Taller de Machine Learning

Natalia Avendaño

• ¿Qué es un árbol de decisión y para qué se utiliza en análisis de datos?

Son algoritmos estadísticos de aprendizaje automático que se utiliza para procesar grandes volúmenes datos que nos permiten crear modelos predictivos a través de la clasificación según ciertas características o propiedades que tengan los datos.

• ¿En qué se diferencia Random Forest de un solo árbol de decisión?

Estructura: Un árbol es un modelo único; Random Forest combina múltiples árboles

Predicción: El árbol sigue un camino único; Random Forest promedia/vota entre todos sus árboles.

Sobreajuste: Los árboles individuales se sobreajustan fácilmente; Random Forest lo reduce significativamente.

Aleatoriedad: El árbol usa todos los datos y características; Random Forest introduce aleatoriedad en muestras y características.

Interpretabilidad: El árbol es fácil de interpretar; Random Forest es más complejo pero más preciso.

Robustez: El árbol es sensible a cambios en datos; Random Forest es más estable y generaliza mejor.

• Explique brevemente cómo funciona la regresión logística y en qué tipo de problemas se aplica.

La regresión logística usa la función sigmoide para transformar una combinación lineal de variables en probabilidades entre 0 y 1. En lugar de predecir valores directos, calcula la probabilidad de que una observación pertenezca a una clase específica.

Tipos de problemas donde se aplica:

- Determinar si un email es spam o no, si un paciente tiene una enfermedad, si un cliente comprará un producto

- Categorizar texto en múltiples temas, diagnóstico médico entre varias enfermedades
- Evaluar probabilidad de default en créditos, riesgo de accidentes
- **Predecir** respuesta a campañas publicitarias, segmentación de cliente

• ¿Cuál es la principal diferencia entre regresión lineal y regresión logística?

Regresión lineal predice **números continuos** (como precios, temperaturas, edades)
Regresión logística predice **categorías** o **probabilidades** (como sí/no, spam/no spam)

• Mencione dos medidas estadísticas básicas que se utilizan para describir un conjunto de datos.

Media (promedio) y Desviación estándar.

• ¿Qué es un kernel en el contexto de métodos de aprendizaje automático y cuál es su función principal?

Un kernel es como un "traductor matemático" que toma datos que no se pueden separar fácilmente y los transforma a un espacio donde sí se pueden separar.

Función principal: Permite que algoritmos simples (que solo pueden trabajar con líneas rectas) resuelvan problemas complejos (con curvas y formas irregulares).

• ¿Qué es DBScan y qué tipo de agrupamiento realiza?

DBSCAN (Density-Based Spatial Clustering) es un algoritmo de clustering que agrupa datos basándose en la **densidad** - es decir, busca zonas donde hay muchos puntos juntos.

Tipo de agrupamiento que realiza:

Clustering basado en densidad:

- No necesitas decirle cuántos grupos quieres (como en K-means)
- Puede encontrar clusters de formas irregulares (no solo círculos)
- Identifica automáticamente puntos atípicos