Supervised Machine Learning

K Nearest Neighbours

Supervised Machine Learning

K Nearest Neighbors based classifications -

a. Is an instance-based method for predicting class or value of a given query. It does not construct a general internal model

b. Classification is computed from a simple majority vote of the nearest neighbors of each point

c. New data point is assigned a class which has the most data points in the nearest neighbors of the point

d. Suited for classification where relationship between features and target classes is numerous, complex and difficult to understand and yet items in a class tend to be fairly homogenous on the values of attributes

e. Not suitable if the data is noisy and the target classes do not have clear demarcation in terms of attribute values

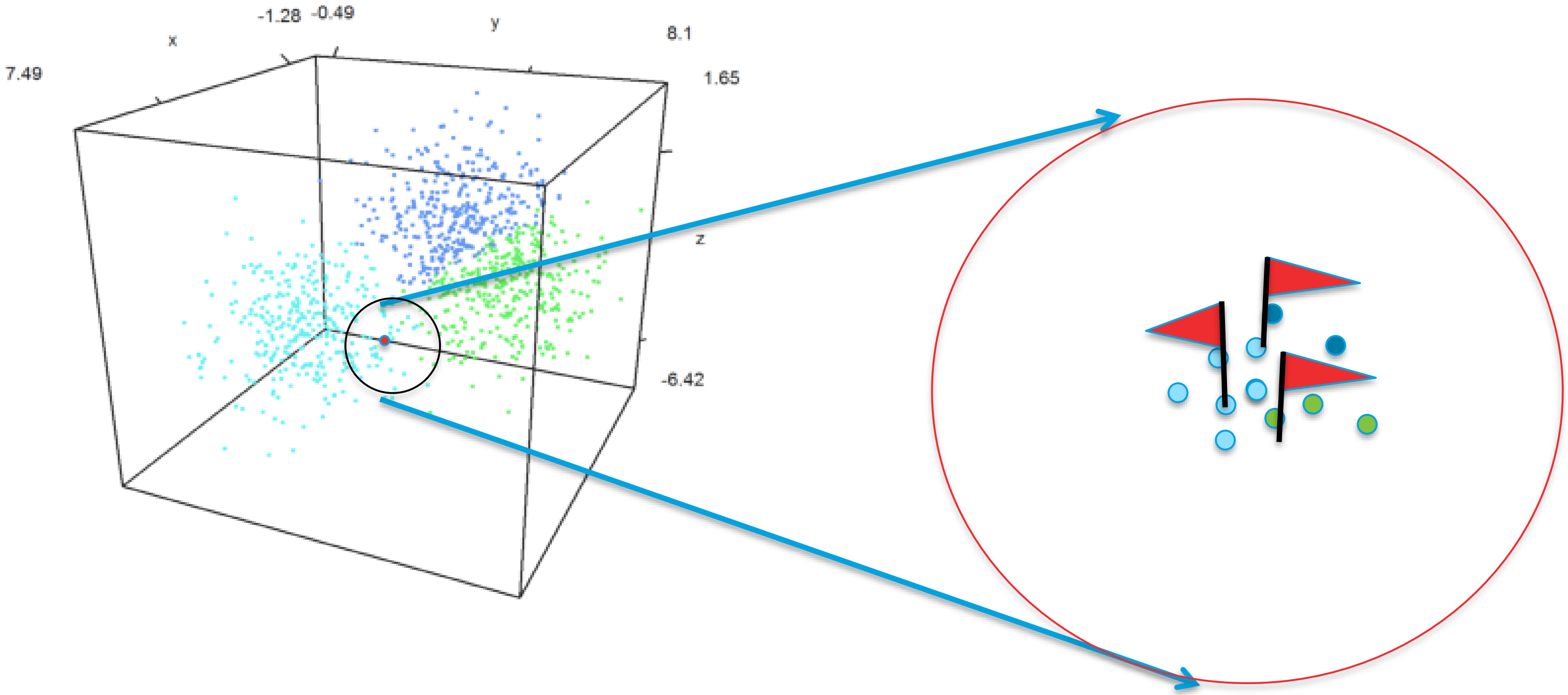
Supervised Machine Learning

K Nearest Neighbors based classifications -

f. The training data is represented by the scattered data points in the feature space

g. The color of the data points indicate the class they belong to

h. The grey point is the query point who's class has to be fixed

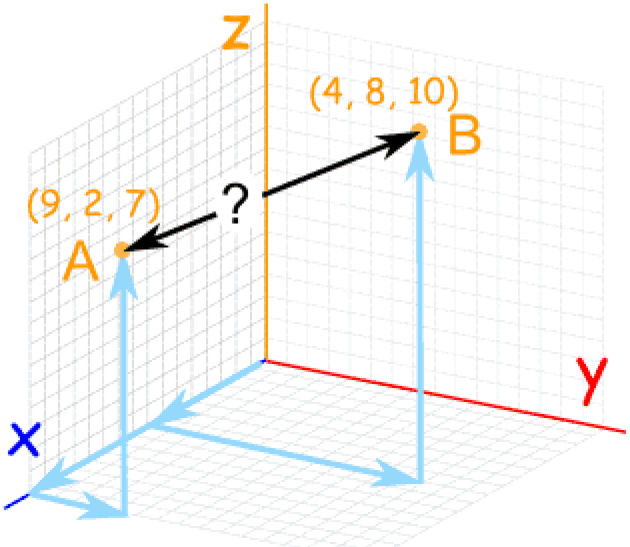


Supervised Machine Learning

K Nearest Neighbors based classifications -

i. Measuring similarity with distance between the points using Euclidian method

j. Other distance measurement methods include Manhattan distance, Minkowski distance, Mahalanobis distance, Bhattacharya distance etc.



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Supervised Machine Learning

K Nearest Neighbors based classifications -

a. Scikit-learn implements two different nearest neighbors classifiers – K Nearest Neighbor Classifier and Radius Neighbor Classifier

b. Radius Neighbor Classifier implements learning based on number of neighbors within a fixed radius r of each training point, where r is a floating point value specified by the user

c. Determining the optimal K is the challenge in K Nearest Neighbor classifiers. In general, larger value of K suppresses impact of noise but prone to majority class dominating

d. Radius Neighbor Classifier may be a better choice when the sampling is not uniform. However, when there are many attributes and data is sparse, this method becomes ineffective due to curse of dimensionality

Ref: http://scikit-learn.org/stable/modules/neighbors.html#classification

Supervised Machine Learning

K Nearest Neighbors based classifications -

e. The Neighbors based algorithm can also be used for regression where the labels are continuous data and the label of query point can be average of the labels of the neighbors

f. The approach to find nearest neighbors using distance between the query point and all other points is called the brute force. Becomes time costly (O(N^2)) and inefficient with increase in number of points

g. KD Tree based nearest neighbor approach helps reduce the time from the order of N^2 to DNlogN where D is number of dimensions. This methods becomes ineffective when D is large dur to curse of dimensionality

Supervised Machine Learning

K Nearest Neighbors based classifications -

a. The distance formula is highly dependent on how features / attributes / dimensions are measured.

b. Those dimensions which have larger possible range of values will dominate the result of the distance calculation using Euclidian formula

c. To ensure all the dimensions have similar scale, we normalize the data on all the dimensions / attributes

d. There are multiple ways of normalizing the data. We will use Z-score standardization

$$z_i = \frac{x_i - \bar{x}}{s}$$

Supervised Machine Learning

K Nearest Neighbors based classifications –

There are many distance calculation formulas in Scikit-learn package-

1. Minkowski distance
2. Euclidean distance
3. Manhattan distance
4. Mahalanobis distanc
5. Cosine similarity

Ref:
http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html
http://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/

Supervised Machine Learning

K Nearest Neighbors based classifications -

Advantages -

1. Makes no assumptions about distributions of classes in feature space
2. Can work for multi classes simultaneously
3. Easy to implement and understand
4. Not impacted by outliers

Dis-advantages -

1. Fixing the optimal value of K is a challenge
2. Will not be effective when the class distributions overlap
3. Does not output any models. Calculates distances for every new point (lazy learner)
4. Computationally intensive (O(D(N^2))), can be addressed using KD algorithms which take time to prepare

Supervised Machine Learning

K Nearest Neighbors based classifications -

Lab- 3 Model the given data to predict type of breast cancer

Description – Sample data is available at
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

The dataset has 10 attributes listed below

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10

Sol: KNN+Breast+Cancer+Modeling.ipynb