kNN

11 October 2025 08:09

- kNN => K Nearest Neighbors
- Classification Algorithm
- Works well with multi class
- Relies on proximity of neighboring data points (i.e. the distance of new data point to be predicted from others).
- K defines the number of neighbors to see, for prediction.
- It is a non-parametric (don't assume the data distribution) and instance based (stores all training instances in memory) algorithm.
- Distance measures -
    - Euclidean Distance (L2 Norm)
        - Use when features are on the same scale and are continuous
    - Manhattan Distance (L1 Norm)
        - Use when sparse data (like text).
        - More robust to outliers than Euclidean.
    - Minkowski Distance (Generalized)
        - A **generalized** form of both Euclidean and Manhattan distances. Controlled by a parameter **p.**

Euclidean Distance :

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

For $p = (1,2)$ and $q = (4,6)$:

$$D = \sqrt{(4-1)^2 + (6-2)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

Manhattan Distance :

$$D(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$

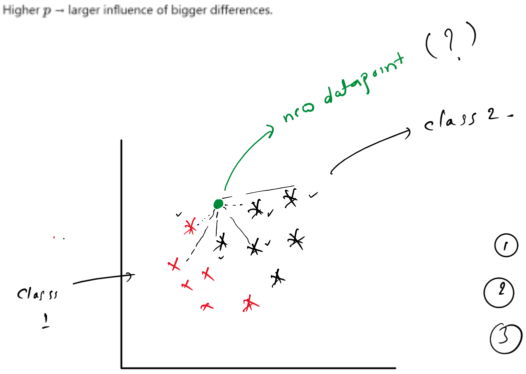
For $p = (1,2)$ and $q = (4,6)$:

$$D = |4-1| + |6-2| = 3 + 4 = 7$$

Minkowski Distance :

$$D(p,q) = \left(\sum_{i=1}^{n}|p_i - q_i|^p\right)^{1/p}$$

- $p = 1 \rightarrow$ Manhattan Distance
- $p = 2 \rightarrow$ Euclidean Distance
- Higher $p \rightarrow$ larger influence of bigger differences.



① K = 5

② Calculate distance.

③ Majority vote.

Red — 1
Black — 4

New point is a class 2 point

K = 1

↳ Underfit.

K : very large

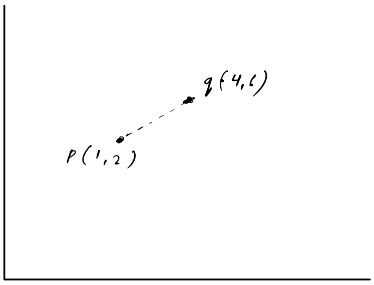
↳ Overfit

K = 6

↳ Binary classification

↳ Tie while majority vote.

→ Binary classification
→ Tie while majority vote.



① Euclidean distance.

$$D(p,q) = \sqrt{(4-1)^2 + (6-2)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

② Manhattan distance.

$$D(p,q) = |4-1| + |6-2| = 3+4 = 7.$$

③ Minkowski distance.