Aprendizaje Automático (2014-2015)

Grado en Ingeniería Informática

Universidad de Granada

Cuestionario T2

Mª Cristina Heredia Gómez

3 de mayo de 2015

Índice

Índice de figuras

1. (1.5 puntos) Leer el apartado 4.4.2 del libro ISLR

1.1. Deducir las expresiones 4.13 y 4.14 a partir de la expresión 4.12. Justificar cada uno de los pasos dados.

En primer lugar, aplicamos el logaritmo a toda la fórmula, y luego, aplicamos las propiedades de los logaritmos. Como el logaritmo de una división es la diferencia de los logaritmos, obtenemos un miembro, al que llamo (*) , y otro miembro, al que llamo (**). Luego los desarrollamos y simplificamos por separado y por último, los restamos:

$$\log(p_k(x)) = \log\left(\frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma}\cdot e^{\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi}\sigma}\cdot e^{\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}}\right) =$$

$$= \underbrace{\log\left(\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right)}\right)}_{(*)} - \underbrace{\log\left(\sum_{l=1}^{k}\pi_l \frac{1}{\sqrt{2\pi}\sigma}\cdot e^{\left(\frac{-(x-\mu_l)^2}{2\sigma^2}\right)}\right)}_{(**)} =$$

$$(*) = \log\left(\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right)}\right) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \log\left(e^{\frac{-(x-\mu_k)^2}{2\sigma^2}}\right)$$

$$= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right) \longrightarrow \frac{-(x-\mu_k)^2}{2\sigma^2} = \frac{-(x^2+\mu_k^2-2x\mu_k)}{2\sigma^2}$$

Aplicamos binomio

↳ cambiando el signo, tenemos:

$$\frac{-x^2}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \frac{\cancel{2}x\mu_k}{\cancel{2}\sigma^2}$$

$$(**) = \log\left(\sum_{l=1}^{k}\pi_l\right) + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \left(\frac{-(x-\mu_l)^2}{2\sigma^2}\right)$$

Ser cte, por lo que lo omitimos

↳ operamos como en el apartado anterior.

$$\frac{-(x^2+\mu_l^2-2x\mu_l)}{2\sigma^2} = \frac{-x^2}{2\sigma^2} - \frac{\mu_l^2}{2\sigma^2} + \frac{\cancel{2}x\mu_l}{\cancel{2}\sigma^2}$$

↳ ctes se omiten

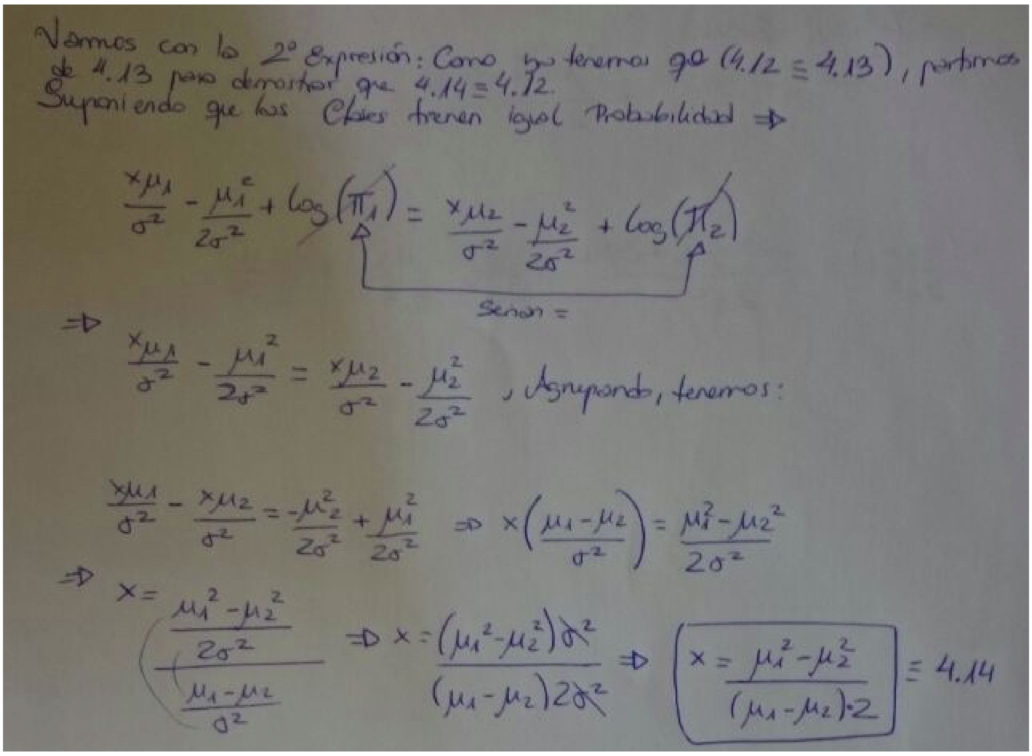
Finalmente tenemos que (*) − (**) =

$$= \left(\log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{x^2}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2}\right) - \left(\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{x^2}{2\sigma^2}\right) =$$

$$= \boxed{\log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2}}$$

Figura 1.1: deducción de la expresión 4.13 a partir de la expresión 4.12

Para deducir la 2º expresión , como ya hemos deducido la 4.13, vamos a partir de esa. Para ello, supondremos que las clases tienen igual probabilidad, obteniendo así que los logaritmos se nos van, pues serían iguales y podrían restarse. Después, simplemente queda agrupar y sacar factor común:



Vamos con la 2º Expresión: Como ya tenemos que (4.12 ≡ 4.13), partimos de 4.13 para demostrar que 4.14 = 4.12.
Suponiendo que las Clases tienen igual Probabilidad ⇒

$$\frac{x\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \cancel{\log(\pi_1)} = \frac{x\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \cancel{\log(\pi_2)}$$

Serán =

$$\Rightarrow \frac{x\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} = \frac{x\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

, Agrupando, tenemos:

$$\frac{x\mu_1}{\sigma^2} - \frac{x\mu_2}{\sigma^2} = -\frac{\mu_2^2}{2\sigma^2} + \frac{\mu_1^2}{2\sigma^2} \Rightarrow x\left(\frac{\mu_1-\mu_2}{\sigma^2}\right) = \frac{\mu_1^2-\mu_2^2}{2\sigma^2}$$

$$\Rightarrow x = \frac{\frac{\mu_1^2-\mu_2^2}{2\sigma^2}}{\frac{\mu_1-\mu_2}{\sigma^2}} \Rightarrow x = \frac{(\mu_1^2-\mu_2^2)\sigma^2}{(\mu_1-\mu_2)2\sigma^2} \Rightarrow \boxed{x = \frac{\mu_1^2-\mu_2^2}{(\mu_1-\mu_2)\cdot 2}} \equiv 4.14$$

Figura 1.2: deducción de la expresión 4.14 a partir de la expresión 4.12

1.2. Interpretar la relación que existe entre las expresiones 4.12 y 4.13 en el contexto de la clasificación de una observación.

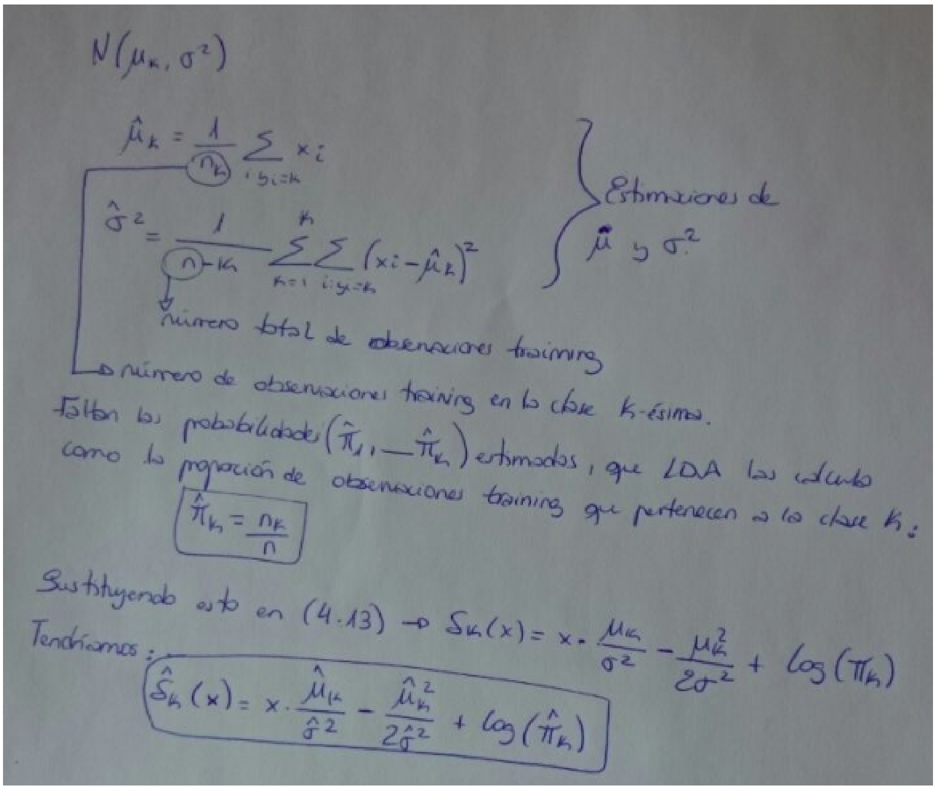
En el contexto de la clasificación de una observación, tenemos que la relación que existe entre las expresiones (4.12) y (4.13) es que (4.12) nos da la probabilidad de que un valor pertenezca a una clase k, y (4.13) nos da un valor para cada clase k que haya, por lo tanto el valor que nos de (4.13) para una clase dada, será a su vez la clase con más probabilidad de que la observación pertenezca a ella. O dicho de otra forma, clasificar una observación en la clase donde (4.12) es más grande equivale a clasificar una observación en la clase donde (4.13) es más grande.

1.3. Interpretar la relación que existe entre las expresiones 4.12 y 4.14 en el contexto de la clasificación de una observación.

En el contexto de la clasificación de una observación, tenemos que la relación que existe entre las expresiones (4.12) y (4.14) es que (4.12) nos da la probabilidad de que un valor pertenezca a una clase k, mientras que (4.14) calcula la frontera de decisión de dos clases con la misma varianza, por lo que las clases con valor más grande tienen más probabilidad. O, en otras palabras, si las observaciones en la clase k-ésima vienen dadas por una distribución N(μ_k, σ^2), el clasificador de Bayes asignará la observación a la clase donde la función discriminante o discriminativa sea máxima.

2. (2 puntos) Supongamos k clases distintas representadas por distribuciones Normales univariantes de media μ_k y varianza σ_k^2 respectivamente. Deducir la regla de Bayes asociada a asignar una nueva observación a la clase más probable. ¿Es una función lineal o cuadrática? Justificar la respuesta

N(μ_k, σ^2)
Tenemos que, a no ser que estemos muy seguros de que la observación (X) sigue una distribución Gaussiana, tendríamos que estimar los parámetros $\mu_1 ... \mu_k$, $\pi_1 ... \pi_k$ y σ^2


$N(\mu_k, \sigma^2)$

$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$

$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$

Estimaciones de $\hat{\mu}$ y σ^2

→ número total de observaciones training

→ número de observaciones training en la clase k-ésima.

Faltan las probabilidades $(\hat{\pi}_1, \ldots \hat{\pi}_k)$ estimadas, que LDA las calcula como la proporción de observaciones training que pertenecen a la clase k:

$\hat{\pi}_k = \frac{n_k}{n}$

Sustituyendo esto en (4.13) → $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$

Tendríamos: $\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$

Figura 2.1: regla de Bayes asociada a asignar una nueva observación a la clase más probable

Y tenemos que se asignará una observación (X = x) a la clase donde $\delta\hat{(x)}$ sea mayor. Este clasificador sería una función lineal, pues las funciones discriminantes de las que se compone $\delta\hat{(x)}$ son lineales.

3. (1 punto)

3.1. En promedio, ¿qué fracción de gente con un cociente de probabilidades (odds) de 0.37 de no pagar el gasto de su tarjeta de crédito, realmente no pagará? Justificar la respuesta

sabemos que log(p1/(1-p1))=$\beta_0 + \beta_1 * X$ por lo que p1/(1-p1)=exp($\beta_0 + \beta_1 * X$)
tenemos que p1=0.37 =>(0.37/ (1-0.37))=0.58 por lo que **un 58 % realmente no pagará**.

3.2. Supongamos que un individuo tiene 16 % de posibilidades de no pagar con su tarjeta de crédito. ¿Cuánto vale el cociente de probabilidades (odds) de que no page? Justificar la respuesta

sabemos que p1/(1-p1)=exp($\beta_0 + \beta_1 * X$) Nos dicen que la probabilidad de que no pague es 0.16 Por lo tanto, el cocciente de probabilidades es: p1/(1-p1)=0.16/(1-0.16)=0.19. Luego tiene un **19 % de posibilidades de que no page**.

4. (1.5 puntos) Supongamos que tenemos datos de un grupo de estudiantes de un curso de AA. Las muestras se componen de X1 =horas estudiadas, X2 =puntuación de formación del estudiante (valor de 1-4), e Y = haber obtenido un sobresaliente . Ajustamos un modelo de regresión logística y estimamos como coeficientes $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.

4.1. Estimar la probabilidad de que un estudiante que estudia 40 horas y su puntuación en formación es de 3.5, obtenga un sobresaliente .

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

Figura 4.1: función logística en regresión logística

Por lo que en nuestro caso sería:

$$P(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x+\beta_2 x_2}} = \frac{e^{-6+0.05\cdot 40+1\cdot 3'5}}{1+e^{-6+0.05\cdot 40+3'5}} = \frac{e^{-0'5}}{1+e^{-0'5}} = \boxed{0'37}$$

Figura 4.2: probabilidad de que el estudiante saque sobresaliente

Por lo que tendríamos que el alumno, estudiando 40h y teniendo una formación de 3.5 sobre 4, tendría un **37 % de posibilidades de sacar un sobresaliente en la asignatura**.

4.2. ¿Cuantas horas necesita estudiar un estudiante para tener al menos un 50 % de posibilidades de obtener un sobresaliente?

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

Figura 4.3: cocciente de probabilidades tomando logaritmos

Ahora nos centraremos en esa función, añadiendo nuestas variables y coeficientes:

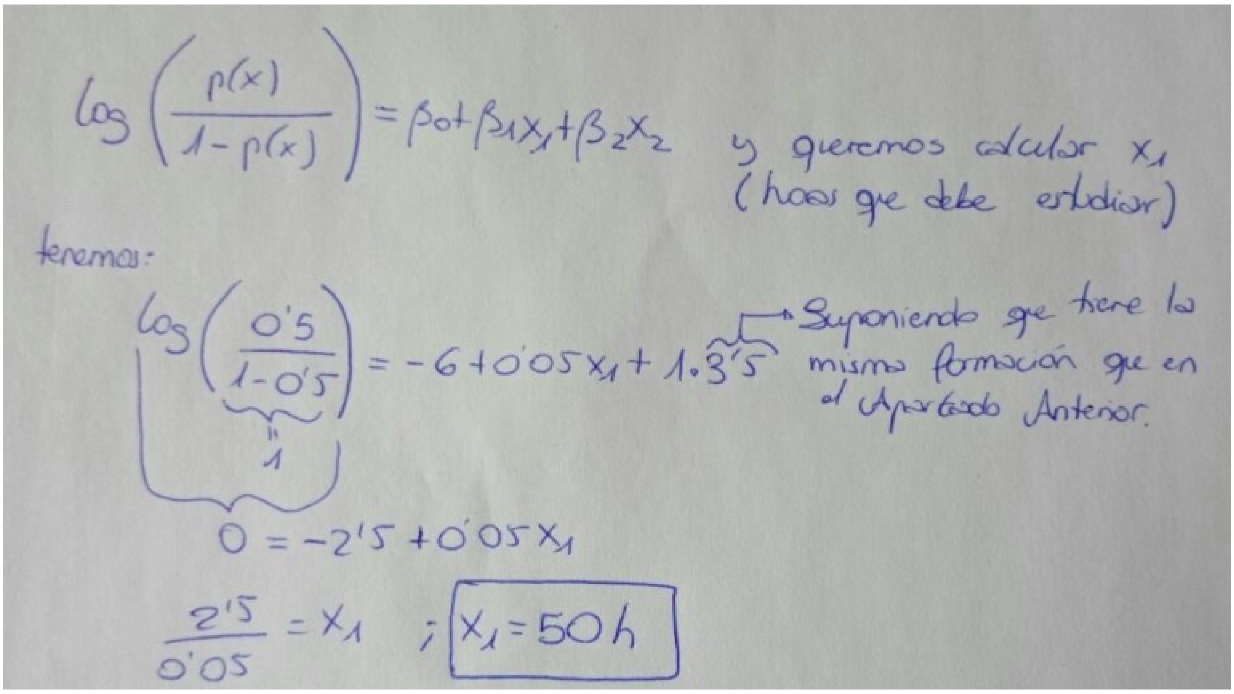


Figura 4.4: horas necesita estudiar un estudiante para tener al menos un 50 % de posibilidades de sacar sobresaliente

Por lo que tendríamos que el alumno, teniendo una formación de 3.5 sobre 4, tendría que dedicarle **50 horas a la asignatura para tener un 50 % de posibilidades de sacar un sobresaliente en AA**.

5. **(2 puntos) Supongamos que deseamos predecir, a partir de X (el porcentaje de beneficio del último año) si una determinada inversión en bolsa dará dividendo en este año (Si o No). Analizamos un conjunto de compañías y descubrimos que el valor medio de X para las compañías que dieron dividendo en el pasado fue X=10, mientras que la media de aquellas que no lo dieron fue X=0. Además la varianza estimada de X para estos dos 2 conjuntos de compañías fue $\sigma = 36$. Finalmente, el 80 % de las compañías dieron dividendos.**

5.1. Suponiendo que X sigue una distribución Normal, predecir la probabilidad de que una compañía de dividendo este año dado que su porcentaje de beneficios del último año fue X=4

Esta vez, usaremos la función que ya conocíamos por otros apartados (4.12, en el libro). Comenzamos por extraer los datos del problema: $\mu_1 = 10$, $\pi_1 = 0{,}8$, $\mu_0 = 0$, $\pi_0 = 1 - 0{,}8 = 0{,}2$, $\sigma^2 = 36$, $X = 4$

$\boxed{\text{2DA}}$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}}{\sum_{l=1}^{k} \pi_l \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}}$$

$$\Rightarrow p_k(4) = \frac{\pi_1 \underbrace{\frac{1}{\sqrt{2\pi}\sigma}}_{cte} \cdot e^{\left(-\frac{1}{2\sigma^2}(4-\mu_1)^2\right)}}{\pi_1 \underbrace{\frac{1}{\sqrt{2\pi}\sigma}}_{cte} \cdot e^{\left(-\frac{1}{2\sigma^2}(4-\mu_1)^2\right)} + \pi_2 \underbrace{\frac{1}{\sqrt{2\pi}\sigma}}_{cte} \cdot e^{\left(-\frac{1}{2\sigma^2}(4-\mu_0)^2\right)}} \qquad \sigma^2 = 36$$

quito $\frac{1}{\sqrt{2\pi}\sigma}$ por ser constante, y queda:

$$= \frac{0'8 \cdot e^{\left(-1/72 \cdot (4-10)^2\right)}}{0'8 \cdot e^{\left(-1/72 \cdot (4-10)^2\right)} + 0'2 \cdot e^{\left(-1/72 \cdot (4-0)^2\right)}} =$$

$$= \frac{0'8 \cdot e^{-0'5}}{0'8 \cdot e^{-0'5} + 0'2 \cdot e^{-0'22}} = \boxed{0'7514}$$

Figura 5.1: probabilidad de que una compañía de dividendo este año

Tenemos que la probabilidad de que una compañía de dividendo este año cuando sus beneficios del pasado año fueron X=4 es de **75.14 %**

6. BONUS-1 (2.5 puntos) Suponga que tomamos un conjunto de datos y lo dividimos en dos partes iguales, una para entrenamiento y otra para test y aplicamos dos técnicas de clasificación distintas. Primero usamos Regresión Logística y obtenemos una tasa de error del 20 % sobre los datos de entrenamiento y del 30 % sobre los datos de test. A continuación usamos k-NN con k=1 y obtenemos una tasa promedio de error (promedio entre entrenamiento y test) del 18 %. Basándose en estos resultados, ¿qué método deberíamos preferir usar para clasificar nuevas observaciones? Justificar adecuadamente la respuesta.

Con RLG, tenemos un 20 % de error sobre datos training y 30 % sobre los datos test, por tanto, tendríamos un $\frac{20+30}{2}$=**25 %** de error medio con RLG, mientras que usando KNN con k=1, obteníamos una tasa promedio de error de **18 %**. Por tanto, si con KNN con k=1 seguimos obteniendo un error menor que el 25 % de la RLG, **deberíamos de preferir usar KNN (con k=1) para clasificar nuevas observaciones**.