UNIVERSIDAD ICESI

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA
Aprendizaje Automático 1
Examen 1

Nombre: _____ Código: _____ Nota: _____

Profesor: **Santiago Ortiz - Henry Velasco** Grupo: **01** Fecha: _____ de 20__

Notas:

- Todas las respuestas, gráficas, tablas y operaciones deben ser debidamente justificadas.
- La información que sea obtenida de alguna fuente debe ser citada y referenciada en el documento a entregar.

1) Considere el conjunto de datos "`data1`" del fichero `data_exam1.xlsx`.

   - Realice un análisis exploratorio de datos ¿Considera que podría generar un modelo de regresión lineal con variable categórica (sin interacción) para la variable $\mathbf{Y}$? Justifique. Si la respuesta a la pregunta es SI, genere un modelo de regresión sin interacción e interpretelo.
   - Realice un gráfico de dispersión para $\mathbf{Y}$ vs $\mathbf{X}$, considerando para cada observación su respectivo valor en la variable `Ind` ¿Hay evidencia muestral que sugiera un cambio en la tasa media de cambio de $\mathbf{Y}$ condicionado a incrementos unitarios de $\mathbf{X}$? ¿Considera que un modelo con interacciones sería más adecuado? Si la respuesta a estas preguntas es afirmativa, genere el respectivo modelo, interprete detalladamente los resultados y valide los supuestos del modelo propuesto $\left(\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)\right)$.

2) Considere el conjunto de datos "`data2`" del fichero `data_exam1.xlsx`

   - De acuerdo al análisis del ítem anterior proponga una transformación (raiz, potencia, logarítmica, sinusoidal, etc.) para alguna de las variables y justifique por qué. Dado lo anterior, proponga un modelo de regresión lineal, interpretelo y valide los supuestos del modelo $\left(\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)\right)$.

3) Considere el conjunto de datos "`Wine Quality`" del fichero `datos.xls`. Defina como variable respuesta ($\mathbf{Y}$) la columna `Densidad` y elimine las variables `pH, Sulfatos, Cloruros, Acidez Volátil, Acidez Fija y Calidad de Vino`.

   - Estandarice las variables, calcule las matrices de correlación de Pearson $\left(\hat{\boldsymbol{\rho}}_{(\mathrm{P})}\right)$, Kendall $\left(\hat{\boldsymbol{\rho}}_{(\mathrm{K})}\right)$ y Spearman $\left(\hat{\boldsymbol{\rho}}_{(\mathrm{Sp})}\right)$ y compárelas ¿Qué diferencia encuentra entre las estructuras de dependencias obtenidas?
   - Realice una partición de los datos tipo 80–20, donde el primer 80 % de los datos es una muestra de entrenamiento y el restante 20 % una muestra de prueba/predicción. Luego, construya 3 modelos RLM con las matrices estimadas en el primer ítem $\left(\hat{\boldsymbol{\beta}}_{(\cdot)} = \hat{\boldsymbol{\rho}}^{-1}_{(\cdot)XX}\hat{\boldsymbol{\rho}}_{(\cdot)XY}$ y $\hat{\beta}_0(\cdot) = \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}} - \hat{\boldsymbol{\mu}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{(\cdot)}\right)$. Compare e interprete los valores de los coeficientes de regresión obtenidos por cada método.
   - Realice una predicción con los datos de prueba de acuerdo a los modelos ajustados y calcule el RMSE $\left(\sqrt{\mathrm{MSE}}\right)$ de la predicción ¿Cuál de los modelos lineales propuestos predice mejor?
   - Valide los supuestos teóricos de cada modelo $\left(\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)\right)$ y concluya (Recuerde que, independiente del modelo que estime, siempre $\hat{\varepsilon}_i = y_i - \hat{y}_i$ para $i = 1, \ldots, n$).

- Realice un análisis del diagrama de dispersión del conjunto de datos ¿Se evidencian comportamientos totalmente lineales? Si la respuesta es negativa, sugiera y realice transformaciones de variables (Ejemplo: $\exp(X_i)$, $\sqrt{X_i}$, $\log(X_i)$, X_i^2, $\frac{1}{X_i}$, etc.) y justifique el por qué de esa transformación. Finalmente, genere un modelo RLM e interprételo detalladamente.

4) Se tiene un conjunto de datos que registra la cantidad de anuncios publicitarios en redes sociales que realiza una empresa y su correspondiente retorno de inversión en ventas. Se desea determinar si existe una relación lineal significativa entre la cantidad de anuncios publicitarios y el retorno de inversión. El conjunto de datos **''publicidad.csv''** consta de 200 observaciones y 4 variables que representan los gastos en publicidad (en miles de dólares) y las ventas (en miles de unidades) de un producto en un mercado específico: - `TV`: Gasto en publicidad en televisión. - `Radio`: Gasto en publicidad en radio. - `Newspaper`: Gasto en publicidad en periódicos. - `Sales`: Número de unidades vendidas (en miles)

   - Graficar el retorno de inversión (variable "`Sales`") vs la cantidad de anuncios publicitarios por canal ("`TV`", "`Radio`", "`Newspaper`"). Para ello use la función scatter_matrix() del paquete pandas e interprete los graficos de las variables dos a dos, teniendo en cuenta que nuestra variable respuesta es "`Sales`".
   - Calcular el coeficiente de correlación entre todas las variables y mediante un mapa de calor represente estas correlaciones. ¿Interprete las estructuras de dependencia encontradas?
   - Teniedo en cuenta el punto anterior, elija solo una variable explicativa ("`TV`", "`Radio`", o "`Newspaper`"; la más conveniente) para modelar las ventas ("`Sales`"), ajuste el modelo de regresión lineal simple y encuentra la ecuación de la recta. ¿Cuál es el valor del coeficiente de determinación R^2? ¿Cómo se interpreta este valor?
   - Realiza una predicción del retorno de inversión esperado cuando se realizan 5 anuncios por el canal de la variable escogida en el ítem anterior. ¿Cuál es el intervalo de confianza del 95 % para la predicción?

Pautas

- Entregar un documento de `RMarkdown/Jupyter/etc` (en PDF) con la solución y rutinas de código empleadas (fecha máxima de entrega: Domingo 6 de Octubre hasta las 23:30). En Intu habrá un buzón de entrega.
- El documento debe contener todos los procedimientos, códigos y gráficos necesarios que den debida justificación a lo realizado. Sin embargo, consolide el documento única y exclusivamente con información relevante, evite mostrar salidas de códigos innecesarias, warnings, errores, etc.
- Realizar en equipos conformados por 3-4 participantes (mandatorio).