Tarea 2: Introducción a Ciencia de Datos.

Rojas Gutiérrez Rodolfo Emmanuel

2 de septiembre de 2021

1. Ejercicios

A lo largo de esta tarea, $M_{n\times p}(\mathbb{R})$ denota el espacio de todas las matrices de dimensión $n\times p$ con coeficientes en los reales, adicionalmente si $A \in M_{n\times p}(\mathbb{R})$ entonces A' denota a la matriz transpuesta de A. Por último, la notación I_p esta reservada para la matriz identidad de dimensión $p \times p$. Por otro lado, se considerará que un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ ha sido fijado. Finalmente y para comenzar con la resolución del ejercicio 1, se deja una última observación concerniente únicamente a dicho ejercicio.

Observación 1 (Ejercicio 1)**.** *Dado un vector aleatorio p-variado con valores en $\mathbb{R}^p$, es posible, gracias al isomorfismo existente entre $\mathbb{R}^p$ y $M_{p\times 1}(\mathbb{R})$, ver a x como un elemento aleatorio que toma valores en $M_{p\times 1}(\mathbb{R})$, con lo que expresiones del tipo Bx con $B \in M_{p\times p}(\mathbb{R})$ o x' cobran sentido. De igual manera, tendrá sentido el hablar de la norma euclideana en $\mathbb{R}^p$ de expresiones del tipo Bx, más aún, hay dos formas de calcular $\|Bx\|$ la primera, obteniendo la suma del cuadrado de las entradas de la matriz Bx y calculando la raíz cuadrada de dicha suma, y la segunda mediante la siguiente igualdad:*[1]

$$\|Bx\| = (x'B'Bx)^{1/2}.$$

Todo lo anterior, evidentemente se cumple para vectores no aleatorios en $\mathbb{R}^p$, por ejemplo, el vector de medias de algún vector aleatorio con valores en $\mathbb{R}^p$. Por último, se recuerda al lector que la norma de Frobenius en $M_{p\times p}(\mathbb{R})$ se define como:

$$\|A\|_F = \mathrm{Tr}(A'A)^{1/2}, \quad A \in M_{p\times p}(\mathbb{R}).$$

Y que efecto, es una norma. △

Ejercicio 1:

Sea x un vector aleatorio p-dimensional, $x' = (x_1, \ldots, x_p)$, con $\mathbb{E}[x] = 0$ y $\mathrm{Var}(x) = V$. Suponga que usted esta interesado en minimizar

$$\mathbb{E}\left\{\left(x_1 - \sum_{j\neq 1}^{p} b_{ij}x_j\right)^2\right\}.$$

[1] Note que por un lado estamos viendo a Bx como un vector aleatorio con valores en $\mathbb{R}^p$, mientras que, por el otro lado estamos viéndolo como un elemento aleatorio en $M_{p\times p}(\mathbb{R})$

Más aún, suponga que no solo se esta interesado en expresar a x_1 en términos del resto de variables, como en la expresión anterior, sino que queremos expresar a cada de las componentes de x en términos de las demás. Esto es, considere el problema de minimizar con respecto a $\{b_{ij} : i, j \in \{1, \ldots, p\},\ i \neq j\}$ a

$$L = \mathbb{E}\left[\sum_{i=1}^{p}\left(x_i - \sum_{j\neq i}^{p} b_{ij}x_j\right)^2\right].$$

Muestre que, si B es un una matriz en $M_{p\times p}(\mathbb{R})$, tal que:

$$B_{ij} = \begin{cases} b_{ij} & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}, \quad i, j \in \{1, \ldots, p\},$$

entonces

$$L = \mathbb{E}\left[\sum_{i=1}^{p}\left(x_i - \sum_{j\neq i}^{p} b_{ij}x_j\right)^2\right] = \mathbb{E}\left\{\|x - Bx\|^2\right\}.$$

donde $\|\|$ denota a la norma euclideana en $\mathbb{R}^p$. Y, muestre que $L = \left\|(I - B)V^{1/2}\right\|_F^2$, donde, $\|\|_F$ denota a la norma de Frobenius en $M_{p\times p}(\mathbb{R})$.

Observación 2: *Si z es un vector con media μ y varianza Σ, entonces $\mathbb{E}[z'Az] = \mu' A\mu + \mathrm{Tr}(A\Sigma)$.* △

Solución. Sean x un vector aleatorio como en el enunciado de ejercicio, $\{b_{ij} : i, j \in \{1, \ldots, p\},\ i \neq j\} \subseteq \mathbb{R}$ y defina $B \in M_{p\times p}(\mathbb{R})$ como aquella matriz en $M_{p\times p}(\mathbb{R})$ con coeficientes:

$$B_{ij} = b_{ij}, \quad i, j \in \{1, \ldots, p\},\ i \neq j, \tag{1}$$

y

$$B_{ii} = 0, \quad \text{para cada } i \in \{1, \ldots, p\}. \tag{2}$$

Entonces, por lo comentado en la Observación 1, es posible ver a $x - Bx = (I_p - B)x$ como un elemento aleatorio en $M_{p\times 1}(\mathbb{R})$ tal que para $i \in \{1, \ldots, p\}$, cumple que:

$$\begin{aligned}((I_p - B)x)_{i1} &= \sum_{j=1}^{p}(I_p - B)_{ij}x_{j1} = \sum_{j=1}^{p}((I_p)_{ij} - B_{ij})x_j \\ &= \delta_{ii} - B_{ii}x_i + \sum_{j\neq i}^{p}(\delta_{ij} - B_{ij})x_j,\end{aligned}$$

donde, $\delta_{ij} = \mathbb{1}_{\{i=j\}}$. Así, usando (1) y (2) es posible continuar la igualdad anterior como:

$$\begin{aligned}
&= (\delta_{ii} - B_{ii})x_i + \sum_{j\neq i}^{p}(\delta_{ij} - B_{ij})x_j \\
&= (1-0)x_i + \sum_{j\neq i}^{p}(0 - b_{ij})x_j \\
&= x_i - \sum_{j\neq i}^{p} b_{ij}x_j.
\end{aligned}$$

De este, modo

$$((I_p - B)x)_{i1} = x_i - \sum_{j\neq i}^{p} b_{ij}x_j, \quad i \in \{1,\ldots,p\}.$$

Por lo cual

$$\|x - Bx\| = \|(I_p - B)x\| = \sqrt{\sum_{i=1}^{p}((I_p - B)x)_{i1}^2} = \sqrt{\sum_{i=1}^{p}\left(x_i - \sum_{j\neq i}^{p} b_{ij}x_j\right)^2}.$$

Lo que implica que

$$\|x - Bx\|^2 = \sum_{i=1}^{p}\left(x_i - \sum_{j\neq i}^{p} b_{ij}x_j\right)^2.$$

Así, tomando esperanzas a ambos lados de la igualdad anterior,[2] se obtiene:

$$L = \mathbb{E}\left\{\sum_{i=1}^{p}\left(x_i - \sum_{j\neq i}^{p} b_{ij}x_j\right)^2\right\} = \mathbb{E}[\|x - Bx\|^2].$$

Que es la primer igualdad solicitada. Por otro lado, de la Observación 1 se sigue que

$$\|x - Bx\|^2 = \|(I_p - B)x\|^2 = x'(I_p - B)'(I_p - B)x,$$

Así, tomando valores esperados a ambos lados de la igualdad anterior,[3] se sigue que

$$L = \mathbb{E}[\|x - Bx\|^2] = \|(I_p - B)x\|^2 = \mathbb{E}[x'(I_p - B)'(I_p - B)x]. \tag{3}$$

Finalmente, de la Observación 2 tomando $A = (I-B)'(I-B)$, $z = x$ y recordando que por hipótesis $\mathbb{E}[x] = 0$ y $\mathrm{Var}(x) = V$, se tiene que:

$$\mathbb{E}[x'(I - B)'(I - B)x] = \mathrm{Tr}((I - B)'(I - B)V).$$

[2]Las cuales existen por la no negatividad de las variables implicadas.
[3]Nuevamente, por la no negatividad de las v.a's implicadas este paso no debería producir ruido.

Ahora, dado que $V \in M_{p\times p}(\mathbb{R})$ es una matriz de covarianzas, entonces, V debe ser una matriz semidefinida positiva, por lo que, existe una matriz simétrica $V^{1/2} \in M_{p\times p}(\mathbb{R})$ tal que[4] $V = V^{1/2}V^{1/2}$, así

$$\begin{aligned}\mathbb{E}[x'(I-B)'(I-B)x] &= \mathrm{Tr}((I-B)'(I-B)V^{1/2}V^{1/2}) \\ &= \mathrm{Tr}(V^{1/2}(I-B)'(I-B)V^{1/2}) \\ &= \mathrm{Tr}((V^{1/2})'(I-B)'(I-B)V^{1/2}) = \left\|(I-B)V^{1/2}\right\|_F^2,\end{aligned}$$

donde, la segunda igualdad se debe a que $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ para cada par A, B de matrices en $M_{p\times p}(\mathbb{R})$, la tercera igualdad se debe a la simetría de $V^{1/2}$ y la última se sigue de la definición de la norma de Froebenius. Así, por la igualdad anterior y por (3) es posible concluir que:

$$L = \left\|(I-B)V^{1/2}\right\|_F^2.$$

La segunda igualdad solicitada, lo que concluye el ejercicio.

■

Ejercicio 2:

Considere el conjunto de datos en la tabla 1. El conjunto mostrado es obviamente ficticio, no obstante suponga que las observaciones son individuos, donde y es igual a uno si el individuo correspondiente a dicha observación padecía cierta enfermedad y cero en otro caso. Además, las predictoras son fiebre (V_1), tos (V_2), piel enrojecida (V_3) y flujo nasal (V_4). Donde, 1 significa que el síntoma esta presente y 0 si no.

a) Para el conjunto de datos 1, calcule todas las probabilidades requeridas para el clasificador Bayesiano ingenuo.

b) Al aplicar el clasificador a los 6 individuos, ¿Hay errores de clasificación?

c) Suponga tres pacientes, uno con tos y fiebre, otro con flujo nasal y fiebre y un tercero con flujo nasal y piel enrojecida. ¿Como los clasifica la herramienta desarrollada.

Obs	V_1	V_2	V_2	V_4	Y
1	0	1	1	1	1
2	0	1	0	1	1
3	1	0	1	0	1
4	0	0	0	1	0
5	0	0	0	0	0
6	0	1	1	0	0

Cuadro 1: Datos.

[4] Ver anexo.

Observación 3 (Ejercicio 2.)**.** *En ocasiones, a lo largo de este ejercicio se referirá como enfermos a los individuos que padezcan la enfermedad mencionada en el enunciado del ejercicio y, como sanos a los que no la padezcan, sin importar sus demás condiciones de salud. Por último, el código implementado para la realización de este ejercicio, puede consultarse en el script adjunto a este documento.* △

a)

Solución. En este caso, se quiere discernir entre dos poblaciones los enfermos denotados por 1 y los sanos denotados por 0. Como herramienta para realizar dicha clasificación se cuenta con 6 muestras, tabla 1, la observación $i = 1, \ldots, 6$ corresponde con las mediciones tomadas para el individuo i. La notación en la tabla indica lo siguiente:

V_1 el individuo presenta fiebre: 1 en caso positivo, 0 e.o.c.

V_2 el individuo presenta tos: 1 en caso positivo, 0 e.o.c.

V_3 el individuo presenta irritación en la piel: 1 en caso positivo, 0 e.o.c.

V_4 el individuo presenta flujo nasal: 1 en caso positivo, 0 e.o.c.

Y el individuo padece la enfermedad: 1 en caso positivo, 0 e.o.c.

A modo de ejemplo, el primer renglón de la tabla 1 indica que: el individuo 1 presentaba tos, irritación en la piel y flujo nasal, además, de que el padecía de la enfermedad.

La clasificación se hará haciendo uso del clasificador Bayes Naive. A modo de resumen, recuerde que si se tienen k poblaciones y p predictores $X = (X_1, \ldots, X_p)$, entonces el clasificador Bayesiano óptimo esta dada por

$$g^*(x) = \text{argmax}_{j\in\{1,\ldots,k\}} \mathbb{P}[Y = k|X = x], \quad x \in \mathbb{R}^p.$$

En nuestro caso, se tienen 2 poblaciones y 4 predictores, así, si denota por $V = (V_1, \ldots, V_4)$ entonces el clasificador bayesiano óptimo queda dado por:

$$g^*(v) = \text{argmax}_{i\in\{1,2\}} \left\{\mathbb{P}[Y = i|V = v]\right\}, \text{ donde } v = (v_1, v_2, v_3, v_4) \in \{0, 1\}^4.$$

Lo que es equivalente a:

$$g^*(v) = \text{argmax}_{i\in\{1,2\}} \left\{\mathbb{P}[Y = i]\mathbb{P}[V = v|Y = i]\right\}, \text{ donde } v = (v_1, v_2, v_3, v_4) \in \{0, 1\}^4.$$

Luego, el clasificador Bayes Naive se obtiene del supuesto de independencia entre los predictores, es decir, esta dado por

$$g^*(v) = \text{argmax}_{i\in\{1,2\}} \left\{\mathbb{P}[Y = i] \prod_{k=1}^{4} \mathbb{P}_k[v_k|Y = i]\right\}, \text{ con } v_k \in \{0, 1\} \text{ para cada } k \in 1, \ldots, 4.$$

donde, dado que los predictores solo toman dos valores, $\mathbb{P}_k[v_k|Y = i]$ indica la probabilidad de que $V_k = v_k$ dado que $Y = i$ con $k \in 1, \ldots, 4$ e $i \in \{0, 1\}$. En otros palabras, si un individuo padece la enfermedad, entonces, la probabilidad de que dicho individuo presente el síntoma $k \in \{1, \ldots, 4\}$ es $\mathbb{P}_k[1|Y = 1]$, mientras que, si un individuo esta sano, entonces, la probabilidad de que dicho individuo presente el síntoma $k \in \{1, \ldots, 4\}$ es $\mathbb{P}_k[1|Y = 0]$. Por último, en clase se vio

que el clasificador anterior podía replantearse de forma equivalente de la siguiente manera. Sea $V = (v_1, v_2, v_3, v_4)$ el conjunto de síntomas de un individuo codificado en ceros y unos, y sean

$$LD = \log(\mathbb{P}[Y = 1]) + \sum_{j=1}^{4} \log\left(\frac{\mathbb{P}_k[Y = 1|v_k]}{\mathbb{P}[Y = 1]}\right),$$
$$LI = \log(1 - \mathbb{P}[Y = 1]) + \sum_{j=1}^{4} \log\left(\frac{1 - \mathbb{P}_k[Y = 1|v_k]}{1 - \mathbb{P}[Y = 1]}\right). \tag{4}$$

donde, $\mathbb{P}_k[Y = 1|v_k]$ es la probabilidad de que el individuo padezca la enfermedad dado que su registro para el síntoma V_k fue $v_k \in \{0, 1\}$. Entonces, si $LD > LI$ se clasifica al individuo como enfermo y en otro caso se clasifica como sano. Por lo cual, las probabilidades necesarias para construir el clasificador Bayes Naive son:

$$\mathbb{P}[Y = 1], \mathbb{P}_k[Y = 1|v_k], \quad v_k \in \{0, 1\}, \ k \in \{1, 2, 3, 4\}.$$

Las probabilidades anteriores se estimaron de los datos, siguiendo las directrices del Ejemplo contenido en las paginas $58 - 68$ de las notas de clase. Es decir, $\mathbb{P}[Y = 1]$ se estimó como la proporción de enfermos en las seis observaciones realizadas, de este modo:

$$\mathbb{P}[Y = 1] = \frac{\text{Número de enfermos en los datos}}{\text{Total de observaciones}} = \frac{3}{6} = \frac{1}{2}.$$

Mientras que

$$\mathbb{P}_k[Y = 1|v_k] = \frac{\text{Número de enfermos en los datos con registro } v_k\text{, para el padecimiento } V_k}{\text{Total de pacientes en los datos con registro } v_k\text{, para el padecimiento } V_k.}$$

Por ejemplo, para calcular $\mathbb{P}_1[Y = 1|0]$, observe la tabla 1 y note que el total de enfermos con un registro de 0 en el síntoma V_1 es 2, mientras que el total de observaciones de individuos con registro 0 para el síntoma V_1, es igual 5, por lo cual:

$$\mathbb{P}_1[Y = 1|0] = \frac{2}{5}.$$

Bajo esta metodología, se calcularon con ayuda del software R las restantes probabilidades $\mathbb{P}_k$, las cuales se presentan en el siguiente resumen

$\mathbb{P}_k[Y = 1\|v_k]$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$v_k = 0$	$\frac{2}{5}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$v_k = 1$	1	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$

Cuadro 2: Probabilidades estimadas con base en los datos de la tabla 1, para el modelo Naive Bayes.

Así, se da por concluido este primer inciso.

■

b)

Solución. Considere la notación del inciso anterior. Haciendo uso de las probabilidades en la Tabla 2, se calcularon los valores de LD y LI para cada una de las 6 observaciones contenidas en la Tabla 1 de datos, todo esto se realizó haciendo uso de R con lo que se obtuvo el siguiente resumen

$Obs/NaiveBayes$	LI	LD	Clase NBayes	Clase Real	Error
1	-1.674	-0.018	1	1	No
2	-0.989	-0.704	1	1	No
3	-5.099	-0.481	1	1	No
4	-0.303	-1.390	0	0	No
5	0.383	-2.076	0	0	No
6	-0.989	-0.704	1	0	Sí

Cuadro 3: Clasificador Bayes-Naive aplicado a los datos de la Tabla 1, es decir, al conjunto de datos de entrenamiento.

Como puede ver, sobre el conjunto de datos en los cuales el clasificador esta basado, solo se comete un error de clasificación, lo que da por terminado el inciso ■

c)

Solución. Para este último inciso, siga considerando la notación de los incisos anteriores. Luego, suponga que llegan tres individuos nuevos que refieren los siguientes síntomas:

1. Individuo 1: tos y fiebre, sin ningún otro síntoma.
2. Individuo 2: flujo nasal y fiebre, sin ningún otro síntoma.
3. Individuo 3: flujo nasal y piel enrojecida, sin ningún otro síntoma.

siguiendo las ideas de los incisos anteriores, los síntomas de los cuatro nuevos sujetos se pueden codicar como: $(1, 1, 0, 0)$ para el primero de ellos, $(1, 0, 0, 1)$ para el segundo y $(0, 0, 1, 1)$ para el último de ellos. Utilizando estas codificaciones es posible obtener haciendo uso de las probabilidades estimadas en la Tabla 2, los valores de LD y LI para cada uno de los nuevos individuos, y por ende, la clase asignada para cada uno de ellos por el clasificador Naive Bayes desrrollado. Nuevamente, esta tarea se hizo con una función programada en R, con la que se obtuvo el siguiente resumen

$Nvo.Individuo/NaiveBayes$	LI	LD	Clase NBayes
1	-5.099	-0.481	1
2	-5.099	-0.481	1
3	-0.989	-0.704	1

Cuadro 4: Clasificador Bayes-Naive aplicado a los datos de los nuevos sujetos.

Se destaca que todos los individuos fueron catalogados en la clase 1, es decir, todos se catalogaron como enfermos. Esto llamo mi atención, ya que todos estos últimos sujetos tenían algo en común, presentaban al menos dos síntomas, por lo que, pareciera que la herramienta desarrollada tiende a clasificar una persona como enferma, a partir de que esta presenta dos de los cuatro síntomas

medidos, lo cual tiene sentido. De este modo, con el fin de corroborar esta hipótesis se construyó la tabla 5, en la que se presentan todas las posibles combinaciones de síntomas y las clasificaciones dadas por nuestro clasificador Naive Bayes

Combinación de síntomas	Clase NBayes
$(0,0,0,0)$	0
$(1,0,0,0)$	1
$(0,1,0,0)$	0
$(0,0,1,0)$	0
$(0,0,0,1)$	0
$(1,1,0,0)$	1
$(1,0,1,0)$	1
$(1,0,0,1)$	1
$(0,1,0,1)$	1
$(0,1,1,0)$	1
$(0,0,1,1)$	1
$(1,1,1,0)$	1
$(1,0,1,1)$	1
$(1,1,0,1)$	1
$(0,1,1,1)$	1
$(1,1,1,1)$	1

Cuadro 5: Todas las combinaciones posibles de síntomas.

Analizando la Tabla 5 se corrobora nuestra hipótesis y se pueden obtener otras conclusiones: El síntoma que parece estar más relacionado con la enfermedad es la fiebre, pues es el único síntoma que de presentarse en un individuo en ausencia de los demás síntoma, provocará que este sea catalogado como enfermo. Todos los demás síntomas, al presentarse de manera individual dan un diagnostico negativo para esta enfermedad. Y, como era de esperarse, la ausencia de síntomas también es catalogada como ausencia de la enfermedad. ■

2. Anexo: Existencia de la matriz raíz cuadrada.

Sea $V \in M_{p\times p}(\mathbb{R})$ una matriz semidefinida positiva, en particular se cumple que V es simétrica por lo que del Teorema Espectral, se sigue la existencia de una matriz ortogonal P y una matriz diagonal D, ambas en $M_{p\times p}(\mathbb{R})$, tales que, $V = PDP'$ donde, la diagonal de D contiene a los valores propios de V. Luego, dado que V es semidefinida positiva, todos sus valores propios son mayores o iguales a cero, así, sea $D^{1/2}$ la matriz diagonal que contiene a la raíz cuadrada de cada uno de los elementos de la diagonal de D, se sigue que $D = D^{1/2}D^{1/2}$ y que $V = PD^{1/2}D^{1/2}P'$. Luego, dado que P es ortogonal $P'P = PP' = I$, así, $V = PD^{1/2}P'PD^{1/2}P' = (PD^{1/2}P')^2$. De este modo, tome $V^{1/2} = PD^{1/2}P'$, es claro que $V^{1/2}$ es simétrica y por lo anteriormente probado es igualmente claro que $V = V^{1/2}V^{1/2}$.