Intuición sobre KNN

Luis R. Furlán

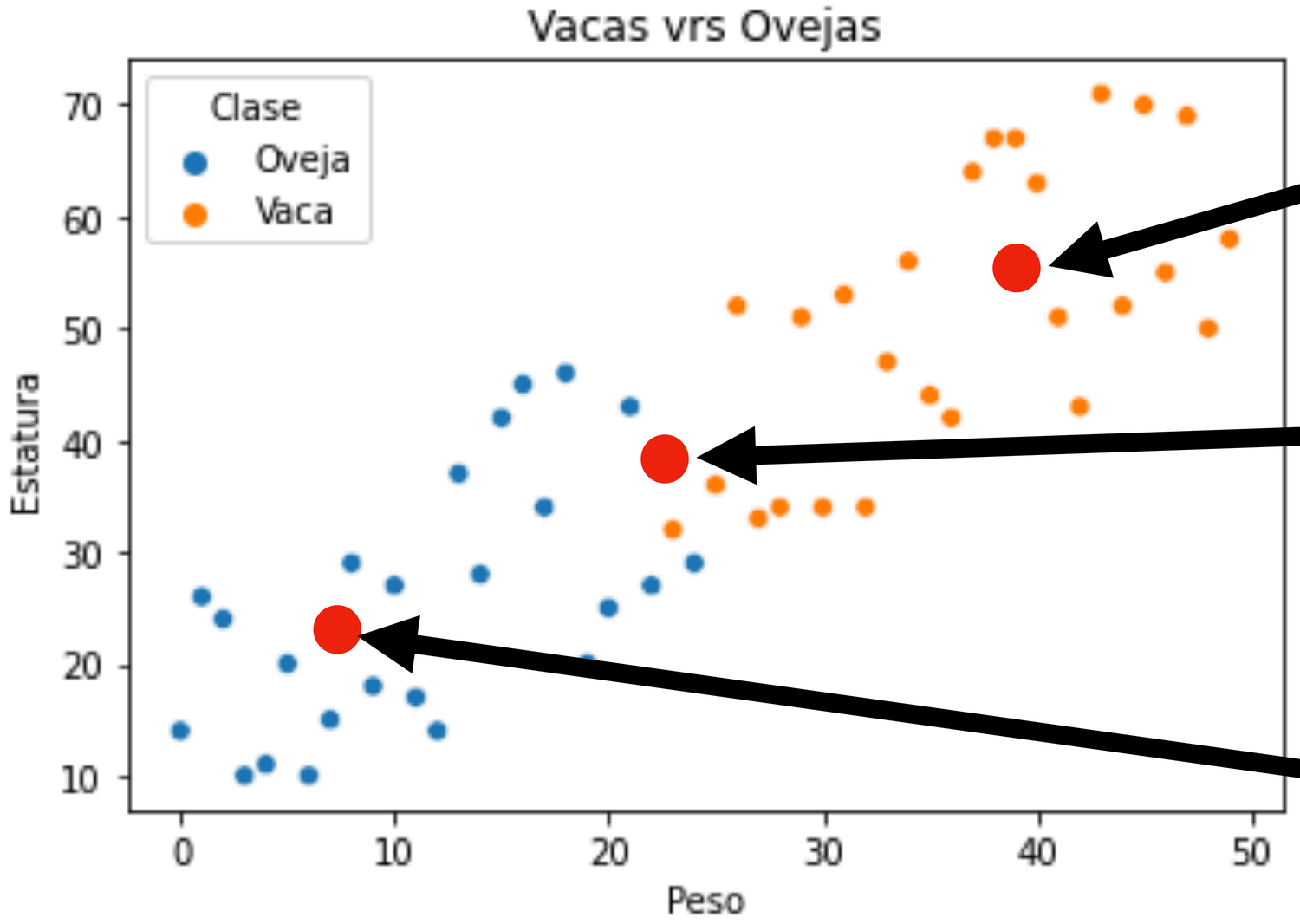
Agosto de 2021

KNN

KNN (siglas en inglés de K vecinos más próximos) es un algoritmo de clasificación que opera bajo un principio muy simple.

Se puede ver mejor por medio de un ejemplo.

Imaginemos que tenemos datos ficticios sobre ovejas y vacas, conteniendo estatura y peso.



Punto nuevo:
Es Vaca u Oveja?

Punto nuevo:
Es Vaca u Oveja?

Punto nuevo:
Es Vaca u Oveja?

KNN

Algoritmo de entrenamiento:

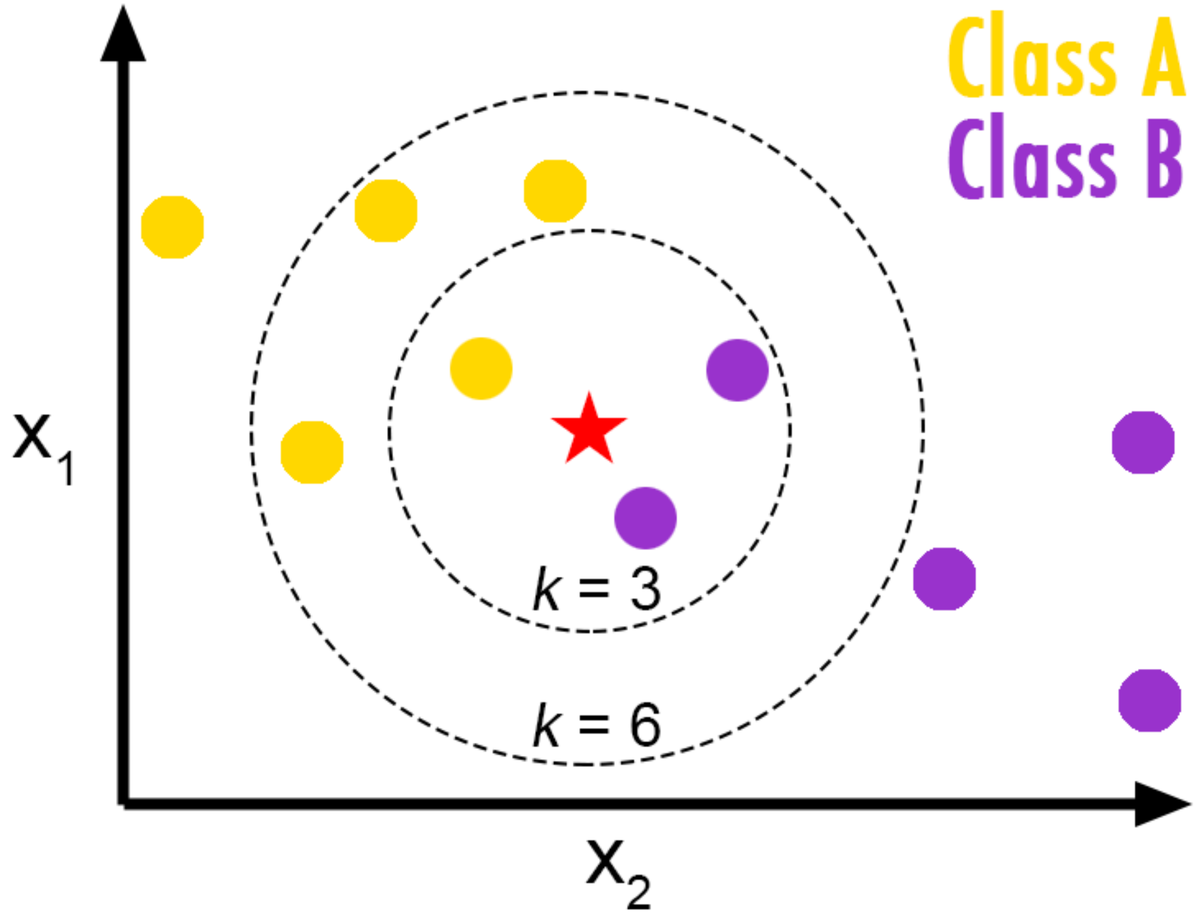
1. Almacenar todos los datos

Algoritmo de predicción:

1. Calcular las distancias desde x hacia todos los puntos de los datos
2. Ordenar los puntos en los datos por distancia (ascendente)
3. Predecir la etiqueta de mayoría de los “K” puntos más cercanos

KNN

La selección de un valor de K afecta la asignación de clase a un punto nuevo;

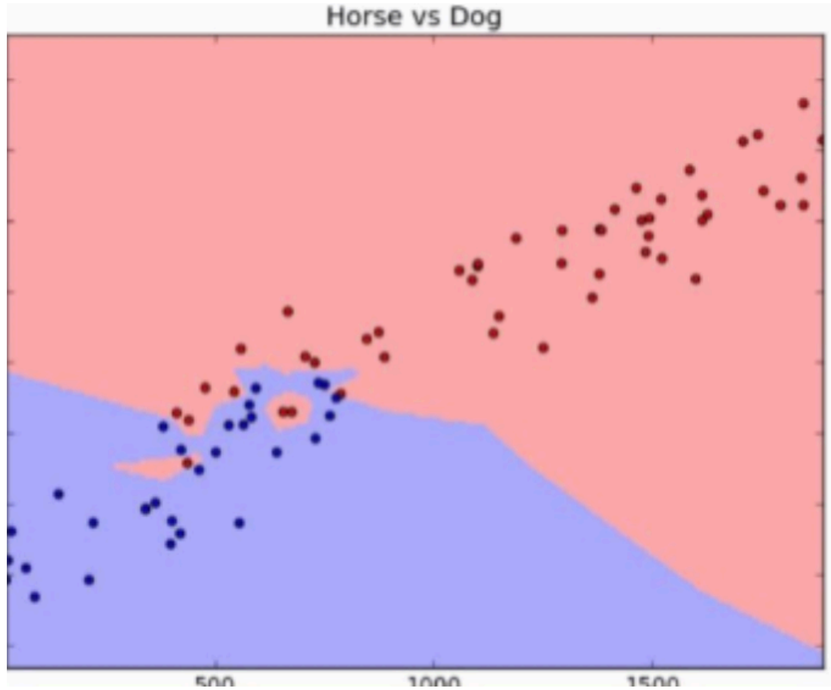


KNN

La selección de un valor de K afecta la asignación de clase a un punto nuevo;

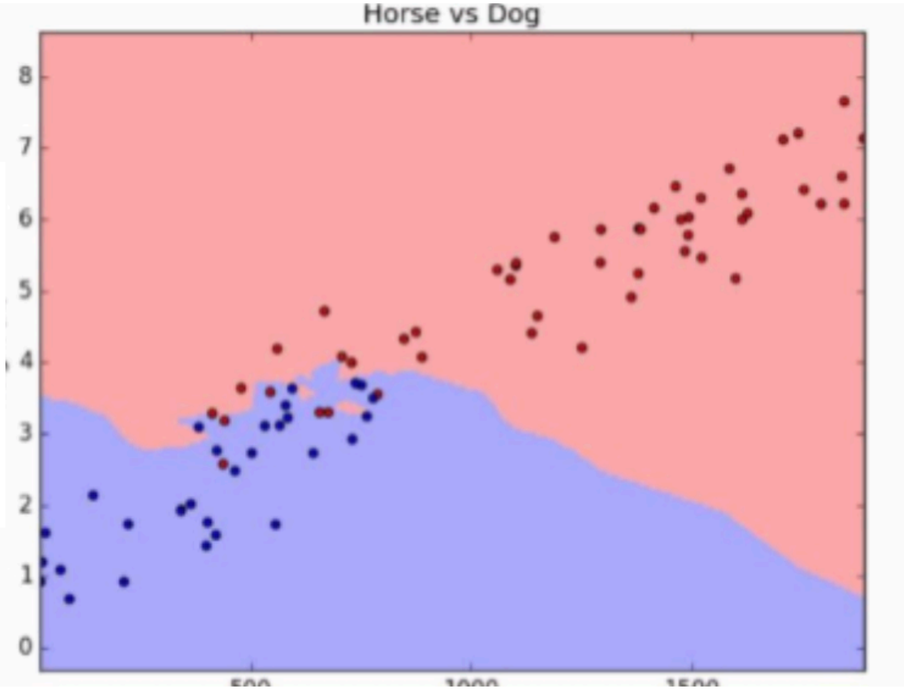
K = 1

Vacas vrs Ovejas



K = 5

Vacas vrs Ovejas

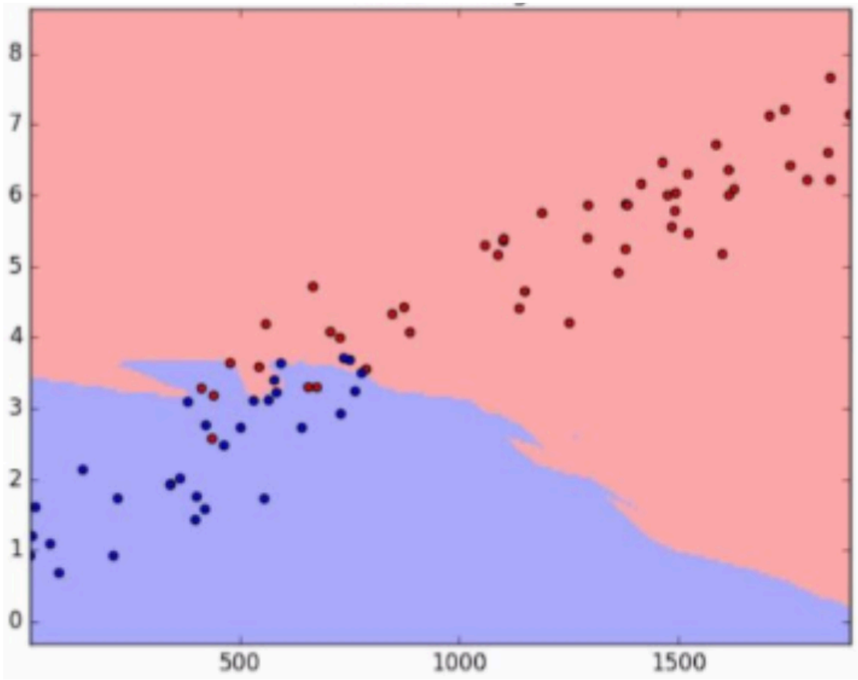


KNN

La selección de un valor de K afecta la asignación de clase a un punto nuevo:

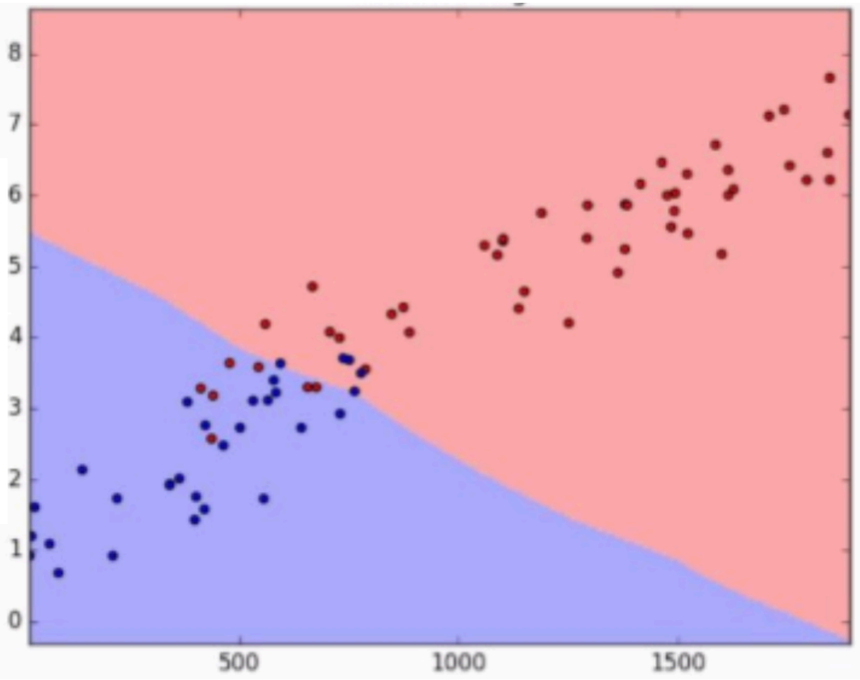
K = 10

Vacas vrs Ovejas



K = 50

Vacas vrs Ovejas



KNN

Pros

- Muy simple
- El entrenamiento es trivial
- Trabaja con cualquier número de clases
- Es fácil agregar más datos
- Pocos parámetros
  - K
  - Métrica de distancia

KNN

Contras

- Costo alto de predicción (peor para conjuntos grandes de datos)
- No es muy bueno con datos de alta dimensionalidad
- La variables (features) categóricas no funcionan bien

KNN

Ejemplo con Python

Una prueba común que le ponen a un Científico de Datos cuando se está entrevistando para un puesto nuevo, es que le den datos anonimizados y le piden que los clasifique sin conocer el contexto de los datos.

Simularemos un escenario similar donde nos dan un conjunto de datos “clasificado”, en donde no sabemos qué es lo que las columnas representan, pero hay que usar KNN para clasificarlo!