Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2022

23 de Noviembre, 2023

1. Un cliente que vende artículos online le proporciona un dataset con reseñas de los productos que vende en su plataforma. El cliente le comenta que quiere diferenciar los productos buenos de los malos en base a las reseñas. Para eso él etiqueto manualmente una gran porción de las reseñas por lo que el dataset indica los textos y sus respectivas categorías binarias. Con sus conocimientos de NLP que pasos seguiría para entrenar un clasificador binario?
2. Qué problemas pueden surgir si uno imputa valores faltantes utilizando la moda? Qué problemas pueden surgir si uno imputa valores faltantes utilizando la media?
3. Ilustre con un ejemplo que problemas pueden surgir si uno no realiza un muestreo estratificado a la hora de crear sus conjuntos de train y test. Asuma que su dataset esta desbalanceado.
4. Usted entrena un modelo lineal para predecir la cantidad de likes de una publicación en Instragram utilizando una regresión lineal. El modelo le devuelve los parámetros del mismo, por lo que la ecuación resulta:

Usted sabe que los coeficientes son y y . Interprete el modelo a través de sus parámetros.

1. Suponga que usted tiene dos datasets 2-dimensionales y decide graficar scatterplots para cada uno obteniendo las siguientes figuras:

A screenshot of a computer game

Description automatically generated

Usted decide utilizar dos algoritmos de clustering para categorizar los datos: K-Means y DBSCAN. Para que dataset utilizaría cada uno? Justifique su respuesta.

1. Suponga que usted tiene un dataset con variables independientes *X* y una variable dependiente *y.* Usted realiza 3 experimentos y mide la performance utilizando un modelo lineal:
   1. En el primer experimento no realiza ningún pre-procesamiento.
   2. En el segundo decide usar un MinMaxScaler (normaliza/escala las columnas).
   3. En el tercero decide usar un StandardScaler (estandariza las columnas).

Usted no encuentra diferencias de performance entre el experimento 1 y 2 pero si encuentra diferencia respecto al 3. Por qué razón sucede esto?

1. Usted trabaja en el área comercial de un supermercado. El supermercado tiene muchas áreas que se dedican a la venta de productos distintos. Usted es el gerente comercial y le pide a los encargados de área que le reporten sus niveles de venta a fin de mes. El gerente general a su vez le solicita a usted que cada área llegue a un nivel determinado de ventas a fin de mes. Qué gráfica cree usted apropiada para mostrarle al gerente general los resultados mensuales? Dibuje la gráfica y justifique su respuesta.
2. Series de tiempo:
   1. Qué significa que una serie de tiempo sea estacionaria?
   2. Analizando gráficamente, son las dos series de tiempo que se muestran a continuación estacionarias?

A graph of different types of lines

Description automatically generated with medium confidence

1. Qué diferencias existen entre supervised y unsupervised learning? De ejemplos de cómo serían los datasets, mencione un modelo y métricas que utilizaría para cada tipo de tarea.
2. Usted trabaja en una empresa de software de streaming televisivo. Quiere recomendar a los usuarios de la plataforma contenido similar a lo consumido por las personas con gustos mas similares a uno.

Su empresa recaba información de las películas que cada usuario ha visto en el pasado y lo almacena en una tabla como se muestra en la figura.

1. Como transformaría los datos para poder ser utilizados por un modelo?
2. Que modelo entrenaría?
3. Como haría las recomendaciones?

A table with numbers and letters

Description automatically generated