KNN

K-nearest neighbours (KNN) is a type of supervised learning algorithm used for both regression and classification.

KNN is called a non-parametric algorithm because it makes no assumptions about the underlying data distribution. Instead of learning fixed parameters, it uses the entire training dataset to make predictions based on similarity.

It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.

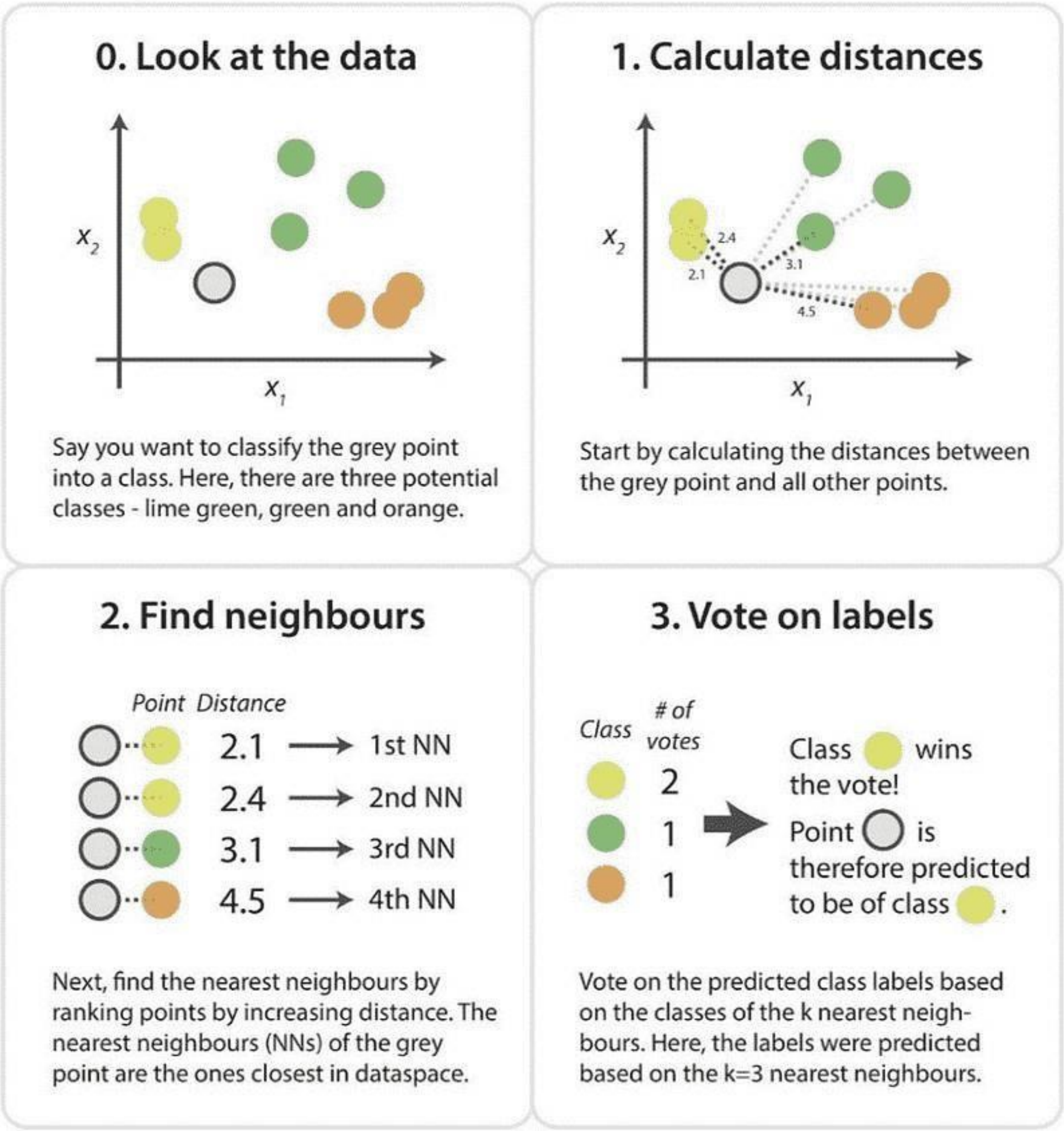
How does K-NN work in Classification Problem

The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the value of K.
- Step-2: Calculate the distance between new data point and all data points in training data.
- Step-3: Take the K nearest neighbours as per the calculated distance.

  Or

  Take K data points (neighbours) whose calculated distance from new data point is minimum.
- Step-4: Among these k neighbours, count the number of the data points belong to each category.
- Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.
- Step-6: Our model is ready.


0. Look at the data
x_2
x_1
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.
1. Calculate distances
x_2
2.4
2.1
3.1
4.5
x_1
Start by calculating the distances between the grey point and all other points.
2. Find neighbours
Point Distance
2.1 → 1st NN
2.4 → 2nd NN
3.1 → 3rd NN
4.5 → 4th NN
Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.
3. Vote on labels
Class
# of votes
2
1
1
Class wins the vote!
Point is therefore predicted to be of class .
Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

How does K-NN work in Regression Problem

The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the value of K.
- Step-2: Calculate the distance between new data point and all data points in training data.
- Step-3: Take the K nearest neighbours as per the calculated distance.

Or

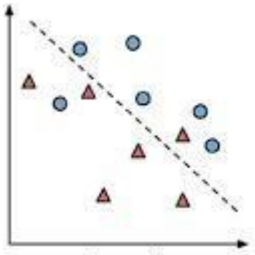
Take K data points (neighbours) whose calculated distance from new data point is minimum.

- Step-4: Then, take average of those K selected minimum distances.
- Step-5: And that will be the predicted value.
- Step-6: Our model is ready.

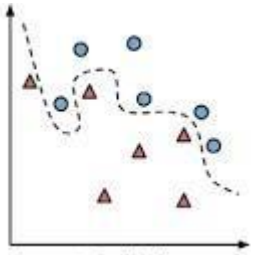
How to select the value of K in the K-NN Algorithm?

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for “K”, so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.
- The impact of selecting a smaller or larger K value on the model
- Larger K value: The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.
- Smaller k value: The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.



Larger k



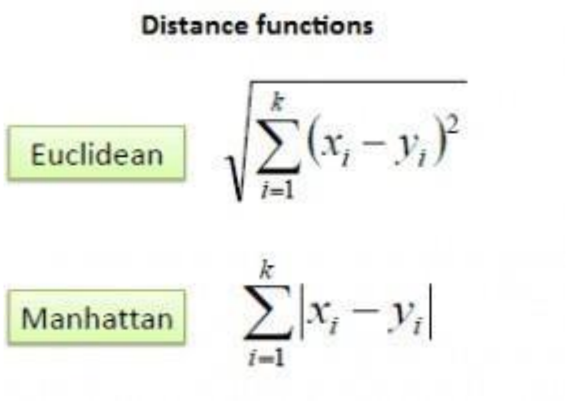
Smaller k

Calculating distance:

There are various methods for calculating the distance are — Euclidean (for continuous), Manhattan (for continuous) and Hamming distance (for categorical).

Euclidean Distance: Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

Manhattan Distance: This is the distance between real vectors using the sum of their absolute difference.



Distance functions

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

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

Hamming Distance: It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0 , Otherwise D=1.

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

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

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

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex sometimes.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

KNN: Important Interview/Exam Questions and Answers

5. What does the "K" in KNN signify?

In the KNN algorithm, "K" represents the number of nearest neighbors considered when making a prediction for a data point.

- In classification, the class is determined by a majority vote among the K nearest neighbors.
- In regression, the output is typically the mean or median of the values of the K nearest neighbors.

Choosing the right value of K is crucial:

- Small K → model may be too sensitive to noise (overfitting).
- Large K → may smooth out important patterns (underfitting).

10. Is KNN a parametric or non-parametric algorithm?

KNN is a non-parametric algorithm.

- Non-parametric means the algorithm makes no strong assumptions about the underlying data distribution.
- It does not learn an explicit model during training — it simply stores the training data and performs calculations during prediction time.
- This makes KNN flexible, but also computationally expensive during inference.

16. What is the curse of dimensionality and how does it affect KNN?

The curse of dimensionality refers to the problems that arise when data has too many features (dimensions).

- In high-dimensional space, all points tend to appear equidistant, making distance-based methods like KNN less effective.
- Volume increases exponentially with dimensions, which leads to sparse data and reduces the density of useful neighbors.
- It can lead to poor accuracy and increased computational cost.

Mitigation strategies:

- Apply dimensionality reduction (e.g., PCA).
- Use feature selection to retain only the most relevant features.

17. How does KNN deal with missing values?

KNN does not natively handle missing values — missing data must be addressed before applying KNN.

Common approaches:

1. Imputation:
   a. Mean/median/mode imputation for numerical or categorical features.
   b. KNN-based imputation (where missing values are filled based on values of nearest neighbors).
2. Removing rows or columns with too many missing values.

Note: Missing data should be handled prior to distance calculations, as missing values can distort similarity.

35. When would you prefer KNN over other algorithms?

KNN might be preferred in the following scenarios:

- When interpretability is important — it's easy to understand and explain.
- When the decision boundary is irregular and non-linear, KNN can model it without needing parametric assumptions.
- In low-dimensional datasets where distance metrics are meaningful.
- When you have a small dataset and don't need fast predictions.

However, KNN may not scale well to large datasets or high dimensions.

37. Is KNN sensitive to noise or outliers?

Yes, KNN is very sensitive to noise and outliers:

- A noisy data point or outlier can mislead the classification if it falls within the K nearest neighbors of a test point.
- Small values of K are more prone to noise, as each individual neighbor has more influence.
- Outliers can also affect distance calculations, especially with unscaled features.

Solutions:

- Use larger K values to reduce the effect of individual noise.
- Apply data cleaning or outlier detection techniques.
- Use weighted KNN, where closer neighbors have more influence.