Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2021 - S1

30 de julio, 2021

1. Explique el método Ordinal encoding, para qué sirve y un caso de uso.
2. Explique qué es el modelo Bag of Words en procesamiento de lenguaje natural.
3. Considere las siguientes métricas de clasificación: Precision y Recall.
   1. Suponga que una empresa de servicios televisivos, como Netflix, se enfoca en que los usuarios tengan una buena experiencia al usar su servicio. Para eso, les es muy importante no recomendar películas y series que al usuario no le gusta. Qué métrica cree que la empresa debería maximizar a la hora de entrenar su algoritmo? Justifique su respuesta.
   2. Considere que una película que al usuario le gusta es una predicción positiva. Suponga que usted trabaja para un banco prediciendo fraudes. Al banco le interesa clasificar correctamente si un cliente dada su información bancaria cometerá un fraude o no. Qué métrica cree que la empresa debería maximizar a la hora de entrenar su algoritmo? Justifique su respuesta. Considere que un usuario que cometerá fraude es una predicción positiva.
4. Enumere y explique los distintos componentes de una serie de tiempo. Puede apoyarse en la ecuación vista en clase: xt = ft + st + ct + et.
5. Explique tres características de un dataset que usted chequearía a la hora de hacer un análisis exploratorio.
6. Explique el método de imputación de valores faltantes llamado imputación por la media. Explique dos problemas que pueden surgir si uno imputa valores faltantes usando imputación por la media.
7. Si usted tiene que buscar clusters en un conjunto de datos que tiene muchas columnas categóricas que a su vez cada una tiene alta cardinalidad (es decir, muchos valores distintos), qué problema puede surgir si uno usa One Hot Encoding? Explique el problema justificando su respuesta.
8. Suponga que usted entrenó un modelo predictivo usando validación de resustitución. Qué es probable que suceda si a futuro usted quiere predecir nuevos datos con mayor diversidad de con los que entrenó?
9. 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.
10. El modelo de regresión logística es un modelo simple que no siempre obtiene la mayor performance respecto a otros algoritmos más complejos como las redes neuronales. Sin embargo, la regresión logística es ampliamente utilizada. Qué ventajas posee la regresión logística?