K-Nearest Neighbourhood (K-NN) Algorithm

Arunima Dutta

What is K-NN?

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

What is K-NN?

- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. 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.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Where to use K-NN?

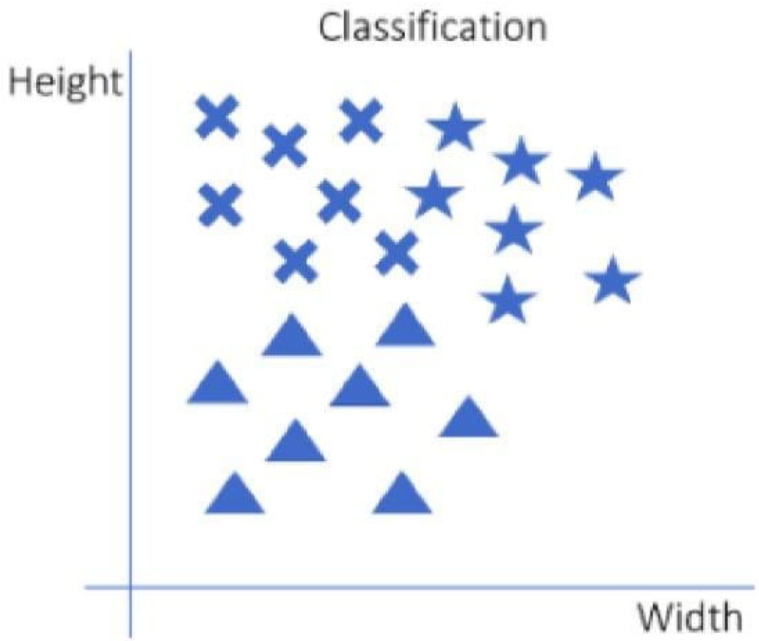
- KNN can be used in both regression and classification predictive problems.However, when it comes to industrial problems, it's mostly used in classification since it fairs across all parameters evaluated when determining the usability of a technique
  - Prediction Power
  - Calculation Time
  - Ease to Interpret the Output

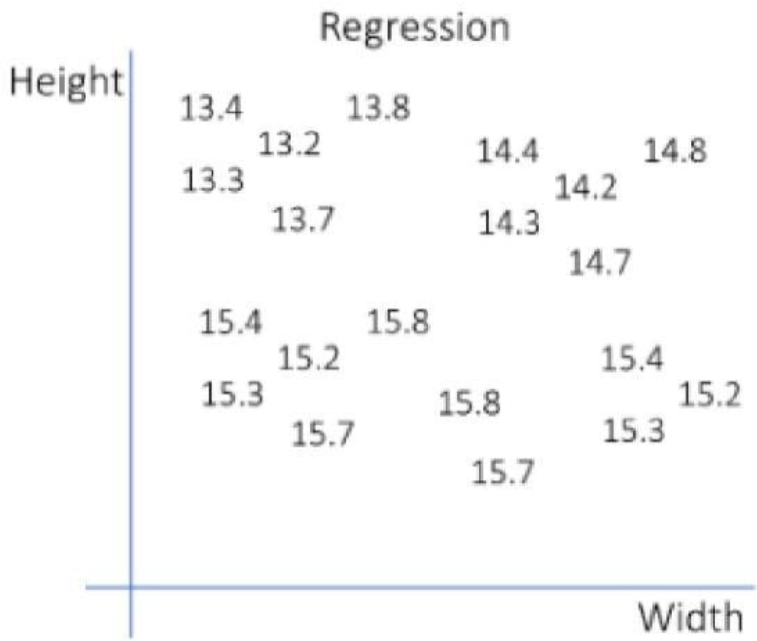
Calculating credit ratings -K-NN can help when calculating an individual's credit score by comparing it with persons with similar traits.

Use in daily life

- KNN does better than more powerful classifiers and is used in places such as genetics, data compression, and economic forecasting.
  - In political science - classing a political voter to "vote party A" or "vote party B", or to a "will vote" or "will not vote".
  - Banking system - K-NN can be used to predict if a person is fit for loan approval.Or if he or she has similar traits to a defaulter.
  - Calculating credit ratings -K-NN can help when calculating an individual's credit score by comparing it with persons with similar traits.
  - Other areas that use the K-NN algorithm include Video Recognition, Image Recognition, Handwriting Detection, and Speech Recognition.

K-NN is a Supervised Learner





K-NN Algorithm History

- KNN was born out of research done for the armed forces. Fix and Hodge -two officers of USAF School of Aviation Medicine - wrote a technical report in 1951 introducing the K-NN algorithm.
- Thomas Cover and Peter Hart proved an upper bound error rate with multiclass k-NN classifications in 1967 (twice the Bayes error rate).
- In 1970: Edward Hellman examined "the (k,k?) nearest neighbor rule with a reject option". In 1975: Fukunaga and Hostetler made refinements with respect Bayes error, In 1976: Dudani; in 1978 Bailey and Jain published distance weighted approaches
- In 1983: Adam Jozwik introduced: A learning scheme for a fuzzy k-NN rule,
- In 2000: Bermejo and Cabestany published: Adaptive soft k-nearest-neighbour classifiers.

K-NN Algorithm History

- K-NN is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data.K-NN was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.

K-NN Algorithm

- Example: Suppose there is a Speed of the car of company A, who does various speed of different road conditions and get Road condition on that.

Weather(x_1) condition	Speed(x_2)	Road (target variables) Condition
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good
3	7	???

K-NN Algorithm

- Step-1: Determine number of nearest neighbour k=3(suppose)
- Step-2: Calculating distance from the query instance

Weather(x_1) condition	Speed(x_2)	Square Distance
7	7	$\sqrt{(7-3)^2+(7-7)^2}=4$
7	4	$\sqrt{(7-3)^2+(7-4)^2}=5$
3	4	$\sqrt{(3-3)^2+(7-4)^2}=3$
1	4	$\sqrt{(3-1)^2+(7-4)^2}=3.6$

K-NN Algorithm

- Step-3: Sort the Distance and determine nearest neighbour from based on K^{th} minimum distance

Weather(x_1) condition	Speed(x_2) Distance	Rank	Is it included in 3-NN
7	7	3	Yes
7	4	4	No
3	4	1	Yes
1	4	2	Yes

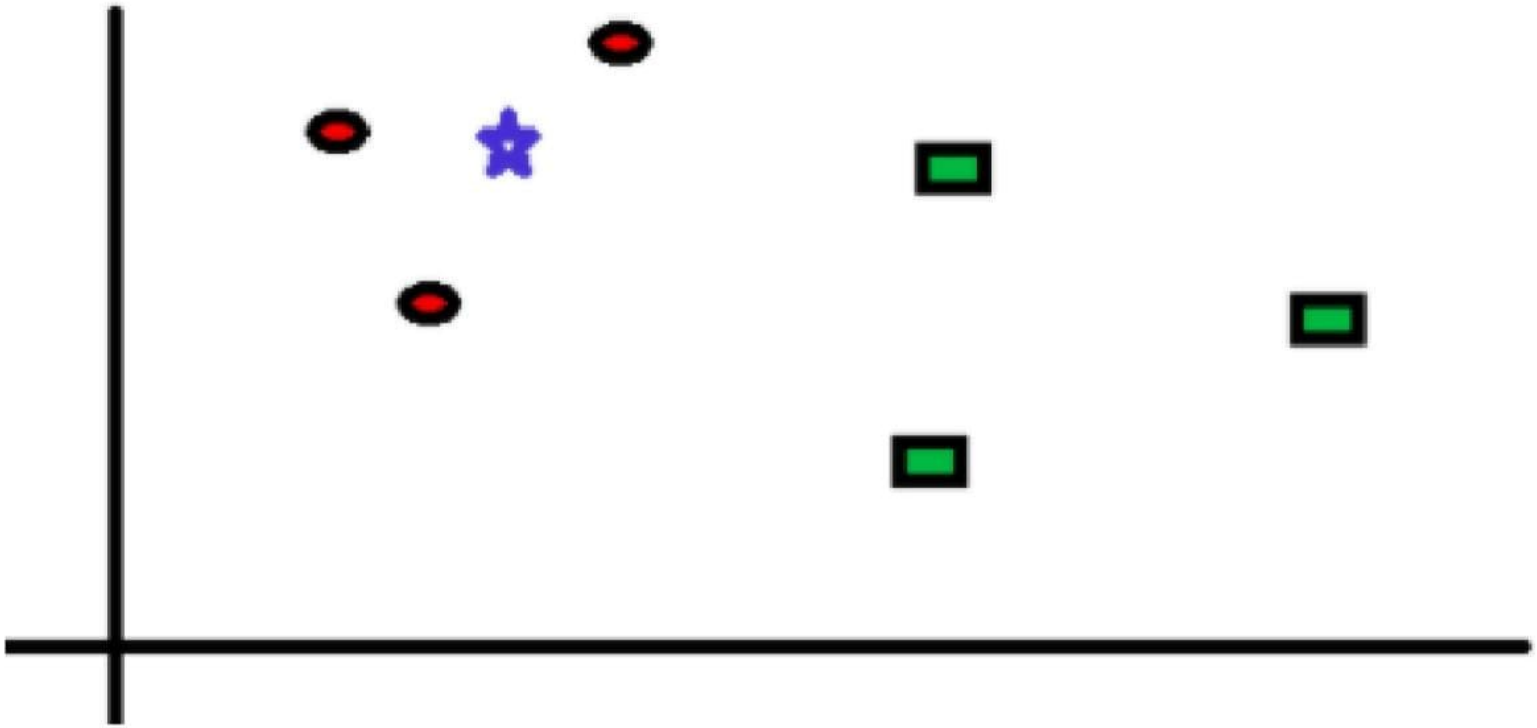
- Step-4: Gather the category(Y) of the NN

Weather(x_1) condition	Speed(x_2) Distance	Rank	Is it included in 3-NN	Y=category of NN
7	7	3	Yes	Bad
7	4	4	No	-
3	4	1	Yes	Good
1	4	2	Yes	Good

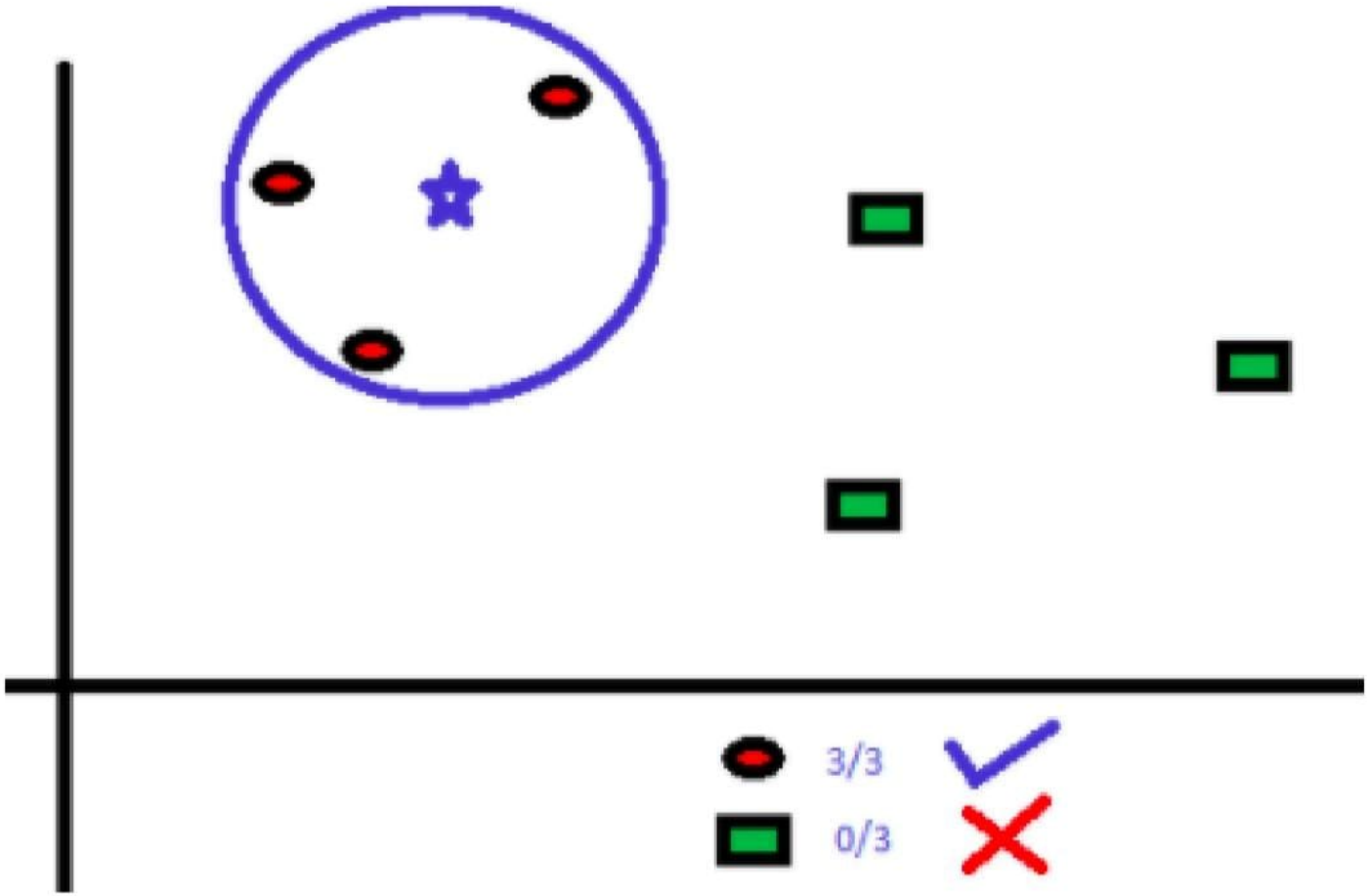
K-NN Algorithm

- Step-5: Using simple majority of the category of the NN as the prediction value of the query instance(we get 2 good, 1 bad, we conclude that new data $x_1 = 3$ and $x_2 = 7$ included in the Good category)

K-NN Algorithm



K-NN Algorithm



Advantages of K-NN Algorithm

- K-NN algorithm is widely used for different kinds of learnings because of its uncomplicated and easy to apply nature.
- The training phase of K-nearest neighbor classification is much faster compared to other classification algorithms.
- There is no need to train a model for generalization.
- No assumptions about data - no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.
- High accuracy - you do not need to compare with better-supervised learning models.
- There are only two metrics to provide in the algorithm. value of k and distance metric.
- Work with any number of classes not just binary classifiers.
- It is fairly easy to add new data to algorithm.

Disadvantages of K-NN algorithm

- The cost of predicting the k nearest neighbours is very high.
- Require high memory - need to store all of the training data.
- Sensitive to the scale of the data and irrelevant features.
- Doesn't work as expected when working with big number of features/parameters i.e., not suitable for large dimensional data.
- Hard to work with categorical features.

Number of neighbors(K) in K-NN

- Still no optimal number of neighbors suits all kind of data sets.
- Each dataset has it's own requirements. In the case of a small number of neighbors, the noise will have a higher influence on the result, and a large number of neighbors make it computationally expensive.
- Research has also shown that a small amount of neighbors are most flexible fit which will have low bias but high variance and a large number of neighbors will have a smoother decision boundary which means lower variance but higher bias.

Numbers of k in K-NN

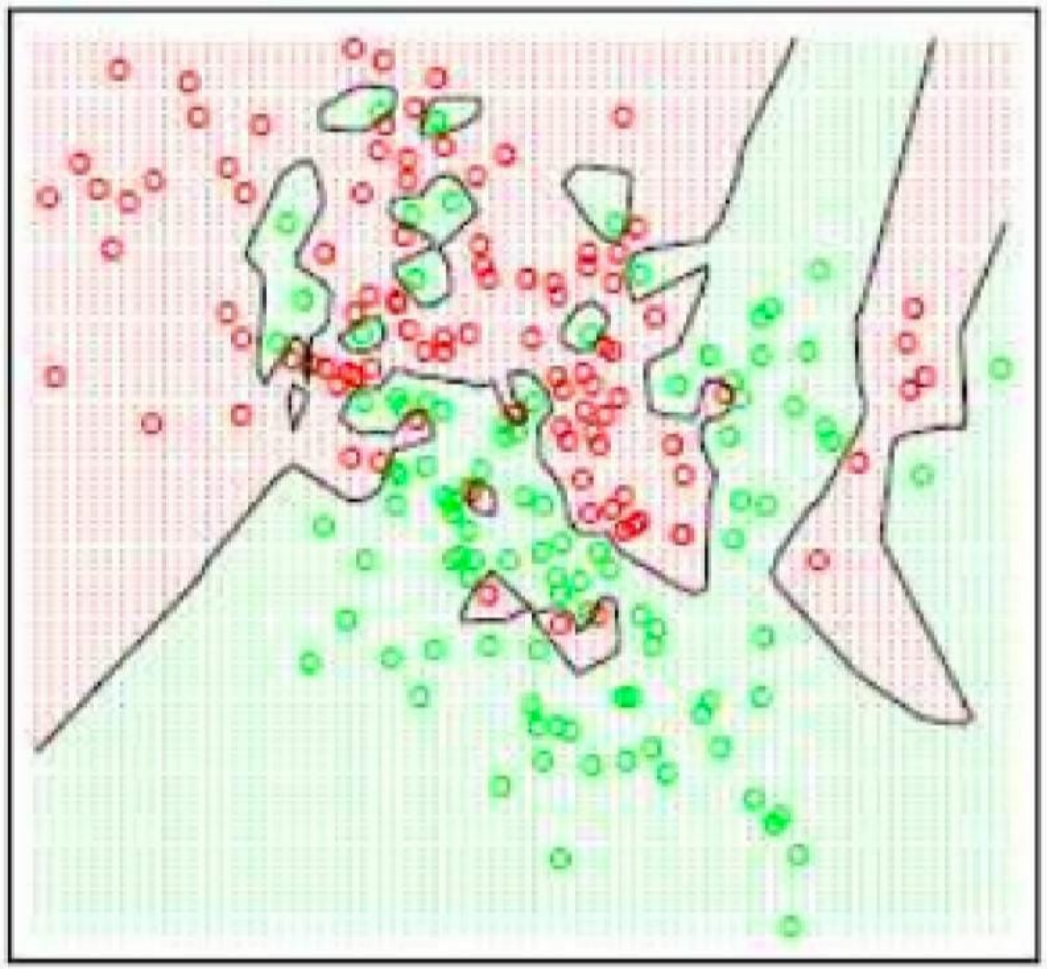


Figure: K-1

Numbers of k in K-NN

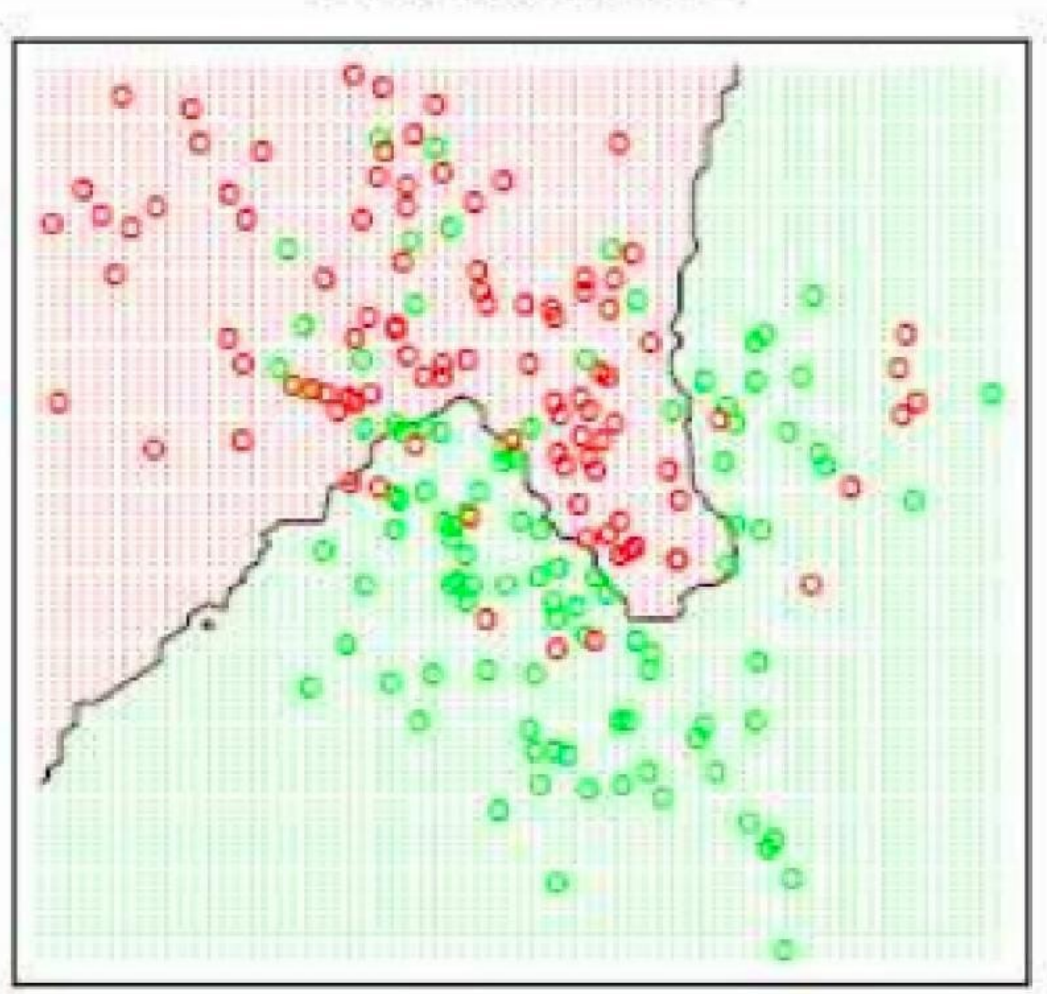


Figure: K-15

Than You all.