Applying K Nearest Neighbour on Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—In machine learning, an K nearest neighbouris a classifier based on the idea that the in order to predict a class of a given datapoint we use a majority vote by its k neighbours.We study the performance of the KNN for the classification of the Breast Cancer Wisconsin Dataset.The system developed performs with 96.09% accuracy for the test sample in a two class classification**

1. Introduction

The KNN algorithm is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Despite its simplicity, KNN can outperform more powerful classifiers and is used in a variety of applications such as economic forecasting, data compression and genetics.

KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consiting of training observations (x,y) and would like to capture the relationship between x and y. More formally, our goal is to learn a function h:X maps Y so that given an unseen observation x, h(x) can confidently predict the corresponding output y.

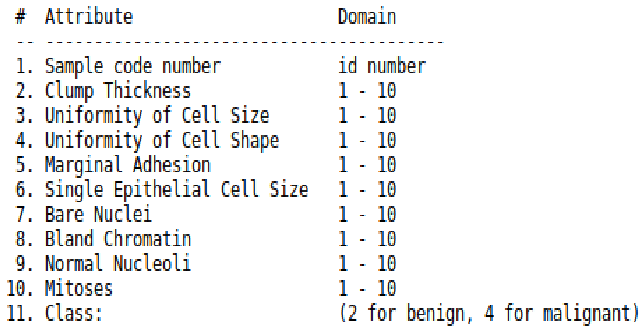
2. Methodology

The KNN classifier is also a non parametric and instance-based learning algorithm.Non-parametric means it makes no explicit assumptions about the functional form of h, avoiding the dangers of mismodeling the underlying distribution of the data. For example, suppose our data is highly non-Gaussian but the learning model we choose assumes a Gaussian form. In that case, our algorithm would make extremely poor predictions.Instance-based learning means that our algorithm doesnt explicitly learn a model. Instead, it chooses to memorize the training instances which are subsequently used as knowledge for the prediction phase. Concretely, this means that only when a query to our database is made (i.e. when we ask it to predict a label given an input), will the algorithm use the training instances to spit out an answer. KNN is non-parametric, instance-based and used in a supervised learning setting.

It is worth noting that the minimal training phase of KNN comes both at a memory cost, since we must store a potentially huge data set, as well as a computational cost during test time since classifying a given observation requires a run down of the whole data set. Practically speaking, this is undesirable since we usually want fast responses.

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1



#	Attribute	Domain
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
11.	Class:	(2 for benign, 4 for malignant)

Figure 1. Breast Cancer Wisconsin Dataset Attributes

4. Algorithm

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given unseen observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by:

$$d(x,x') = \sqrt{\left(x_1 - x_1'\right)^2 + \left(x_2 - x_2'\right)^2 + \ldots + \left(x_n - x_n'\right)^2}$$

but other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance.

.

5. Experiments

The training data of this exercise is formed by a set of Breast Cancer Wisconsin Dataset and each instance belong to one of two different classes; one of the classes is Malignant and the other is Benign.We will use KNN classifier from scikit learn to classify the data int these classes.Now we will set up KNNs parameters.We have to define no. og neighbours i.e k , and give the training data.After this we will fit our data in the KNN classifier.Now we will use the test data to test the model's accuracy using score method.Also we can see the optimum no. of neighbours using the cross validation by using scikit learn's cross val score function.

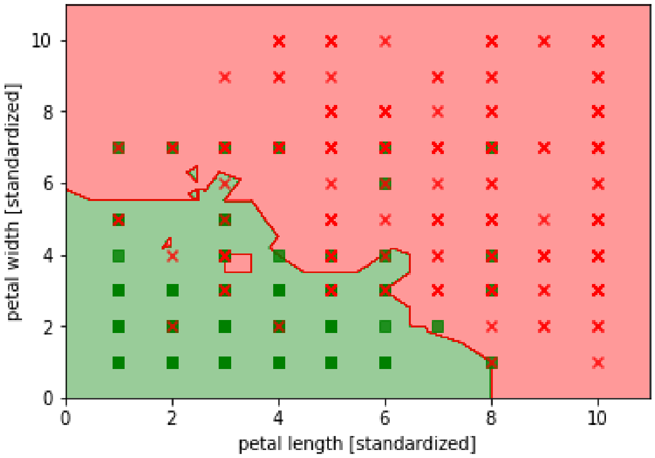


Figure 2.

6. Conclusion

After the classification we can see that the classification was very effective,.This indeed suggest that the KNN model is ver apt to classifify the data into two class Benign and Malignant for the breast cancer dataset The system developed performs with 96.09% accuracy for the test sample in a two class classification.Also we used cross val score to check on different parameters value.We got a best result of 96.09% for k =9