Clasificación no lineal

Inteligencia Artificial

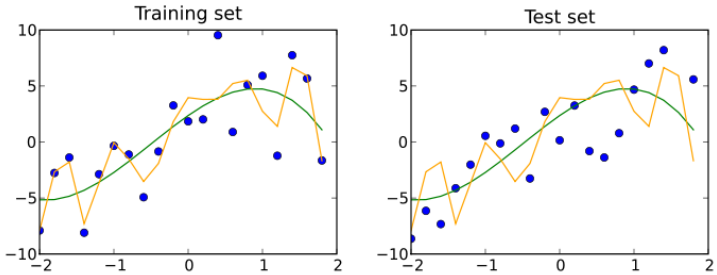
Marco Teran

2021 - Bogotá

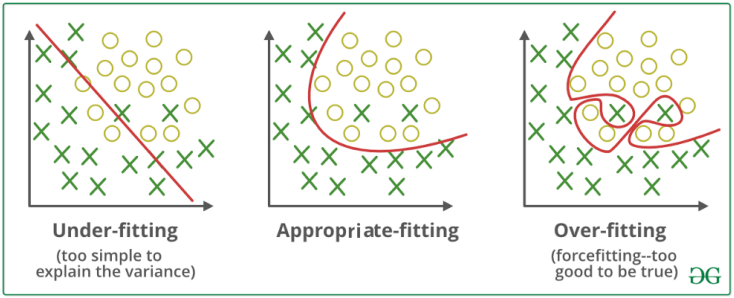
Contenido

Complejidad de un modelo

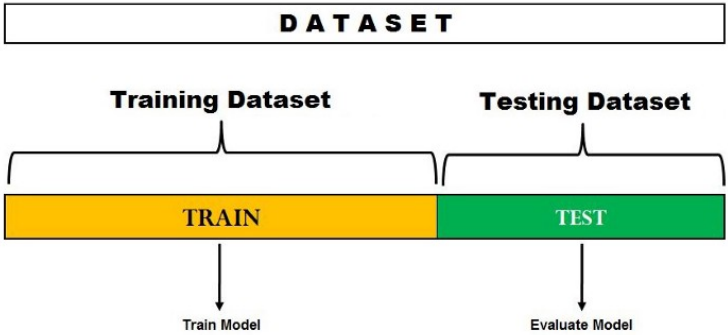
¿Qué es Underfitting y Overfitting?



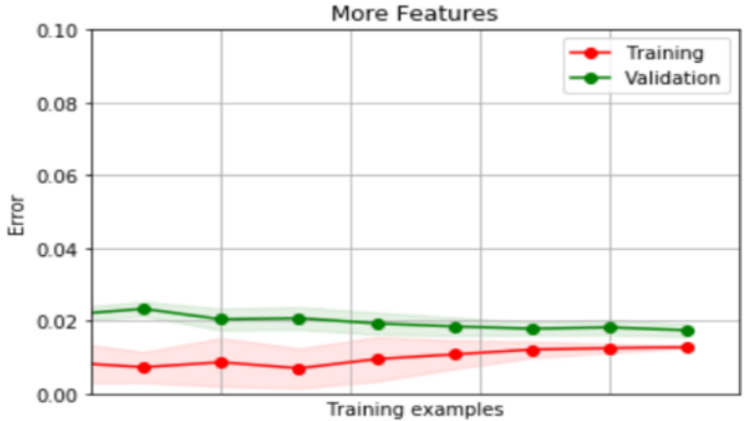
Ajuste


Under-fitting
(too simple to explain the variance)
Appropriate-fitting
Over-fitting
(forcefitting--too good to be true)

Ajuste

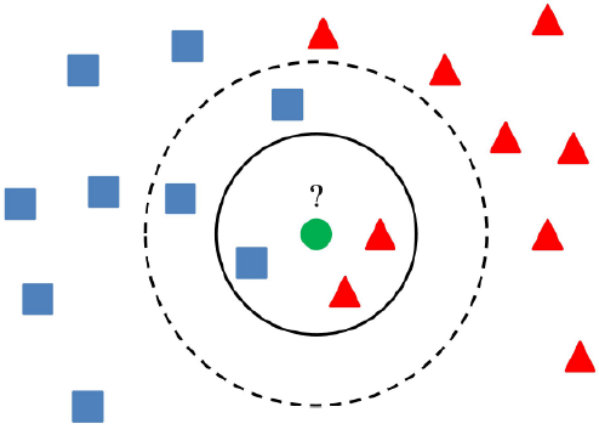

DATASET
Training Dataset
Testing Dataset
TRAIN
TEST
Train Model
Evaluate Model

Complejidad del modelo



K-Nearest-Neighbor

¿Cómo funciona kNN?



K-Nearest-Neighbor

- K-Nearest-Neighbor es un algoritmo basado en instancia de tipo supervisado de Machine Learning.
- Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos)
- Al ser un método sencillo, es ideal para introducirse en el mundo del Aprendizaje Automático
- Sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basado en esa clasificación.

¿Qué es el algoritmo k-Nearest Neighbor?

Es un método que simplemente busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean:

- **Supervisado:** esto -brevemente- quiere decir que tenemos etiquetado nuestro conjunto de datos de entrenamiento, con la clase o resultado esperado dada "una fila" de datos.
- **Basado en Instancia:** Esto quiere decir que nuestro algoritmo no aprende explícitamente un modelo (como por ejemplo en Regresión Logística o árboles de decisión). En cambio memoriza las instancias de entrenamiento que son usadas como "base de conocimiento" para la fase de predicción

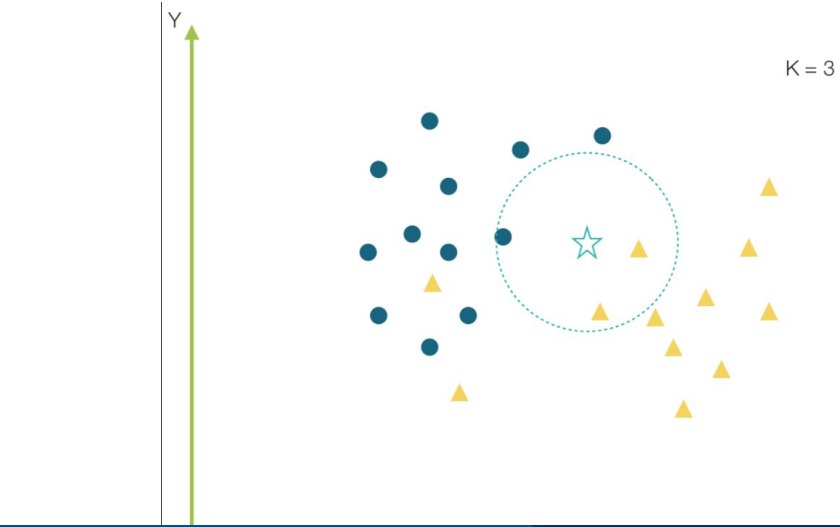
¿Dónde se aplica k-Nearest Neighbor?

Aunque sencillo, se utiliza en la resolución de multitud de problemas, como en sistemas de recomendación, búsqueda semántica y detección de anomalías.

Pros y contras

- Como pros tiene sobre todo que es sencillo de aprender e implementar.
- Tiene como contras que utiliza todo el dataset para entrenar "cada punto" y por eso requiere de uso de mucha memoria y recursos de procesamiento (CPU). Por estas razones kNN tiende a funcionar mejor en datasets pequeños y sin una cantidad enorme de features (las columnas).

¿Cómo funciona kNN?



¿Cómo funciona kNN?

1. Calcular la distancia entre el item a clasificar y el resto de items del dataset de entrenamiento.
2. Seleccionar los "k" elementos más cercanos (con menor distancia, según la función que se use)
   - Las formas más populares de "medir la cercanía" entre puntos son la distancia Euclidiana (la "de siempre") o la Cosine Similarity (mide el ángulo de los vectores, cuanto menores, serán similares).
3. Realizar una "votación de mayoría" entre los k puntos: los de una clase/etiqueta que «dominen» decidirán su clasificación final.

¿Cómo funciona kNN?

Distancia Euclidiana

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Distancia Manhattan

$$\sum_{i=1}^{k} |x_i - y_i|$$

Distancia Minkowski

$$[\sum_{i=1}^{k} (|x_i - y_i|)^4]^{\frac{1}{4}}$$

Muchas gracias por su atención

¿Preguntas?

Contacto: Marco Teran
webpage: marcoteran.github.io/