Segundo Examen Parcial

Análisis Estadístico Multivariado

ITESO
Departamento de Matemáticas y Física

Nombre:

Fecha:

Algunas referencias para responder a este examen

1. The hundred-page machine learning book. La sección 5.3 hace un tratamiento breve de la racionalidad detrás de la división de un conjunto de datos.
2. Introduction to statistical learning. La sección 3.2 hace un tratamiento breve de la regresión lineal múltiple, que no es sino una extensión de la regresión lineal simple.
3. El mismo libro de ISLR, par el tema de PCA se puede consultar la sección 12.2.
4. Yo les encomiendo a usar LLM como chatGPT para resolver dudas. Con prompts apropiados, las respuestas pueden ser bastante útiles. Por ejemplo: `how can I compare several models in Python using the AIC?`.

1. Regresión Lineal Simple y Múltiple

Elija una base de datos para proponer un modelo de regresión simple y un modelo de regresión múltiple y conteste las siguientes preguntas:

a) ¿Qué supuestos debe cumplir un modelo que describa la relación lineal entre dos variables? Describa en qué consiste cada uno de ellos.
b) Con base a los datos que eligió, escriba un enunciado planteando el objetivo o el problema a resolver.
c) Obtenga el modelo de regresión simple y escriba su ecuación. De una descripción de cómo estos parámetros impactan sobre la variable dependiente.
d) ¿Por qué es importante separar los datos en 80% para entrenamiento y 20% para prueba?
e) De una interpretación de los resultados obtenidos (Summary)
f) ¿Cuál sería el valor de $T_{\text{crítico}}$ con el que contrastaría el valor de $T_{\text{observado}}$ si se tuviera un nivel de significancia del 0.05?
g) Indique si el modelo lineal se ajusta a los datos basado en las predicciones obtenidas. Justifique su respuesta.

h) ¿Qué son los "outliers"? ¿Cómo influyen en el análisis de regresión? ¿qué solución propone ante la presencia de estos valores?
i) Escriba un enunciado planteando el problema a resolver.
j) Obtenga el modelo de regresión lineal múltiple y escriba su ecuación. De una descripción de cómo estos parámetros impactan sobre la variable dependiente.
k) Interprete los resultados obtenidos (Summary)
l) ¿Qué modelo de regresión es mejor, el simple o el múltiple? Justifique su respuesta. *Hint:* puedes usar el estadístico conocido como AIC junto con el R^2 para comparar los modelos.

2. Análisis de la Varianza

Elija una base de datos para realizar un ANOVA de dos factores y responda las siguientes preguntas:

a) ¿Qué supuestos se deben cumplir para realizar el ANOVA?
b) ¿Cuál es el objetivo del ANOVA?
c) Indique las variables que va a utilizar y especifique quién es la variable dependiente y las variables independientes
d) Escriba un enunciado planteando el problema a resolver. (Plantear la H0 y H1 del problema a resolver).
e) ¿Cuántos niveles tiene cada factor que eligió?
f) Obtenga el modelo del ANOVA y de una interpretación de los resultados.
g) ¿Cuál sería el valor de $F_{\text{crítico}}$ con el que contrastaría el valor de $F_{\text{observado}}$ si se tuviera un nivel de significancia del 0.05?
h) ¿Para que sirven las pruebas Post-Hoc en el ANOVA?
i) De ser necesario, incluya la prueba de Tukey y de la interpretación de los resultados.
j) Conclusiones generales de análisis de la varianza que acaba de realizar.

3. Análisis de Componentes Principales

Elija una base de datos para realizar el análisis de componentes principales y responda las siguientes preguntas:

a) ¿Cuál es el objetivo de realizar el análisis de componentes principales?
b) ¿Qué es lo que sucede si las variables que se están utilizando para realizar el PCA no están correlacionadas? Justifique su respuesta.
c) ¿Bajo qué circunstancias se recomienda hacer una estandarización de los datos?
d) ¿Qué representan los vectores propios (eigenvectores) de la matriz de varianzas y covarianzas en el análisis de componentes principales?
e) ¿Qué representan los valores propios (eigenvalores) de la matriz de varianzas y covarianzas en el análisis de componentes principales?
f) Obtenga la matriz de vectores propios y de una interpretación de los resultados.
g) En el análisis de componentes principales que está realizando ¿Cuál es número óptimo de componentes principales? Justifique su respuesta

h) ¿Cuáles son las variables que más influyen en los primeros dos componentes principales? *Hint:* buscar los *loadings* e interpretarlos.