Análisis Predictivo y Gestión de Datos

Sesión 4: Primer modelo supervisado - KNN

Oscar Leonardo Rincón León

24 de abril de 2025

Objetivo de la sesión

- Implementar un modelo de clasificación supervisado básico usando el algoritmo K-Nearest Neighbors (KNN).
- Comprender el funcionamiento, limitaciones y requerimientos de este algoritmo.

¿Qué es el aprendizaje supervisado?

- Técnica de aprendizaje automático en la que se entrena un modelo con datos de entrada X y salidas conocidas y.
- El objetivo es aproximar una función f tal que $y \approx f(X)$.
- Se busca minimizar el error entre la predicción del modelo y la salida real: $y = f(X) + \varepsilon$.
- Ejemplos:
  - Predecir si un estudiante abandonará la universidad (clasificación).
  - Estimar el ingreso mensual en función de edad y educación (regresión).

Clasificación vs Regresión

- **Clasificación:** la variable objetivo es una etiqueta o categoría (ej. sí/no, tipo A/B/C).
- **Regresión:** la variable objetivo es continua (ej. ingresos, edad, temperatura).
- KNN puede ser usado para ambos, pero hoy lo aplicamos a clasificación binaria.

¿Qué es el algoritmo KNN? (1/2)

- KNN significa **K-Nearest Neighbors** (K vecinos más cercanos).
- Es un algoritmo de clasificación supervisado que se basa en la distancia entre puntos.
- No construye un modelo explícito, sino que decide al momento de hacer una predicción.
- Se le conoce como un modelo "perezoso" porque no entrena previamente.

¿Qué es el algoritmo KNN? (2/2)

- Para predecir la clase de un nuevo dato, el algoritmo:
  1. Calcula la distancia entre el nuevo punto y todos los puntos del entrenamiento.
  2. Selecciona los k vecinos más cercanos.
  3. Asigna la clase más común entre ellos.
- **Ejemplo:** Si una persona tiene edad e ingreso similares a 5 individuos, y 3 de ellos tienen alta conectividad, se clasificará como " alta conectividad "si $k = 5$.

¿Qué significa entrenar un modelo?

- Entrenar un modelo significa ajustar sus parámetros para que aprenda a predecir una salida y a partir de una entrada X.
- En modelos como regresión logística o redes neuronales, se optimizan parámetros internos.
- En KNN no se entrena un modelo como tal: se almacenan los datos y se calcula la predicción directamente al comparar con los vecinos.
- Por eso se le conoce como modelo "perezoso".

¿Por qué se dice que es un modelo “perezoso”?

- No realiza un entrenamiento intensivo.
- Almacena el conjunto de entrenamiento tal cual.
- Calcula las distancias y decide en el momento de la predicción.
- Ventaja: simple y directo. Desventaja: lento con grandes volúmenes.

Sensibilidad a escalas y ruido

- En KNN, la distancia entre puntos determina qué tan similares son dos casos.
- Si una variable tiene una escala mucho mayor que otra, dominará la distancia total.
- Ejemplo: si ingreso va de 0 a 10.000 y edad de 0 a 100, ingreso tendrá mayor peso si no se escalan.
- El **ruido** (datos atípicos o mal registrados) puede sesgar el resultado si k es pequeño.
- Soluciones:
  - Aplicar escalado de variables.
  - Probar diferentes valores de k para mayor robustez.

¿Cuándo se usa KNN?

- Exploración inicial de patrones.
- Clasificación binaria o multiclase con datos limpios y balanceados.
- Casos donde no se quiere asumir una forma funcional compleja.
- Proyectos con recursos computacionales limitados.

Preprocesamiento necesario para KNN

- Eliminar valores faltantes.
- Codificar variables categóricas.
- Escalar las variables numéricas (MinMaxScaler o StandardScaler).
- Dividir datos en entrenamiento y prueba.

Flujo de trabajo práctico

- Cargar dataset limpio.
- Separar variables predictoras (X) y variable objetivo (y).
- Dividir en entrenamiento y prueba.
- Aplicar escalado.
- Entrenar y evaluar el modelo con KNN.

Cierre

- KNN es un algoritmo simple pero efectivo en tareas de clasificación básica.
- Su desempeño depende fuertemente del preprocesamiento y la elección de k.
- En la próxima sesión lo compararemos con otros modelos más sofisticados.