KNN

Minería de Datos (CC3074) - 2026

KNN

Semestre 01, 2026

Problemática

- Se tiene un conjunto de datos con clases conocidas.
- Llega una nueva instancia sin clase.
- ¿Cómo se determina a qué clase pertenece?

Ejemplo

Un banco tiene registros de clientes: edad, monto de crédito y si pagaron o no.

Llega un nuevo solicitante.

¿Se le otorga el crédito?

Intuición

Las personas se parecen a quienes las rodean.

- Un cliente con perfil similar a quienes pagaron → probablemente pagará.
- Un cliente rodeado de deudores → mayor riesgo.

KNN explota esa idea de proximidad.

"Dime con quién andas y te diré quién eres"

Definición

KNN es un modelo de aprendizaje supervisado.

Se utiliza para clasificación y regresión.

No aprende una función explícita: simplemente memoriza los datos de entrenamiento.

¿Cómo funciona?

1. Se calcula la distancia entre la nueva instancia y todos los datos conocidos.
2. Se seleccionan los K vecinos más cercanos.
3. Se toma una decisión basada en esos vecinos.

El parámetro K

K define cuántos vecinos se consideran.

- K = 3 → se toman los 3 puntos más cercanos.
- K = 5 → se toman los 5 puntos más cercanos.

La elección de K afecta directamente el resultado del modelo.

Clasificación con KNN

En clasificación, cada vecino vota por una clase.

La clase ganadora es la predicción.

Ejemplo visual

Nueva instancia: cliente de 43 años, crédito de Q320,000.

- Vecino 1 → pagó
- Vecino 2 → pagó
- Vecino 3 → no pagó

Resultado: pagará (2 votos contra 1).

K impar

Se recomienda usar valores de K impares.

Así se evitan empates en la votación.

| K | Empate posible | | - | ---------- | | 2 | Sí | | 3 | No | | 4 | Sí | | 5 | No |

Regresión con KNN

En regresión, no se vota por una clase.

Se calcula el promedio de los valores de los K vecinos.

Ejemplo

Nueva instancia: casa de 120 m².

- Vecino 1 → Q850,000
- Vecino 2 → Q920,000
- Vecino 3 → Q780,000

Predicción: (850,000 + 920,000 + 780,000) / 3 = Q850,000

Clasificación vs Regresión

- Clasificación: Se toma la clase más frecuente
- Regresión: Se promedia el valor numérico

Distancia

KNN necesita medir qué tan cercanos están los puntos.

La medida más común es la distancia euclidiana.

Distancia Euclidiana

Para dos puntos A y B con atributos X e Y:

`\(d = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}\)`

Es la distancia en línea recta entre dos puntos.

Escala

Si una variable va de 18 a 60 y otra va de 100,000 a 600,000:

La segunda domina el cálculo de distancia.

El modelo le da más peso a esa variable sin razón conceptual.

Escalado de datos

Antes de aplicar KNN, los datos deben escalarse.

Todas las variables deben estar en el mismo rango.

Normalización Min-Max

Transforma cada valor al rango [0, 1].

```
$$x_{\text{norm}} = \dfrac{x - x_{\min}}{x_{\max} - x_{\min}}$$
```

- El valor mínimo se convierte en 0.
- El valor máximo se convierte en 1.
- El resto se distribuye proporcionalmente.

Ejemplo

| Variable | Valor original | Valor escalado | | -------- | -------------- | -------------- | | Edad 18 | 18 años | 0.00 | | Edad 39 | 39 años | 0.50 | | Edad 60 | 60 años | 1.00 |

Después del escalado, ambas variables tienen el mismo peso.

Elección de K

K es el hiperparámetro más importante de KNN.

No se aprende de los datos: se define antes de entrenar.

Efecto de K

| K pequeño | K grande | | --------- | -------- | | Modelo muy flexible | Modelo más suave | | Sensible al ruido | Menos sensible al ruido | | Puede sobreajustar | Puede subajustar |

Regla general

Una referencia inicial muy usada es:

```
$ K \approx \sqrt{n} $
```

Donde n es el número de instancias de entrenamiento.

Con 200 datos → K ≈ 14. Se redondea al impar más cercano: K = 15.

La mejor práctica

Probar varios valores de K y elegir el que minimice el error en validación.

No existe un valor universalmente correcto.

Validación

Un modelo que funciona bien en entrenamiento no garantiza buen desempeño en datos nuevos.

Se divide el conjunto de datos en dos partes:

- Entrenamiento → para construir el modelo.
- Prueba → para evaluar el modelo.

Métricas de Clasificación

Exactitud (Accuracy)

Proporción de predicciones correctas sobre el total.

```
$ \text{Accuracy} = \frac{\text{predicciones correctas}}{\text{total de instancias}} $
```

- Fácil de interpretar.
- Puede ser engañosa cuando las clases están desbalanceadas.

Matriz de Confusión

Resume todos los resultados del modelo.

| | Predicho: Sí | Predicho: No | | ---------- | ---------- | ---------- | | Real: Sí | VP | FN | | Real: No | FP | VN |

- VP → Verdadero Positivo: predijo Sí y era Sí.
- VN → Verdadero Negativo: predijo No y era No.
- FP → Falso Positivo: predijo Sí pero era No.
- FN → Falso Negativo: predijo No pero era Sí.

Precisión y Exhaustividad

Precisión: De los que se predijeron como positivos, ¿cuántos realmente lo eran?

`$\text{Precisión} = \frac{VP}{VP + FP} $`

Exhaustividad (Recall): De los que realmente eran positivos, ¿cuántos se identificaron?

`$\text{Recall} = \frac{VP}{VP + FN} $`

F1-Score

Combina precisión y exhaustividad en un solo número.

`$ F1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} $`

Se usa cuando se busca un balance entre ambas métricas.

Métricas de Regresión

MAE — Error Absoluto Medio

Promedio de los errores en valor absoluto.

- Está en las mismas unidades que la variable.
- MAE = 5,000 → en promedio se erra en Q5,000.

RMSE — Raíz del Error Cuadrático Medio

Similar al MAE, pero penaliza más los errores grandes.

- Útil cuando los errores grandes son especialmente problemáticos.

R^2

Indica qué proporción de la variabilidad se explica con el modelo.

- $R^2 = 0.85$ → el modelo explica el 85% de la variación.
- Va de 0 a 1. Más cercano a 1 es mejor.

Casos de uso

| Dominio | Aplicación | | ------- | ---------- | | Banca | Clasificar clientes como pagadores o deudores | | Medicina | Diagnóstico basado en síntomas similares | | Comercio | Sistemas de recomendación | | Biología | Clasificación de especies por características | | Inmobiliario | Estimación de precios de propiedades |

¿Cuándo se usa KNN?

- Se tienen pocos atributos (pocas columnas).
- Los datos están bien escalados.
- Se necesita un modelo simple y explicable.
- No se cuenta con muchos datos (KNN puede funcionar con conjuntos pequeños).

Ventajas

- Simple e intuitivo: la lógica es fácil de explicar a personas sin formación técnica.
- No paramétrico: no se asume ninguna distribución de los datos.
- Versátil: se aplica tanto a clasificación como a regresión.
- Sin entrenamiento explícito: el modelo simplemente almacena los datos.

Limitaciones

- Sensible a la escala: se requiere normalización obligatoria.
- Lento con muchos datos: la distancia se calcula contra cada punto conocido.
- Maldición de la dimensionalidad: el rendimiento se degrada con muchas variables.
- Requiere almacenamiento: todos los datos de entrenamiento deben estar disponibles.

Maldición de la dimensionalidad

Con pocas variables, los puntos están relativamente cerca.

Al agregar más variables, el espacio crece exponencialmente.

Los puntos se alejan entre sí → la noción de "vecino cercano" pierde significado.