Segundo Parcial

Noviembre 26 de 2020

Indicaciones generales

- Este es un examen **individual** con una duración de **110 minutos: de 9:00 a 10:25**. Los 5 minutos entre las 10:25 y 10:30 se destinarán a subir la información de forma correcta.
- Debe entrar al aula virtual del curso (por zoom) y encender la cámara web (puede ser mediante el celular).
- Sólo se contestarán preguntas sobre los enunciados del parcial, durante los primeros 10 minutos.
- Las respuestas deben estar totalmente justificadas. Puede determinar los cuantiles con R, pero debe explicar cómo lo hizo.
- En todos los puntos que use `R`, debe justificar el procedimiento tanto en papel (no el código) y debe colocar el código usado en un archivo Rmarkdown con nombre `NombreApellido.Rmd`. Posteriormente, debe adjuntar en `e-aulas` el archivo `.Rmd` y su versión en `.html`.
- Puede enviar fotos rápidas de los procesos de cada punto al acabar el parcial. Posteriormente, puede enviar las imágenes escaneadas (o foto) con buena calidad de los procedimientos detallados de forma clara y ordenada. El archivo `.Rmd` deben adjuntarlo al acabar el parcial.
- Cualquier incumplimiento de lo anterior conlleva la anulación del examen.

1. [25 ptos.] Los vectores aleatorios $\mathbf{X}^{(1)}$ y $\mathbf{X}^{(2)}$ de tamaño (2×1) tienen la media y la matriz de covarianza conjunta:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{bmatrix} = \left[\begin{array}{c} -3 \\ 2 \\ \hline 0 \\ 1 \end{array}\right]$$

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline -\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ \hline 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{array}\right]$$

   Calcule las correlaciones canónicas ρ_1^*, ρ_2^*.

2. [20 ptos.] Suponga que las densidades conjuntas de $\boldsymbol{X} = (X_1, ..., X_p)'$ de dos poblaciones π_1 y π_2 son normales multivariadas con parámetros μ_1 y Σ y μ_2 y Σ respectivamente

   Muestre que:

$$-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_1})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu_1}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu_2})$$
$$= (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})$$

3. [25 ptos.] Considere la matriz de distancias

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0 \end{bmatrix}$$

Agrupe las 4 observaciones mediante clustering jerárquico (procedimiento paso a paso) usando `complete linkage`. Dibuje el dendograma respectivo.

4. [30 ptos.] Cargue el dataset `penguins.RData` (mediante load()). Este dataset contiene información acerca de tres especies diferentes de pingüinos respecto a diferentes variables, coleccionados a partir de tres islas en el archipiélago Palmer, en la Antártida.

   a) [15 ptos.] Usando la longitud del pico (bill length), la profundidad del pico (bill depth) y la logintud de la aleta (flipper length) construya dos modelos de clasificación usando LDA y QDA. ¿Qué modelo es mejor? Recuerde particionar el dataset en un conjunto de training y otro de test.

   b) [15 ptos.] Realice un clustering mediante K-means con todo el dataset de pingüinos (olvidándose de las especies). Compare los clusters obtenidos con las especies verdaderas. ¿Algún cluster contiene más de una especie de pingüino? ¿Se puede asociar algún cluster a alguna especie?