un problema de clasificación donde la variable dependiente toma 2 valores. Elabora. ¿Qué es Y? ¿Qué es X? ¿Qué es $p(X)$? ¿Cómo estimarías $p(X)$ y $C(X)$ en este caso con el método de $K-$ Nearest Neighbours?**

Nos ponemos en el papel de un banco y queremos saber si nuestros clientes son aptos para condecerles un préstamo o no en función de los ingresos medios mensuales. Por este motivo, declaramos como variable independiente $X =$"Ingreso medio mensual (en €)"y como variable dependiente $Y =$"Resolución para el préstamo (0=no apto, 1=apto)".

Si definimos $p(x)$ como la siguiente probabilidad:

$$p(x) = \Pr(Y = 1|X = x).$$

En consecuencia, nuestro clasificador debemos definirlo como:

$$C(x) = \begin{cases} 1 & p(x) \geq 0.5 \\ 0 & p(x) < 0.5 \end{cases}$$

Para estimar $p(x)$ por el método de K-Nearest Neighbours, se utiliza la siguiente función:

$$\widehat{p}(x) = \frac{1}{K} \sum_{x_i \in N_K(x)} y_{x_i},$$

dónde $N_K(x)$ es el conjunto de los K vecinos más cercanos a x y y_{x_i} es 0 o 1, según si se les ha denegado el préstamo o no a los vecinos x_i (el par (x_i, y_{x_i}) corresponde al conjunto de entrenamiento).

4. **Busca un ejemplo de un problema de clasicacion donde la variable dependiente toma mas de 2 valores. Elabora.**

Nos proponemos clasificar el status social de una família según el gasto medio mensual. Por este motivo declaramos como variable independiente: $X =$"Gasto medio mensual (en €)"y como variable dependiente declaramos $Y =$"Status social (0=precario, 1=medio, 2=acomodado)".

Para realizar la clasificación es necesario definir las siguientes probabilidades:

$$p_0(x) = \Pr(Y = 0|X = x),$$
$$p_1(x) = \Pr(Y = 1|X = x),$$
$$p_2(x) = \Pr(Y = 2|X = x).$$

En consecuencia, nuestro clasificador se define como:

$$C(x) = j, \text{ dónde } p_j(x) = \max\{p_0(x), p_1(x), p_2(x)\}.$$

Para estimar $p(x)$ por el método de K-Nearest Neighbours, se utiliza la siguiente función:

$$\widehat{p_j}(x) = \frac{1}{K} \sum_{x_i \in N_K(x)} I[y_{x_i} = j],$$

dónde $N_K(x)$ es, de nuevo, el conjunto de los K vecinos más cercanos a x y y_{x_i} es 0, 1 o 2 dependiendo de la clase social de los vecinos x_i (el par (x_i, y_{x_i}) corresponde al conjunto de entrenamiento).

5. **Un amigo quiere visitarte en Noviembre en Palma. Pregunta por las fechas óptimas y te dice que le gusta el sol. ¿En qué sentido puede verse esto como un problema de clasificación? Elabora. ¿Cómo se podría estimar el problema?**

Escoger las fechas óptimas para éste amigo se puede entender como un problema de clasificación. La idea es predecir que dias de noviembre va a hacer sol y que dias no, y por tanto, clasificar cada día del mes en *sol*, *no sol*.

En este caso, declararíamos como variable independiente X ="Día de Noviembre"y como variable dependiente Y="Hacer Sol" (0 = no sol, 1 = sí sol). La estimación del problema, por ejemplo, se podría hacer mediante las observaciones de los meses de Noviembre de años anteriores y clasificar el día según el tiempo que hizo ese mismo dia en la mayor parte de años. Consideremos la probabilidad que queremos estimar,

$$p(Dia) = \Pr(Y = 1|X = Dia).$$

De esta manera, para obtener una estimación de p podríamos utilizar una función, obtenida a partir las N observaciones de años anteriores, como la siguiente:

$$\widehat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} y_{x_i},$$

dónde x_i indica el mismo dia x del los N años anteriores y y_{x_i} toma valor 1 o 0 dependiendo si hizo sol o no el dia x del año i.

En consecuencia, nuestro clasificador debemos definirlo como:

$$C(x) = \begin{cases} 1 & \widehat{p}(x) \geq 0.5 \\ 0 & \widehat{p}(x) < 0.5 \end{cases}.$$

Con esto, teniendo las observaciones de los mismos días en años anteriores, se puede obtener una clasificación Y dado un día determinado del mes de noviembre. Así, se podrían realizar predicciones de cada día de noviembre para determinar qué días estarán soleados, y así luego decidir el mejor momento para visitar Palma.

6. **Encuentra un ejemplo real de clasificación (simple) donde el error de Bayes es máximo. Y uno donde es mínimo. ¿Qué implica para la estimación?**

En primer lugar, vamos a encontrar un ejemplo dónde el error de Bayes sea máximo. Supongamos que queremos predecir el comportamiendo de cara/cruz de una moneda equilibrada, dada una variable independiente X (cualquiera, por ejemplo, $X =$"Altura del individuo que tira la moneda (en cms)). Así, consideramos la variable dependiente $Y =$"Resultado de una moneda (0=cruz, 1=cara)".

Supongamos que elegimos siempre predecir que sale cruz.

La tasa de clasificación errónea es:

$$1 - E(\max_k \Pr(Y = k|X = x)).$$

Cuando lanzamos una moneda equilibrada tenemos la misma posibilidad de sacar cara que cruz, en consecuencia: $\Pr(Y = k|X = x) = \Pr(Y = k) = 0.5\ \forall k$, es decir, el valor valor de Y no depende de ninguna característica o variable independiente X. Por este motivo la tasa de clasificación errónea es máxima, ya que dada la clasificación que demos, siempre tendremos una tasa de clasificación de 0.5:

$$1 - E(\max_k \Pr(Y = k|X = x)) = 1 - 0.5 = 0.5$$

En segundo lugar, para encontrar un ejemplo dónde el error de Bayes sea mínimo debemos encontrar un ejemplo de sucesos que tengan probabilidad 1 de suceder, veamos por qué.

Supongamos que queremos predecir si mañana va a salir el Sol, dada una variable independiente X (cualquiera, por ejemplo, X ="La temperatura media de la tierra (en grados)". Consideramos Y ="El resultado de si mañana va a salir el sol"(1=va a salir el Sol, 0=no va a salir el Sol) la variable dependiente.

Consideremos nuestro modelo de predicción como $f(X) = 1$, es decir, nuestra predicción no depende de ninguna variable independiente. La tasa de clasificación errónea es:

$$1 - E(\max_k \Pr(Y = k|X = x))$$

Asumiendo que nos encontramos en un modelo determinista, con toda seguridad mañana va a salir el Sol, en consecuencia $\Pr(Y = 1|X = x) = \Pr(Y = 1) = 1$ y $\Pr(Y = 0|X = x) = \Pr(Y = 0) = 0$, por este motivo la tasa de clasificación errónea es mínima, ya que como siempre clasificaremos que mañana va a salir el Sol, la tasa va a ser 0:

$$1 - E(\max_k \Pr(Y = k|X = x)) = 1 - E(\max\{\Pr(Y = 0|X = x), \Pr(Y = 1|X = x)\})$$

$$= 1 - E(\max\{1, 0\}) = 1 - 1 = 0.$$

7. **En un problema de clasificación con 3 características describe cómo son las frontera de decisión de Bayes resultando de una estimación de K-Nearest Neighbours. ¿Para qué sirven? ¿Cómo dependen del K?**

Consideremos un problema de clasificación con 3 características, $X = (X_1, X_2, X_3)$, que explican el valor de la variable cualitativa Y. Una manera de representar las observaciones $x_i = (x_{i1}, x_{i2}, x_{i3})$ de una muestra es en el espacio. Así, la frontera de Bayes resulta una superficie.

En la Figura 3 se muestra un ejemplo de un problema de clasificación con 3 predictores y $Y \in \{0, 1\}$. Los puntos azules y amarillos representan aquellas características x_i que se han clasificado como 1 y 0, respectivamente. La frontera de Bayes, en este ejemplo, divide el espacio en dos regiones. Así, para predecir la clasificación de un nuevo x_0, diremos que $Pr(Y = 1|X = x_0) > 0.5$ si x_0 está por encima de la frontera de Bayes. Análogamente, la probabilidad será estrictamente menor que 0.5 si x_0 se encuentra bajo la frontera y exactamente 0.5 si está sobre ella. Por tanto, la frontera de Bayes sirve para hacer predicciones, es decir, para definir en que regiones se clasificará un x_0 como un valor

u otro. Concretamente, se pueden clasifican como 1 todos aquellos puntos por encima de la frontera y como 0 los puntos sobre y bajo la superfície.

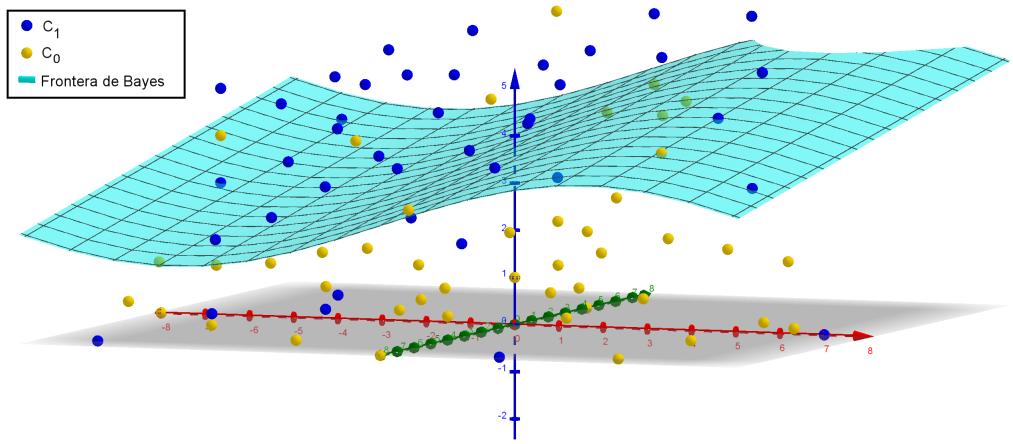

Figura 3: Ejemplo de un problema de clasificación con 3 características, Y binaria y la frontera de Bayes resultante de aplicar K-Nearest Neighbours.

La probabilidad $Pr(Y = j|X = x_0)$ se puede estimar por el método de K-Nearest Neighbours. Así, la frontera de Bayes estimada depende de K.

Sea $N_K(x_0)$ el entorno de x_0 que contiene K vecinos, contendrá demasiados vecinos para valores de K grandes. Esto puede provocar *underfitting*, es decir, que no se detecten regiones de clasificación relevantes como R en la Figura 4. De hecho, si K coincide con el número de observaciones, entonces se clasificará todo el espacio como un único valor. Por otra parte, si K es muy pequeño tenemos *overfitting*, ya que se pueden llegar a detectar varias zonas similares a R que no son significativas.

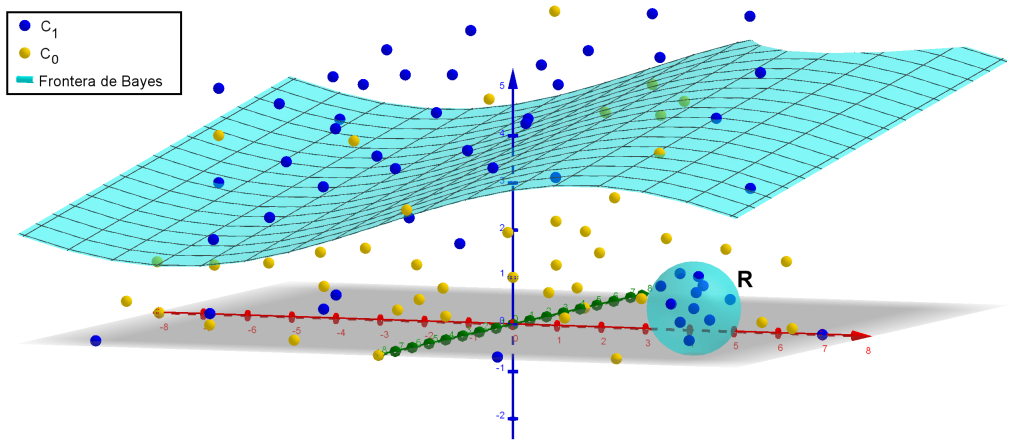

Figura 4: Ejemplo de un problema de clasificación con 3 características, Y binaria y la frontera de Bayes resultante de aplicar K-Nearest Neighbours.