Tarea 6 - Análisis de datos
Giovanni Gamaliel López Padilla

Problema 1

Verifica que la información mutua está bien definida. Es decir

$$I(X,Y) = I(Y,X)$$

Se tiene que:

$$I(X,Y) = log\left(\frac{P(X|Y)}{P(Y)}\right)$$

aplicando la definición de $P(X|Y)$ se obtiene lo siguiente:

$$\begin{aligned} I(X,Y) &= log\left(\frac{P(X|Y)}{P(Y)}\right) \\ &= log\left(\frac{P(X,Y)}{P(Y)P(X)}\right) \\ &= log\left(\frac{P(Y,X)}{P(Y)P(X)}\right) \\ &= log\left(\frac{P(Y|X)}{P(X)}\right) \\ &= I(Y,X) \end{aligned}$$

Problema 2

Sea $\mathbf{X}$=($\mathbf{X_1}$,$\mathbf{X_2}$,$\mathbf{X_3}$)$\sim \mathcal{N}(\mu,\Sigma)$ con

$$\mu^T = (2,-3,1) \qquad \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

a) **Encuentra la distribución de $\mathbf{X_1}$+$\mathbf{X_2}$-$\mathbf{X_3}$.**
Sea $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$, entonces se tiene que:

$$Y = \sum_{i=1}^{3} a_i X_i$$

la cual es una combinación lineal de X_1, X_2, X_3 donde $a = \{1, 1-1\}$. Entonces $Y \sim N(EY, Var(T))$. Calculando EY, se tiene lo siguiente:

$$\begin{aligned} EY &= E(X_1 + X_2 - X_3) \\ &= EX_1 + EX_2 - EX_3 \\ &= 2 - 3 - 1 \\ EY &= -2 \end{aligned}$$

Calculando $Var(Y)$ se tiene lo siguiente:

$$\begin{aligned} Var(Y) &= Var(X_1 + X_2 - X_3) \\ &= Var(X_1) + Var(X_2) + (-1)^2 Var(X_3) + 2Cov(X_1, X_2) + 2Cov(X_2, -X_3) + 2Cov(X_1, -X \\ &= Var(X_1) + Var(X_2) + Var(X_3) + 2Cov(X_1, X_2) - 2Cov(X_2, X_3) - 2Cov(X_1, X_3) \\ &= 1 + 2 + 3 + 2(1) - 2(2) - 2(1) \\ Var(Y) &= 1 \end{aligned}$$

Entonces la distribución $X_1 + X_2 - X_3 \sim \mathcal{N}(-2, 1)$.

b) Calcula $\mathbf{EX_1|X_2{=}2}$

Definamos a X_1 y X_2 como lo siguiente:

$$\begin{cases} X_1 &= \mu_1 + \sigma_1 \mathcal{Z}_1 \\ X_2 &= \mu_2 + \sigma_2 \left(\rho \mathcal{Z}_1 + \sqrt{1 - \rho^2} \mathcal{Z}_2 \right) \end{cases}$$

donde $\mathcal{Z}_1, \mathcal{Z}_2 \sim \mathcal{N}(0, 1)$.
Entonces calculando $EX_2|X_1 = x$ se tiene lo siguiente:

$$E(X_2|X_1 = x) = E(\mu_2 + \sigma_2 \left(\rho \mathcal{Z}_1 + \sqrt{1 - \rho^2} \mathcal{Z}_2 \right) |X_1 = x)$$

por linealidad de la esperanza se obtiene lo siguiente:

$$E(X_2|X_1 = x) = E(\mu_2|X_1 = x) + E(\sigma_2 \rho \mathcal{Z}_1|X_1 = x) + E(\rho \sqrt{1 - \rho^2} \mathcal{Z}_1|X_1 = x)$$

como $\mathcal{Z}_1$ y $\mathcal{Z}_2$ son distribuciones independientes de x entonces:

$$E(X_2|X_1 = x) = \mu_2 + \sigma_2 \rho \mathcal{Z}_1 + \sigma_2 \sqrt{1 - \rho^2} E(\mathcal{Z}_2)$$

como $E\mathcal{Z}_2 = 0$ entonces:

$$\begin{aligned} E(X_2|X_1 = x) &= \mu_2 + \sigma_2 \rho \mathcal{Z}_1 \\ E(X_2|X_1 = x) &= \mu_2 + \frac{\sigma_2}{\sigma_1} \rho (x - \mu_1) \end{aligned}$$

Para este caso se tiene que $\rho = \frac{\sqrt{3}}{3}, \mu_1 = 2, \mu_2 = -3, \sigma_1 = 1, \sigma_2 = \sqrt{3}$, entonces:

$$\begin{aligned}
E(X_2|X_1=2) &= \mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x-\mu_1) \\
&= -3 + \frac{\sqrt{3}}{1}\left(\frac{\sqrt{3}}{3}\right)(2-2) \\
E(X_2|X_1=2) &= -3
\end{aligned}$$

Observando el resultado uno podria llegar a equivocarse y decir que X_2 y X_1 son independientes, pero esto es una coincidendia condiconar a la variable X_1 con su promedio.

c) **Encuentra un vector v tal que $\mathbf{X_2}$ y $\mathbf{X_2}$-$\mathbf{v^T}$ $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ sean independientes.**

Sea $Y = X_2 + v^T \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$, entonces, esta variable es igual a:

$$\begin{aligned}
Y &= X_2 + v^T \begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \\
&= X_2 + \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \\
Y &= X_2 - aX_1 - bX_2
\end{aligned}$$

Se tiene que si X_2 y Y son independientes, entonces:

$$EX_2Y = EX_2EY$$

Calculando EX_2Y se tiene que:

$$\begin{aligned}
E(X_2Y) &= E(X_2^2 - aX_1X_2 - bX_2X_3) \\
&= EX_2^2 + E(-aX_1X_2) + E(-bX_2X_3) \\
&= EX_2^2 - aEX_1X_2 - bEX_2X_3
\end{aligned}$$

como las variables X_i son independientes entre si, entonces:

$$EX_2Y = EX_2^2 - aEX_1EX_2 - bEX_1EX_3 \tag{1}$$

Calculando EX_2EY se tiene que:

$$\begin{aligned}
EX_2EY &= EX_2E(X_2 - aX_1 - bX_3) \\
&= EX_2(EX_2 - aEX_1 - bEX_3) \\
EX_2EY &= (EX_2)^2 - aEX_1EX_2 - bEX_2EX_3
\end{aligned}$$

entonces

$$EX_2EY = (EX_2)^2 - aEX_1EX_2 - bEX_2EX_3 \tag{2}$$

Igualando las ecuacione 1 y 2 se tiene lo siguiente:

$$\begin{aligned} EX_2Y - EX_2EY &= 0 \\ EX_2^2 - aEX_1EX_2 - bEX_1EX_3 - (EX_2)^2 + aEX_1EX_2 + bEX_2EX_3 &= 0 \\ EX_2^2 - (EX_2)^2 &= 0 \\ Var(X_2) &= 0 \end{aligned}$$

entonces, el vector v para que X_2 y Y sean independientes puede ser cualquiera, siempre y cuando $Var(X_2) = 0$. Al ser $X_2 \sim \mathcal{N}(\mu, \sigma)$, la cual requiere que $\sigma > 0$, entonces no existe un vector v.

Problema 3

Supongamos que se quiere estimar el número promedio μ de amigos que alguien tiene en Facebook. Se toma una muestra de personas y ellos eligen al azar algunos de sus amigos en Facebook. Se calcula el promedio del número de amigos que estos amigos tienen. Aunque suponemos independencia, argumenta que en general se va a sobrestimar μ de esta manera.

El número de personas que tiene alguien en Facebook puede depender de la edad, actividades que haga. Entonces puede llegar a crear sesgos. Esto es debido a que se pueden tener amigos mutuos o coincidir en varios grupos. Por ende tener un número de amigos semejante, dando así un promedio de amigos sobrestimado.

Problema 6

Sea X una variable aleatoria que toma valores en {1, 2, 3}. Define $\theta = (\theta_1, \theta_2, \theta_3)$ donde $\theta_i = P(X = i)$. Supongamos que tenemos una muestra con n_i observaciones igual a i. i=1,2,3. Calcula $l(\theta)$ y el estimador de máxima verosimilitud.

Se tiene que la verosimilitud es la siguiente:

$$\mathcal{L} = \prod_i P(X = i)$$

entonces:

$$\begin{aligned} \mathcal{L} &= \prod_i \theta_i^{n_i} \\ \mathcal{L} &= \theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3} \end{aligned}$$

por lo tanto la log-verosimilitud es:

$$l(\theta) = n_1 log(\theta_1) + n_2 log(\theta_2) + n_3 log(\theta_3)$$

donde $n = n_1 + n_2 + n_3$ y $\theta_1 + \theta_2 + \theta_3 = 1$.

Problema 7

Considera el siguiente método para estimar el tamaño (N) de una población de animales de un especie particular. Primero se capturan M animales, los marcan y son puestos de nuevo en libertad. Un tiempo más tarde se capturan animales hasta encontrar un animal marcado. Sea X el número total de animales capturados (X incluye el animal marcado). Después se dejan todos los animales en libertad. Se repite lo anterior de tal forma que se obtenga una muestra $\{x_1, x_2, \ldots, x_n\}$ de X (así este procedimiento puede tardar bastante). Puedes suponer que en cada momento la probabilidad de capturar un animal marcado es siempre igual (así se supone que N es mucho mayor que M).

a) **Demuestre que:**

$$P(X = x) = \frac{M}{N}\left(1 - \frac{M}{N}\right)^{x-1}, \qquad x = 1, 2, \ldots$$

Sea Y una variable aleatoria tal que $Y \sim Bernoulli\left(\frac{M}{N}\right)$, donde $\frac{M}{N}$ es la probabilidad de capturar a un animal marcado.

Al ser cada evento independiente del anterior, entonces se tendria que la probabilidad de capturar a x-1 animales no marcados es:

$$P(X = x - 1) = \left(1 - \frac{M}{N}\right)^{x-1} \tag{3}$$

Y la probabilidad de capturar a un animal marcado en el x-esimo intento es $\frac{M}{N}$, por lo tanto, la probabilidad de capturar a x animales:

$$P(X = x) = \frac{M}{N}\left(1 - \frac{M}{N}\right)^{x-1}, \qquad x = 1, 2, \ldots$$

b) **Demuestra que:**

$$\hat{\Theta}_n = \frac{M}{N}\sum_{i=1}^{n} X_i$$

es el estimador de Máximo verosimilitud. ¿Está insesgado?¿Qué puedes decir si $\mathcal{N} \to \infty$?