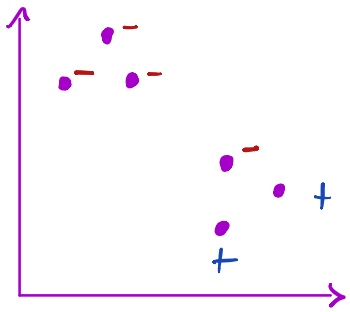
KNN

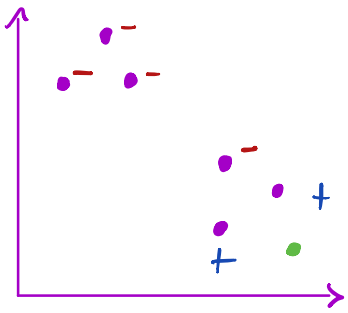
Después de estudiar el regresor logístico, es buena idea tener entre nuestras herramientas un clasificador no lineal.

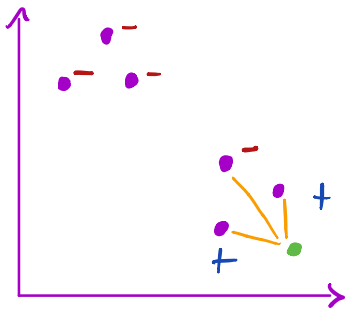
K-nearest-neighbors es un clasificador supervisado basado en distancias. Al entrenar guardamos todos los puntos con sus etiquetas, y al predecir, vemos quienes son los K vecinos más cercanos, y cada uno vota con su etiqueta. La etiqueta con más votos gana.

Ejemplo. Supongamos el siguiente dataset

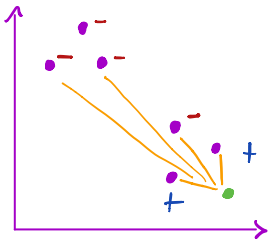


Si queremos clasificar el punto verde en base a los 3 puntos mas cercanos:





Será clasificado como positivo, pero si ahora usamos 5 puntos:



Será clasificado como negativo.

Ojo, este algoritmo tiene variantes:

- El concepto de distancia no siempre es el de distancia euclidiana.
- Pueden pesar más los votos de puntos más cercanos
- Este modelo puede ser usado para clasificación multiclase y datasets de más dimensiones.

Curse of Dimensionality

Este concepto hace referencia a a los fenómenos ocurridos al trabajar con espacios de más altas dimensiones.

Uno de los principales problemas es que en altas dimensiones los dateset tienden a ser más "sparse".