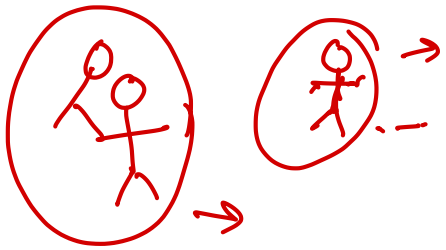
Machine Learning

K-Nearest Neighbours

Overview

- The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems
- However, it is more widely used in classification problems in the industry
- It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection
- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.



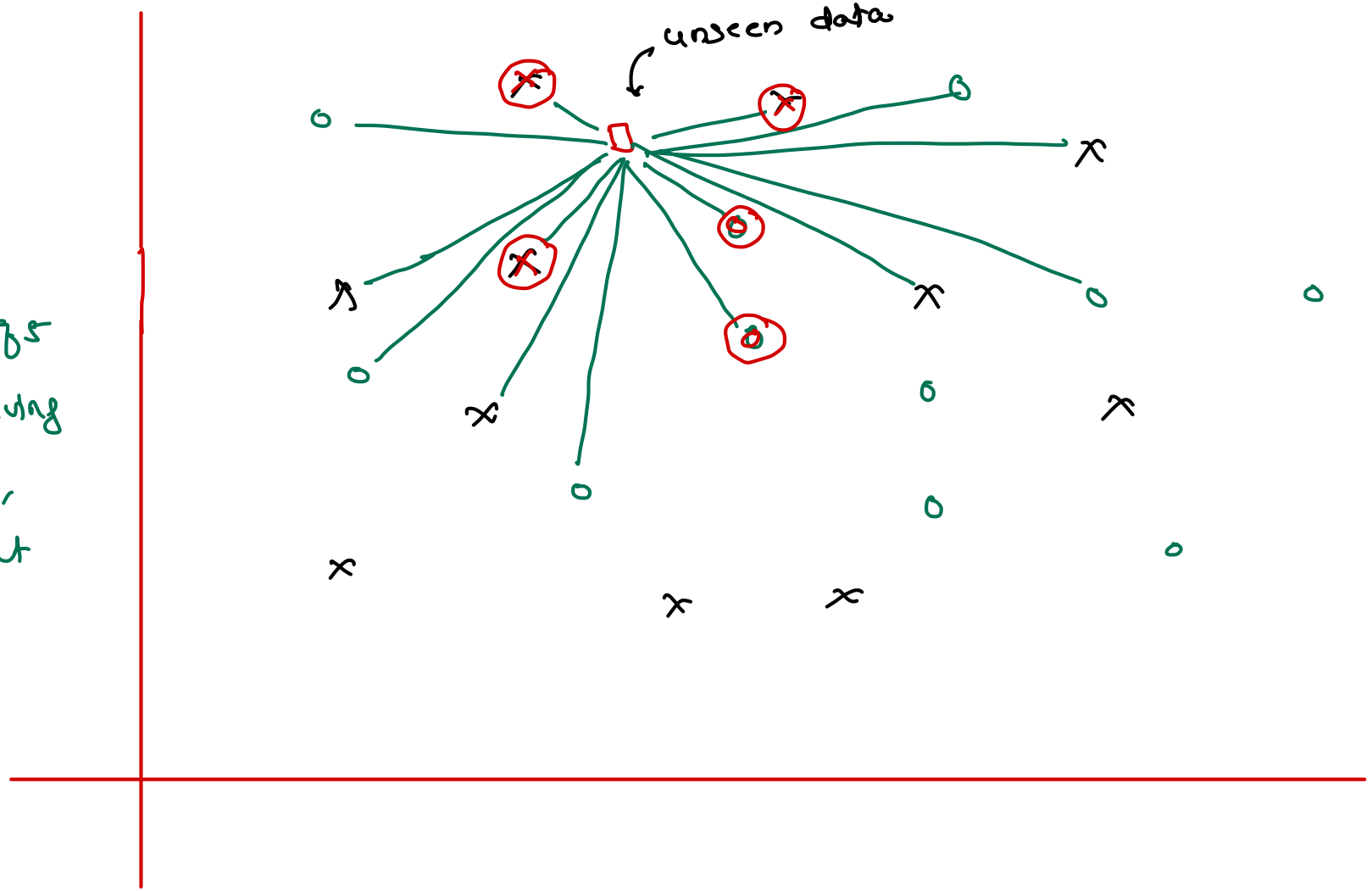
How does it work?

K = 5

x : 3 →

o : 2

because 3 outof 5 paitients having heart attack, newer patient will get an attack



o → not having any heart disease

x → has a heart disease

How does it work ?

- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function
- If K = 1, then the case is simply assigned to the class of its nearest neighbor.
- Note: all three distance measures are only valid for continuous variables.

Distance functions

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

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

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

How does it work?

- In the instance of categorical variables the Hamming distance must be used

Hamming Distance

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

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

How does it work ?

- Choosing the optimal value for K is best done by first inspecting the data
- In general, a large K value is more precise as it reduces the overall noise but there is no guarantee
- Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value
- Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

Advantages

- No assumptions about data — useful, for example, for nonlinear data
- Simple algorithm — to explain and understand/interpret
- High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models
- Versatile — useful for classification or regression (mean)

Disadvantages

- Computationally expensive — because the algorithm stores all of the training data
- High memory requirement
- Stores all (or almost all) of the training data
- Prediction stage might be slow (with big N)
- Sensitive to irrelevant features and the scale of the data

Applications of KNN

- Recommender system
- Relevant document classification
- OCR