



Ejercicio de Feedback Unidades 1, 2 y 3

Técnicas de AI: Regresiones, Deep Learning, otros

Objetivo

El objetivo de este ejercicio se centra en proponer una aplicación práctica a los estudiantes para el desarrollo de las destrezas adquiridas en los módulos de la asignatura que comprenden las áreas de conocimiento de Regresiones Lineal y Logística, así como Random Forest y árboles de decisión.

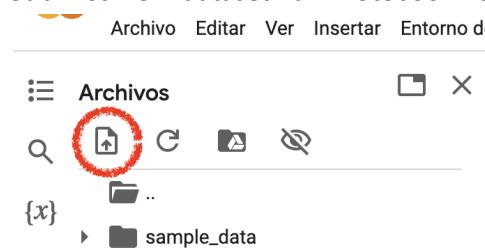
Enunciado

El ejercicio debe ser resuelto tomando como base el Notebook “Feedback_1.ipynb” disponible en la Enseñanza Virtual.

Regresión Lineal

Descargamos el dataframe “precio_casas.csv” disponible en la Enseñanza Virtual. Lo guardamos en una carpeta en Local.

Subimos el dataset al Notebook de Google Collab clicando en el siguiente ítem:



Seleccionamos la ruta donde hemos guardado el csv y clicamos en abrir. Ignoramos el aviso y ejecutamos la celda con las librerías y con la lectura del dataset. Las columnas del dataset son las siguientes:

size: Representa el tamaño de la casa en metros cuadrados

location_index: Índice de calidad de la ubicación

num_rooms: Número de habitaciones en la casa

price: Precio de la casa en miles de dólares

Apartado 1.1

Calcular la media y la varianza de la variable “size” del dataset. Dibujar este histograma incluyendo la media y la desviación estándar y realizar una breve interpretación de la distribución.

Apartado 1.2

Realizar un gráfico de dispersión (scatter plot) de la variable size frente a la variable price, y de la variable location_index frente a la variable price. A priori, en términos de una regresión lineal simple, ¿cuál de las dos variables elegirías como variable predictora para el desarrollo del modelo, tomando como variable objetivo la variable price?

Apartado 1.3

Desarrolla un modelo de regresión lineal con la variable location_index, tomando un 80% del conjunto inicial como conjunto de entrenamiento y “135” como semilla en la división de conjunto de entrenamiento y test. Evalúa los supuestos del modelo. ¿Qué ocurre con el Test Reset de Ramsey? ¿Qué consecuencias en este caso puede tener el resultado del Test Reset de Ramsey en nuestro modelo?

Apartado 1.4

Tomando como variables de entrada todas las variables menos la variable objetivo, desarrolla un modelo de regresión lineal múltiple, de nuevo fijando los parámetros de división de entrenamiento y prueba del apartado anterior. Evalúa los supuestos del modelo e interpreta brevemente los resultados. Calcula el R^2 del modelo. ¿Crees que es un valor de calidad confiable?

Regresión Logística

Descargamos el dataframe “deteccion_cancer.csv” disponible en la Enseñanza Virtual.

Subimos el dataframe al Notebook de Google Collab, siguiendo las indicaciones del primer apartado.

La tabla cuenta con las siguientes columnas:

radius: Radio del nevus del paciente, calculado a partir de mediciones en diferentes imágenes o perspectivas.

texture: Textura del nevus del paciente, que mide la variación en la intensidad o apariencia de la superficie.

concavity: Concavidad de los bordes del nevus del paciente, que indica cuán irregulares o curvados hacia adentro están los bordes.

Diagnosis: Variable objetivo que indica si el nevus que tiene el paciente es maligno (1) o benigno (0).

Apartado 2.1

¿Cuál es la probabilidad de que extrayendo una observación del dataset ocurra que tomemos un nevus maligno? ¿Cuál es el odd asociado a este suceso (extraer una observación donde el

nevus es maligno)? ¿Y el log-odd? (Recuerda que para calcular el logaritmo neperiano de un valor x_0 en Python utilizamos la función `np.log(x_0)`).

Apartado 2.2

Calcula 3 regresiones logísticas simples utilizando las 3 variables predictoras de la tabla (tomando como variable objetivo Diagnosis), formulando en los 3 casos un conjunto de test que suponga el 25% de las observaciones totales con la semilla “135”. En un supuesto de regresión logística simple, basado en la métrica gini, ¿qué variable predice mejor el cáncer?

Apartado 2.3

Realiza una regresión logística múltiple utilizando las 3 variables en un mismo modelo, tomando como parámetros de la regresión logísticas los mismos que en el apartado anterior. Dibuja la curva ROC. Indica la ganancia o pérdida de gini de la regresión logística múltiple con respecto a los tres modelos desarrollados en el apartado anterior.

Apartado 2.4

Si tuvieras que valorar definir un umbral para el modelo desarrollado, ¿en qué métrica te basarías?

Random Forest

Utilizando el mismo dataframe que en el apartado anterior

Apartado 3.1

Entrena un árbol de clasificación, utilizando un conjunto de testeo del 20% (continuando con la semilla “135”), con una profundidad máxima de 7 y un número mínimo de observaciones por split de 10. Obtén sus métricas de precisión, recall y f1-score.

Apartado 3.2

Entrena un Random Forest con 100 estimadores y una profundidad máxima de 6, con el mismo conjunto de entrenamiento y prueba que el apartado anterior. Fija la misma semilla en el desarrollo del modelo (“135”). Dibuja su curva ROC asociada. Obtén sus métricas de precisión, recall y f1-score. Haz una comparativa con los resultados del modelo de árbol de clasificación del apartado anterior. ¿Qué modelo predice mejor de todos los desarrollados en los dos últimos apartados? ¿Cuál crees que pueden ser las causas de que dicho modelo tenga un mejor rendimiento?

Instrucciones de entrega

- **Extensión:** Sin requisitos
- **Nombre del fichero:** Feedback_1_Nombre_Apellido1.ipynb
- **Formato de entrega:** Notebook (.ipynb). Las respuestas y razonamientos deberán ir en celdas de texto o bien en comentarios del código (utilizando “#” para comentar). Se valorará positivamente la limpieza y claridad del Notebook entregado.
- **Fecha límite de entrega:** Lunes, 23 de diciembre

Evaluación

Cada apartado tendrá un valor de 1 punto sobre un total de 10 puntos.



WELCOME
TO
UAX

UAX

Universidad
Alfonso X el Sabio

GRACIAS

