Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

19 de Noviembre, 2024

1. Como interpretaría los coeficientes que resultan de entrenar un modelo de Regresión Logística? Interprete para el caso de un coeficiente negativo, positivo y cero. Cómo es que estos afectan la probabilidad del resultado?
2. Considere los dos siguientes escenarios:
   1. Un científico de datos entrena una regresión lineal para capturar una relación parabólica.
   2. Un científico de datos entrena un algoritmo Random Forest con una gran profundidad y un gran número de árboles.

Cual escenario es probable que resulte en underfitting y cual en overfitting? Explique su respuesta.

1. Considere el siguiente corpus, compuesto por 3 textos:

# List of sentences

sentences = [

"The cat sat on the mat.",

"The dog barked at the cat.",

"The dog sat on the mat."

]

Calcule la representación TF-IDF para las palabras “dog” y “cat”.

1. A screenshot of a white table

   Description automatically generatedCalcule la correlación de Spearman para la siguiente tabla:
2. Comente los errores que pueden apreciarse en los gráficos del anexo y como modificaría los mismos.
3. Una compañía de marketing desea enviar promociones a sus clientes en base sus interacciones en su plataforma de streaming. Su cliente le ofrece un dataset con los patrones de visualizaciones de sus clientes. Además, le informa que la empresa por su conocimiento del rubro estima segmentos de clientes entre 5 y 20 grupos. ¿A qué tipo de problema cree que se está enfrentando? ¿Qué algoritmo utilizaría para resolver este problema y como determinaría la cantidad optima de segmentos?
4. Como se descompone una serie de tiempo? Que modelos utilizaría para capturar las distintas componentes?
5. Suponga que usted quiere utilizar un algoritmo no supervisado para encontrar clusters en su conjunto de datos. Sin embargo, sus datos son categóricos (no numéricos). Que feature engineering debe aplicarse a estas columnas categóricas y por qué?
6. En qué casos puede ser una mala decisión imputar valores faltantes utilizando la media y en cuales la moda?
7. Que ventajas presenta realizar una validación K-fold (con 5 folds) vs una validación donde se realiza un train/test Split de 80/20?

A graph with red lines and numbers

Description automatically generated



