Parte 2
Modelos no paramétricos predictivos: KNN, árboles de clasificación y regresión y modelos de predicción basados en árboles

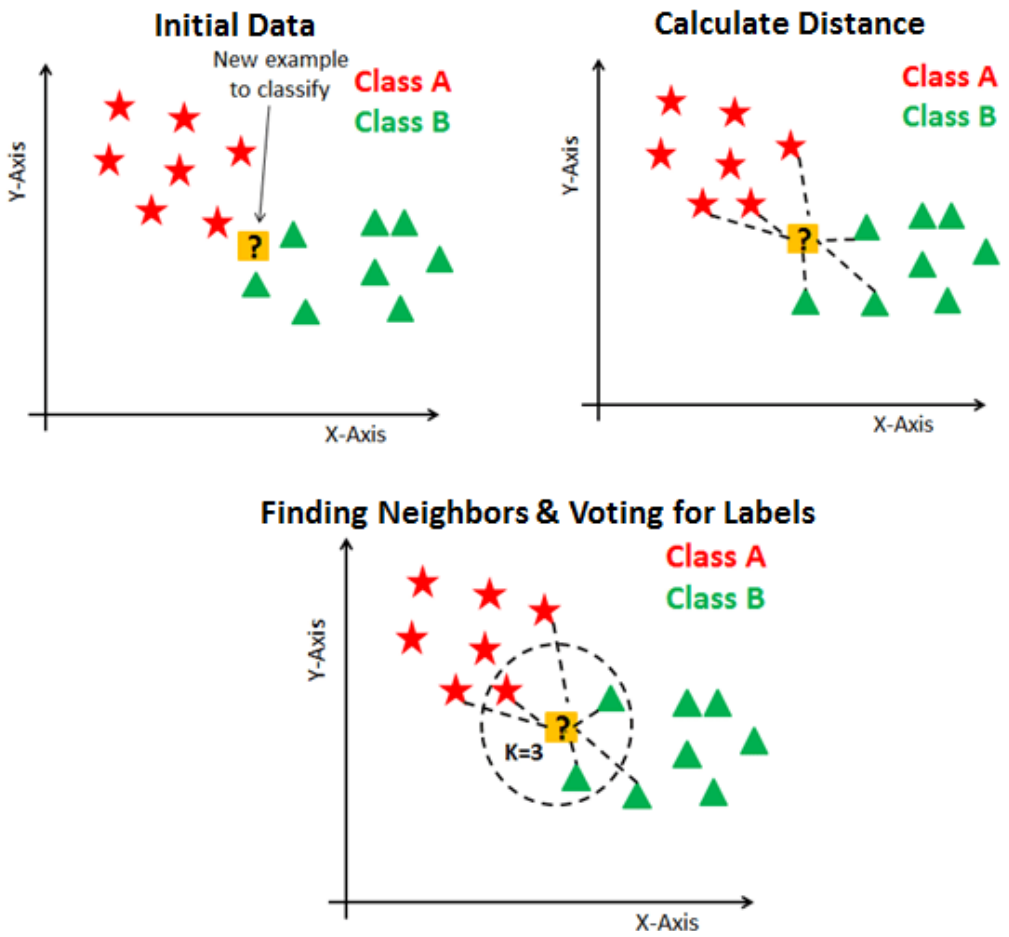
Aprendizaje Estadístico

Junio 2025

Aida Calviño

El modelo del vecino más próximo (más conocido por sus siglas en inglés, KNN) es un modelo de predicción muy **sencillo**, que se basa en la idea de que observaciones con **valores parecidos** de las variables input deben tomar también valores similares de la variable objetivo.

De esta forma, a la hora de predecir el valor de una variable objetivo para una determinada observación, se buscan las k **observaciones más próximas** y se obtiene el valor de predicción como la **media** o la **proporción** de los valores de la variable objetivo, según sea esta cuantitativa o cualitativa, respectivamente.



AJUSTE DEL MODELO

- **¿Cuántos vecinos se deben tener en cuenta?** El número óptimo de vecinos **depende del conjunto de datos** que se vaya a utilizar. No obstante, hay que tener en cuenta que:
  - Valores pequeños de k dan lugar a modelos con sesgo pequeño pero gran varianza.
  - Valores grandes de k dan lugar a modelos con gran sesgo pero varianza pequeña.
  - Lo habitual es probar un conjunto de valores de k y seleccionar el que **menor error** produzca aplicando validación cruzada (repetida o no).
- **¿Cómo se mide la proximidad de las vecinos?** La proximidad de los vecinos se mide a partir de la **distancia entre las observaciones**.
  - Habitualmente se utiliza la **distancia euclídea** pero, dependiendo de la tipología de las variables, se pueden utilizar otros tipos de distancias (Mahalanobis, Manhattan, Jackard, etc.).
  - Se recomienda **estandarizar las variables input** previamente para que todas tengan el mismo peso a la hora de calcular dichas distancias.
  - Si hay variables input categóricas y se desea aplicar alguna distancia "numérica" (como la euclídea), es necesario convertirlas previamente en variables numéricas, lo que se puede conseguir creando variables *dummy*.

VENTAJAS

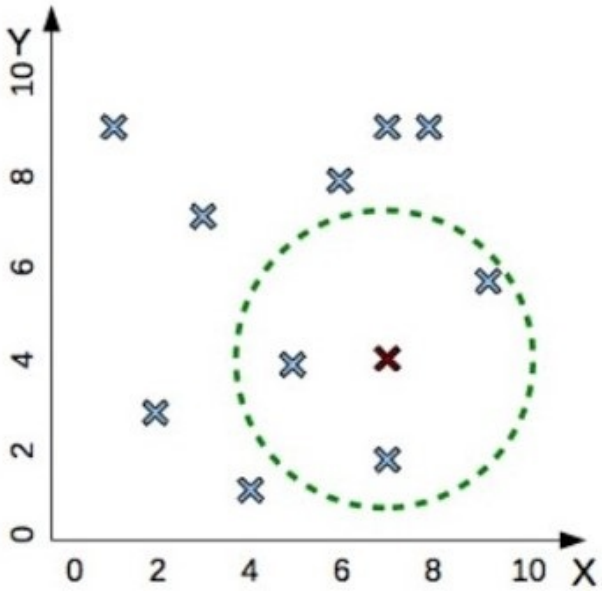
- El modelo es sencillo y fácil de comprender.
- No hay asunciones teóricas.
- Ofrece buenos resultados cuando la estructura de los datos es compleja, lo que incluye relaciones no lineales (ni monótonas) e interacciones.

DESVENTAJAS

- No se puede explicar ni conocer la importancia de las variables.
- No se puede aplicar si existen valores ausentes, por lo que hay que tratarlos previamente.
- Es muy sensible a la presencia de variables “inutiles”, por lo que se hace necesario llevar a cabo una selección de variables previa.
- Es un método computacionalmente intenso.

Why is kNN slow?

What you see



Find nearest neighbors
of the testing point (red)

What algorithm sees

- Training set:

  {(1,9), (2,3), (4,1),
  (3,7), (5,4), (6,8),
  (7,2), (8,8), (7,9), (9,6)}

- Testing instance:

  (7,4)

- Nearest neighbors?

  compare one-by-one
  to each training instance

- n comparisons

SELECCIÓN DE VARIABLES

Como acabamos de mencionar, los resultados del modelo KNN dependen mucho de las variables consideradas a la hora de calcular las distancias. Más concretamente, la presencia de variables explicativas no relacionadas con la dependiente puede dar lugar a **sobreajuste**.

Por ello, resulta recomendable llevar a cabo una **selección de variables** que permita eliminar del modelo aquellas variables explicativas poco o nada relacionadas con la variable a predecir.

Desgraciadamente, los **métodos de selección de variables vistos para regresión no son aplicables** en este caso puesto que no se dispone de parámetros y, por ende, no es posible calcular los estadísticos AIC/BIC que rigen el proceso de selección.

Alternativamente, se puede recurrir al algoritmo ***RFE (Recursive Feature Elimination)***, el cual puede verse como una generalización de los métodos *forward/backward* que puede aplicarse sobre cualquier tipo de modelo.

SELECCIÓN DE VARIABLES: ALGORITMO *RFE*

Este algoritmo persigue generar una secuencia de modelos, que puedan evaluarse posteriormente mediante **validación cruzada**, **introduciendo iterativamente aquellas variables explicativas con mayor poder predictivo** (o, visto de otra manera, eliminando del modelo "completo" aquellas variables explicativas con menor poder predictivo).

El orden anterior se obtiene previamente a la aplicación del algoritmo mediante el R^2 **o el área bajo la curva ROC** (*pairwise* si la variable dependiente no es binaria) obtenida a partir de cada variable independiente de manera separada (por lo que no se tienen en cuenta las posibles interacciones entre variables).

Este algoritmo entra dentro de los denominados *greedy* pues consiste en **seleccionar las mejores variables** en cada iteración sin evaluar otras combinaciones. Por ello, no se asegura que se obtenga la mejor combinación posible (pues no es posible evaluar todas las posibles combinaciones de variables independientes) pero sí se asegura obtener un **resultado suficientemente bueno** (es lo que se llama una heurística).

Una vez encontrada la combinación "óptima" de variables, se debe **determinar el parámetro k óptimo para el modelo**.

ANÁLISIS DE LAS VARIABLES PREDICTORAS

Aunque este tipo de modelo no es interpretable, existe una herramienta que permite conocer la **influencia de las variables *input* sobre la variable objetivo**:

- Gráficos de dependencia parcial (*Partial Dependence Plots-PDP*): Los PDP proporcionan una **forma visual de entender el efecto marginal** que una o varias variables tienen sobre las predicciones del modelo, manteniendo constantes las demás variables.
  - Se toma el conjunto de datos original y se **modifican iterativamente** los valores de la(s) variables(s) de interés según una rejilla de valores ya observados, pero manteniendo el resto.
  - Se predice la variable objetivo para estas observaciones modificadas y se calcula la **predicción media** para cada valor (o combinación de valores) evaluado.
  - Estas predicciones **se representan en un gráfico**, lo que permite observar el efecto promedio que tiene la modificación de dicha(s) variable(s) predictora(s) en la objetivo.