K NEAREST NEIGHBORS ALGORITHM

Lesson Note presented by:
Esther Oduntan

African Leaders University, Rwanda

February 6, 2020

Outline

Introduction

What is K Nearest Neighbor(KNN)

- This is a method used for Classifying objects based on their similarities to each other.
- The objects similar to each other are refered to as "Neighbors"
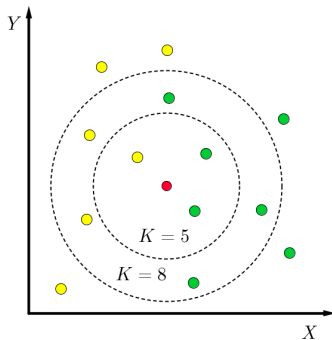- Objects with similar features are close to each other

Figure: Graphical illustration of KNN

K refers to the number of nearest neighbors

Litterature Review

KNN in Machine Learning

What is machine learning ?

- Is an application of Artificial Intelligence that focuses on the developpement of computer programs that can access data and use it to learn for themselves;

Type of Machine Learning:

- Supervised: This is an learning that have an input variable **X** and an output variable **Y** and an algorithm is used to learn the mapping function from the input to the output $\mathbf{Y} = f(\mathbf{X})$

- Unsupervised: Training of the machine with data that have not been labeled, classified or categorized.

Litterature Review

Types of Machine Learning

- K Nearest Neighbor
- Decision Tree
- Logistic Regression
- Neural Network
- Support Vector Machine

This presentation focuses on the K Nearest Neighbor Algorithm

Mathematical Framework of KNN

Preambule

The K Nearest Neighbor Algorithm works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbors and classify the object by a majority vote of its neighbors.

The KNN Algorithm is as follows:

- First Load the data of the problem;
- Initialize K to your chosen number of neighbors;
- Calculate the distance between the query example and the current example from the data. we can use distances like Euclidean, Manhattan etc;
- Add the distance and the index of the example to an ordered collection

Mathematical Framework of KNN

- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances;
- Pick the first K entries from the sorted collection;
- Get the labels of the selected K entries;
- If classification, return the mode of the K labels

The mathematical illustration of KNN is as follows

$$D_{mink}(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

.

- x is the labeled training data points
- y is the target point to be classified
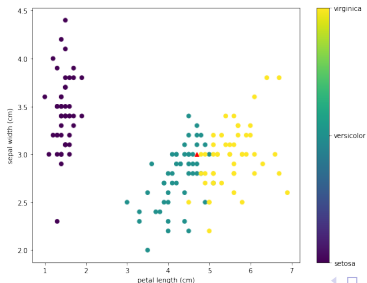- p is the Minkowski parameter

Implementation in Machine Learning

Experimental Design

- Input Dataset: Iris Dataset
- Algorithm: KNeighborsClassifier
- Performance Metric: Accuracy

Results

Features Visualization
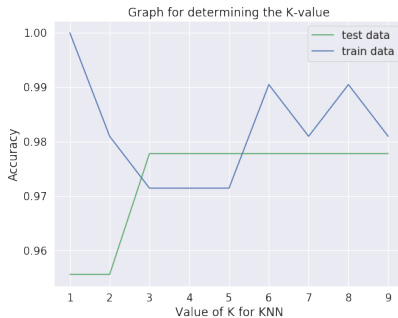
Implementation in Machine Learning

Graph for determining the K-value

Table: Classification Report

	setosa	versicolor	virginica
f1-score	1.0	0.947368	0.965517
precision	1.0	1.000000	0.933333
recall	1.0	0.900000	1.000000

The value of K is 3 and Accuracy is 0.9778

Discussion

The Pros and Cons of K Nearest Neighbors

Advantages

- Pretty intuitive and simple;
- K-NN has no assumptions;
- It constantly evolves;
- No Training Step;
- Easy to implement for multi-class problem;
- ForClassification and Regression;
- One Hyper Parameter.

Disadvantages

- K-NN is a slow algorithm;
- Curse of Dimensionality;
- Needs homogeneous features;
- Optimal number of neighbors;
- Imbalanced data causes problems;
- Outlier sensitivity;
- Missing Value treatment.

Conclusion

In this study, it has been discovered that KNN algorithm is used for classification and regression models. The algorithm gets slower as the number of prediction/ independent variables increases. Hence, it is not appropriate where prediction needs to be made rapidly.