UNIVERSIDAD DEL PACÍFICO
Departamento Académico de Economía
Machine Learning para Economistas
Segundo semestre de 2022
Profesor: F. Rosales, JP: R. Arauco

PRÁCTICA CALIFICADA 1

1. Muestre que una variable aleatoria Gaussiana tiene una distribución de probabilidad perteneciente a la familia de distribuciones exponenciales. (5 puntos)

   Respuesta. Considere $y \overset{iid}{\sim} \mathcal{N}(\mu, \sigma)$. La distribución Gaussiana pertenece a la familia de distribuciones exponenciales dado que

$$
\begin{aligned}
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, && (1)\\
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}}\sigma^{-1} \exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, && (2)\\
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}} \exp\{\log(\sigma^{-1})\} \exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, && (3)\\
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}} \exp\{-\log(\sigma)\} \exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, && (4)\\
f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 - \log(\sigma)\right\}, && (5)\\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2\sigma^2}y^2 - \frac{\mu}{\sigma^2}y + \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}, && (6)\\
& && (7)
\end{aligned}
$$

   donde es claro que la función masa de probabilidad tiene la forma

$$f(y;\theta) = h(y)\exp[\eta(\theta)\cdot T(y) - A(\theta)],$$

   donde $h(y) = \frac{1}{\sqrt{2\pi}}$, $\eta(\mu,\sigma) = (-\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})^\top$, $T(y) = (y, y^2)^\top$, y $A(\mu,\sigma) = -\frac{\mu^2}{2\sigma^2} + \log\sigma$. □

2. Considere el método de clasificación KNN con $K = 3$, la norma $\|\boldsymbol{x}\|_\infty = \max_{i=1,\dots,n}\{|x_i|\}$, y los siguientes datos

Obs.	Data	X_1	X_2	X_3	Y
1	train	1/2	1	3/4	Verde
2	train	1	1	1/4	Verde
3	train	3/4	1/2	3/4	Verde
4	train	1/4	1/4	1/2	Azúl
5	train	1	3/4	1/4	Rojo
6	train	1/2	1/2	3/4	Azúl
7	test	0	0	0	Rojo
8	test	1/2	1/2	1/2	Azúl
9	test	1	1	1	Verde

(a) Clasifique las observaciones en data de prueba. (2 puntos)

(b) Calcule el ratio de error de su clasificación en (a). (1 punto)

Respuesta. Las distancias entre cada observación en data de prueba y las observaciones en data de entrenamiento se muestran en la siguiente tabla: En la última fila de la misma tabla

Obs.	Y	7	8	9
1	Verde	1	1/2	1/2
2	Verde	1	1/2	3/4
3	Verde	3/4	1/4	1/2
4	Azúl	1/2	1/4	3/4
5	Rojo	1	1/2	3/4
6	Azúl	3/4	1/4	1/2
		Azúl	Azúl	Verde

se muestra la clasificación resultante como la etiqueta predominante en los tres vecinos más cercanos. El ratio de error resultante es 1/3, dado que la observación 7 se clasifica mal, pero las 8 y 9 se clasifican bien. □

3. Considere el caso de regresión lineal múltiple de $\boldsymbol{y} \in \mathbb{R}^n$ con respecto a $\boldsymbol{X} \in \mathbb{R}^{n \times p}$

   (a) Muestre que la recta de regresión estimada siempre pasa por el punto $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$. (2.5 puntos)

   (b) Muestre que el R^2 es igual al cuadrado de la correlación entre $\boldsymbol{y}$ y $\hat{\boldsymbol{y}}$. (2.5 puntos)

   (c) Muestre que $\boldsymbol{H}_{ii} \in [0, 1]$, donde $\hat{\boldsymbol{f}} = \boldsymbol{H}\boldsymbol{y}$. (2.5 puntos)

   (d) Muestre que $\sum_{i=1}^{n} \boldsymbol{H}_{ii} = p$. (2.5 puntos)

Respuesta. (a) Sabemos que $\mathbf{1}^\top \boldsymbol{\epsilon} = 0$, luego

$$\begin{aligned} \mathbf{1}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= 0 \\ \bar{y} &= \bar{\boldsymbol{x}}^\top \hat{\boldsymbol{\beta}} \\ \bar{y} &= \hat{f}(\bar{x}). \end{aligned}$$

(b) Sabemos que:

$$\begin{aligned} \mathbf{1}^\top \boldsymbol{\epsilon} &= 0 \\ \mathbf{1}^\top (\boldsymbol{y} - \hat{\boldsymbol{y}}) &= 0 \\ \bar{y} &= \bar{\hat{y}}. \end{aligned}$$

y que

$$\begin{aligned} s_{y,\hat{y}} &= \mathbb{C}\text{ov}(\boldsymbol{y}, \hat{\boldsymbol{y}}) \\ &= \mathbb{C}\text{ov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\epsilon}) \\ &= \mathbb{C}\text{ov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\epsilon}) \\ &= \mathbb{C}\text{ov}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}) \\ &= s_{\hat{y},\hat{y}}. \end{aligned}$$

Luego

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \text{... usando 1er resultado} \\ &= \frac{s_{\hat{y},\hat{y}}}{s_y^2} \\ &= \left(\frac{s_{y,\hat{y}}}{s_y^2}\right)\left(\frac{s_{\hat{y},\hat{y}}}{s_{\hat{y},\hat{y}}}\right) \text{... usando 2do resultado} \\ &= \frac{s_{y,\hat{y}}^2}{s_y^2 s_{\hat{y}}^2} \\ &= \left(\frac{s_{y,\hat{y}}}{s_y s_{\hat{y}}}\right)^2 \\ &= \rho_{y,\hat{y}}^2. \end{aligned}$$

(c) Sabemos que $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$ y que $\boldsymbol{H} = \boldsymbol{H}^\top$, i.e. $\boldsymbol{H}_{ij} = \boldsymbol{H}_{ji}$. Luego,

$$\begin{aligned}
\boldsymbol{H}_{ii} &= \sum_{j=1}^{n} \boldsymbol{H}_{ij}\boldsymbol{H}_{ji} \\
&= \boldsymbol{H}_{ii}^2 + \sum_{j\neq i}^{n} \boldsymbol{H}_{ij}\boldsymbol{H}_{ji} \\
&= \boldsymbol{H}_{ii}^2 + \sum_{j\neq i}^{n} \boldsymbol{H}_{ij}^2.
\end{aligned}$$

En consecuencia, $\boldsymbol{H}_{ii} \geq 0$. Adicionalmente, $\boldsymbol{H}_{ii}(1 - \boldsymbol{H}_{ii}) \geq 0$, y es obvio que $\boldsymbol{H}_{ii} \leq 1$.

(d) Sabemos que $\sum_{i=1}^{n} \boldsymbol{H}_{ii} = \mathrm{Tr}(\boldsymbol{H})$, luego

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{H}_{ii} &= \mathrm{Tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) \\
&= \mathrm{Tr}((\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) \\
&= \mathrm{Tr}(\boldsymbol{I}_p) = p.
\end{aligned}$$

□

4. Considere el caso de regresión lineal de $\boldsymbol{y} \in \mathbb{R}^n$ con respecto a $\boldsymbol{x} \in \mathbb{R}^n$. Suponga que se recogen datos y se separan en data de entrenamiento (70% de obs) y data de prueba (30% de obs). Se ajustan dos modelos a la data: una regresión lineal simple y una regresión cúbica, es decir: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{x}^2 + \beta_3 \boldsymbol{x}^3 + \epsilon$.

   (a) Suponga que la verdadera relación entre $\boldsymbol{x}$ e $\boldsymbol{y}$ es lineal, es decir, $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \epsilon$. Considere el MSE para la regresión lineal, y para la regresión cúbica utilizando data de entrenamiento. ¿Cómo se comparan estas dos cantidades? (1 punto)

   (b) Responda (a) usando la data de prueba en lugar de la data de entrenamiento. (1 punto)

   Respuesta. (a) Se espera que el MSE del modelo de ambos modelos sean muy similares. (b) Se espera que el MSE del modelo lineal sea mucho menor que el del modelo cúbico □