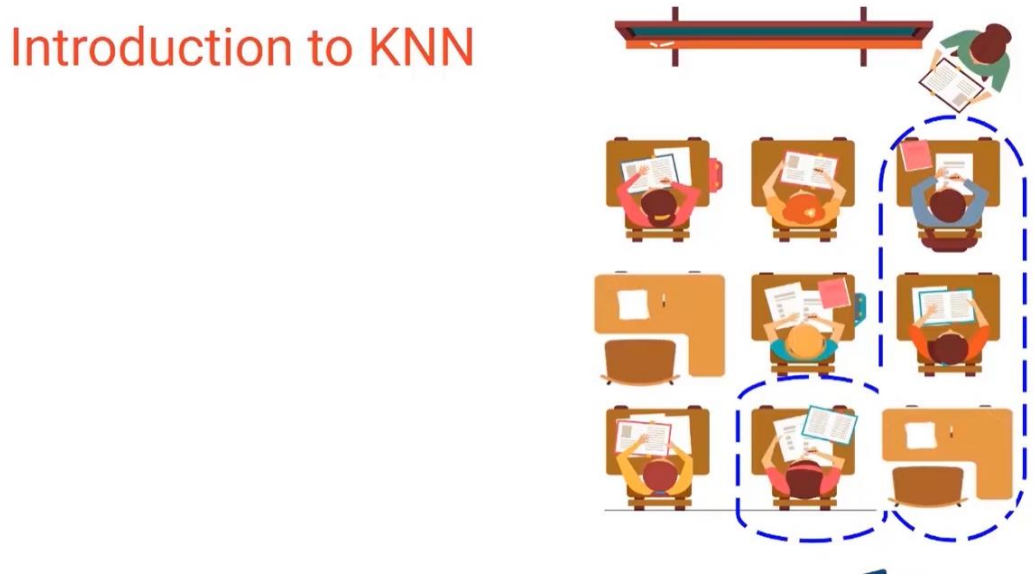
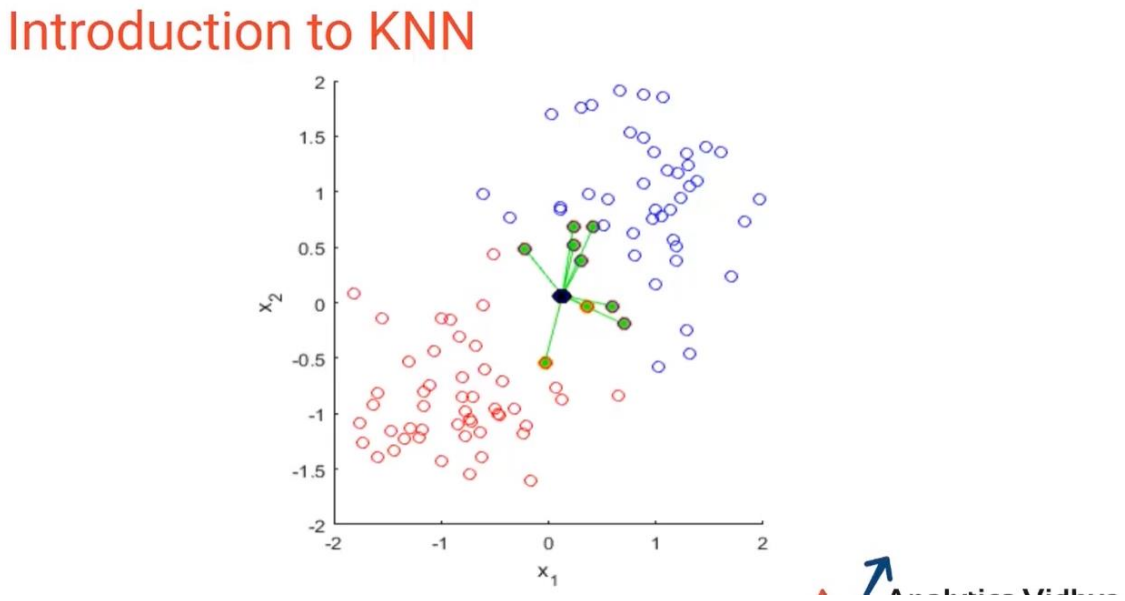
K-Nearest Neighbors (KNN) Algorithm in Python and R

Lazy learning algorithm

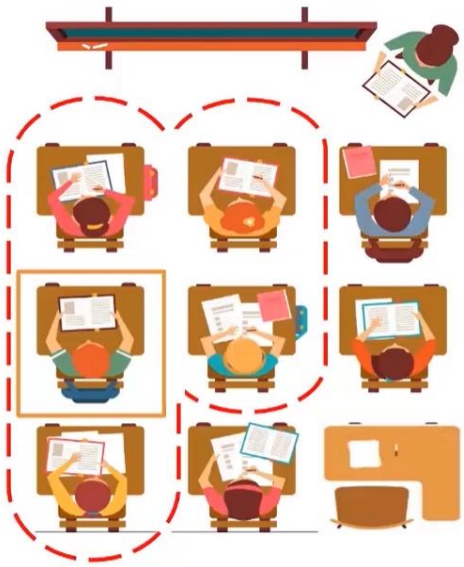
What is KNN?





Introduction to KNN

- Observe the nature of nearest neighbours
- Lazy Learning Algorithm
- Simplest Machine Learning Algorithm



Applications of KNN

Enable fullscreen

KNN can be used for both classification and regression techniques but it is generally used for classification problems.

Examples of real-life use of KNN come in the fields of-

1. Microbiology(for classifying of cells),

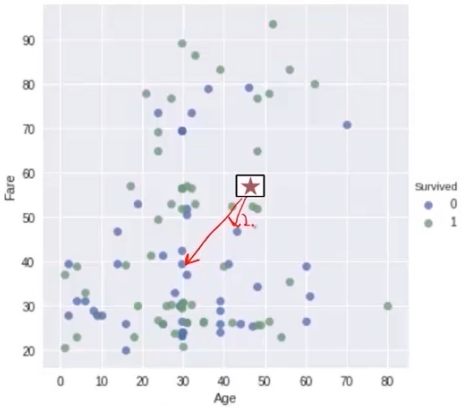
2. Marketing(for customer segmentation)

3. Credit Fraud Analytics and many more...

Steps to build a KNN model

Building Knn model

- Plot the training dataset
- Locate the new “test” instance
- Calculate distance from all train data points



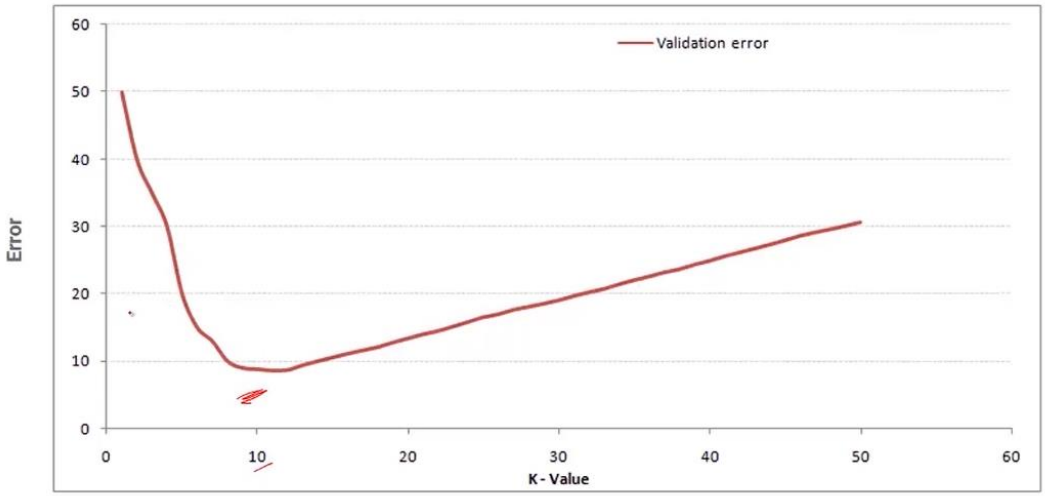
Analytics Vidhya
Learn everything about analytics

Building Knn model

Classification	Regression
1 1 0 0 0 1 0 ...	1 99 22 53 97 ...
New Instance = Mode	New Instance = Mean

Determining right value of k

Determining value of K



How to Calculate distance?

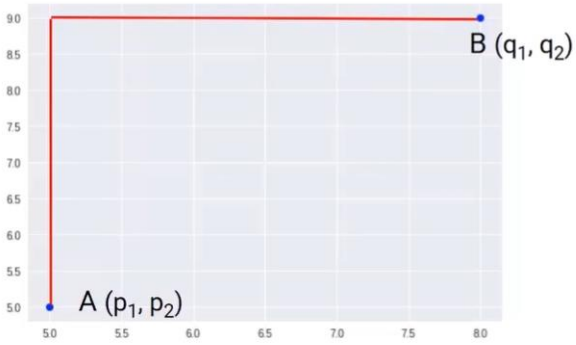
How to Calculate Distance

- Manhattan Distance
- Euclidean Distance
- Minkowiski Distance
- Hamming Distance

Analytics Vidhya
Learn everything about analytics

Manhattan Distance

Sum of Absolute differences between the two points, across all dimensions

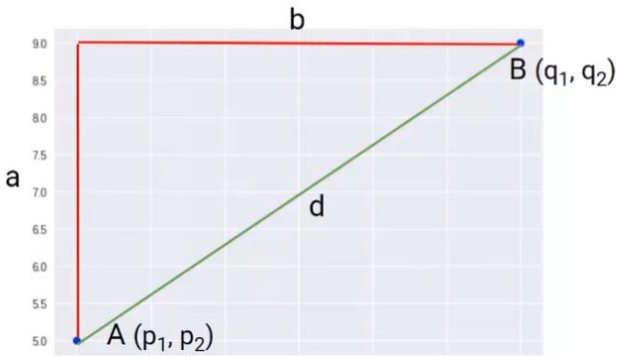


$$d = |p_1 - q_1| + |p_2 - q_2|$$

Analytics Vidhya
Learn everything about analytics

Euclidean Distance

The Shortest distance between two points



$$d = (b^2 + a^2)^{1/2}$$

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

Analytics Vidhya

How to Calculate Distance

- Manhattan Distance $D_m = \sum_{i=1}^{n} |p_i - q_i|$ (k = 1)
- Euclidean Distance $D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2\right)^{1/2}$ (k = 2)

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

Analytics Vidhya
Learn everything about analytics

How to Calculate Distance

- Manhattan Distance $D_m = \sum_{i=1}^{n} |p_i - q_i|$
- Euclidean Distance $D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2\right)^{1/2}$
- Minkowiski Distance $D = \left(\sum_{i=1}^{n} |p_i - q_i|^p\right)^{1/p}$

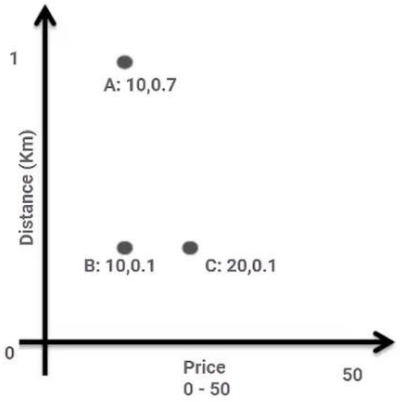
Analytics Vidhya
Learn everything about analytics

Issues with distance based algorithms

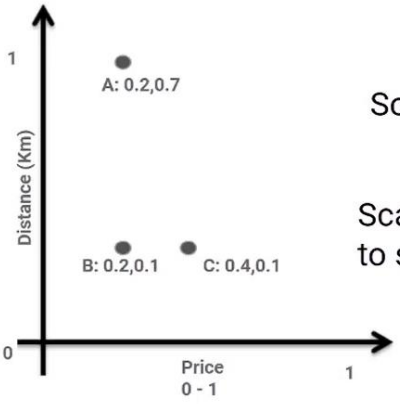
Issues with Distance Based Algorithms

- Takes the distance between points into account
- Fails when variables have different scales

Relative Distance



BA = 0.6 units



BA = 0.6 units

Solution?

Scaling all features to same scale

Analytics Vidhya

Hamming Distance

Total number of differences between two strings of identical length

ID	Gender	Marital Status	Employment Status
A	Male	Married	Self Employed
B	Female	Married	Salaried
C	Male	Unmarried	Unemployed

ID	Gender	Marital Status	Employment Status	Strings
A	0	0	1	0 0 1
B	1	0	2	1 0 2
C	0	1	3	0 1 3

Hamming Distance

Total number of differences between two strings of identical length

ID	Gender	Marital Status	Employment Status
A	Male	Married	Self Employed
B	Female	Married	Salaried
C	Male	Unmarried	Unemployed

ID	Gender	Marital Status	Employment Status	Strings
A	0	0	1	0 0 1
B	1	0	2	1 0 2
C	0	1	3	0 1 3