Project No: 35 Dermatology

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

Dermatology using K-Nearest Neighbors

Date : 25/02/2019

By : Chintada Abhilash

Sec : KEM23,

RegNo : 11601466 , B35

Submitted To : Dr . V . Devendran sir

LOVELY PROFESSIONAL UNIVERSITY,

SCHOOL OF COMPUTERS SCIENCE ENGINEERING .

Introduction :

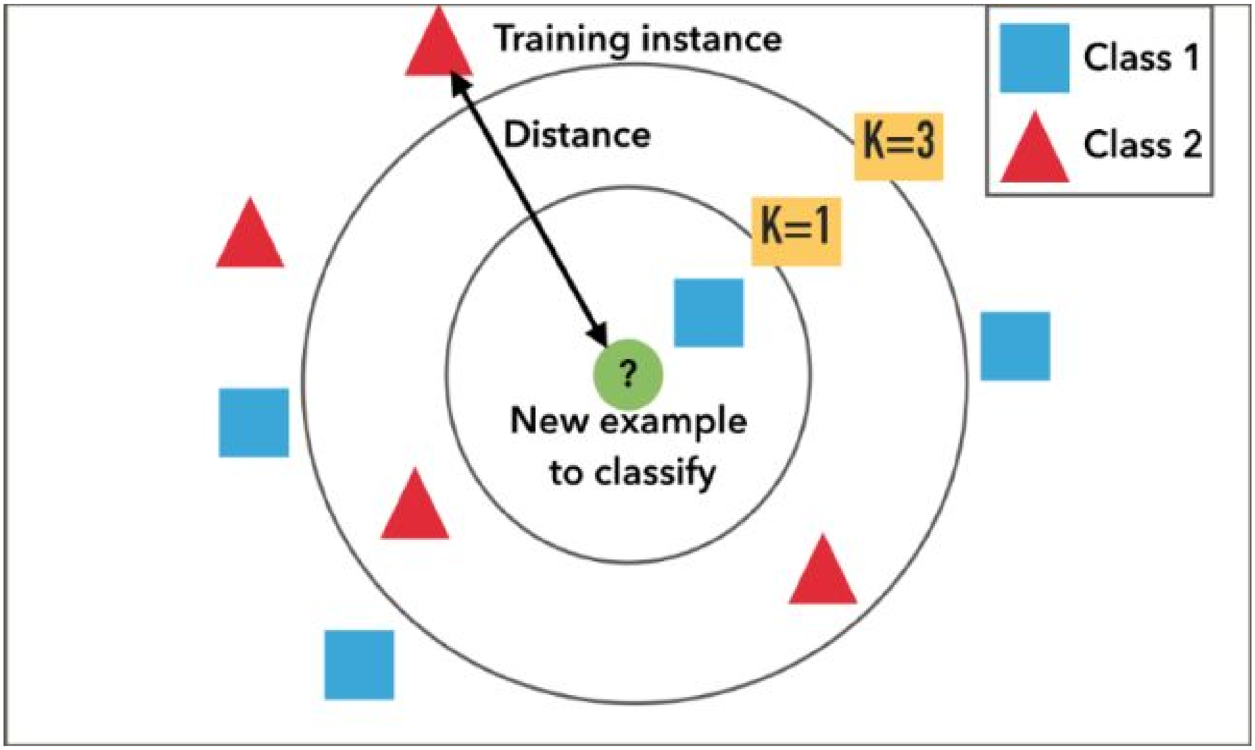
KNN (K - Nearest Neighbors) which is also called as a Lazy Classifier got its name because it does nothing when training

, i.e.. , the fit function does nothing all the mystery behind this Fancy named algorithm happens in the predict function . It takes in a parameter K , which tells the number of neighbors we need to consider before predicting the class of sample . The task is to work on Dermatology data from UCI , Use KNN algorithm on it with various K values ranging from 0 to 10 , finding out the best K value we've got .

The model has been designed from scratch . keeping track of sorted list of euclidean distance of K neighbors and implementing a voting procedure which calculates maximum votes per class in the K neighbors and allocating the class with maximum number of votes .

KNN is widely used in Machine learning and Deep learning as well , KNN is used in facial recognition , facial embedding from Inception V2 model is extracted out of face with a size of 128 . KNN could be used here in theoretically finding the closest match to the face i.e.., Practically finding the closest vector embedding to the current facial embedding . KNNs are also very accurate when it comes to different implementations

KNN has few types like Ball Tree KNN and K-D tree . but we're gonna implement a normal KNN in the current minor project .



Considering the above example , when K = 1 the new example to classify would be classified as Class 1 , when K = 3 the new example to classify would be classified as Class 2 .

So what could be the best value of K ?? well , K is a Hyperparameter which means you can only find out the best value by experimenting and previous experience . using an even K might lead to 1 - 1 clash , we can come out of that trap by Implementing the sorted list according to euclidean distance (In this case , there's no profit using K = 2 , the above problem would be solved when K = 1) . Thus the favourable values of K are Odd number . I like the number 3 by the way .

About Data :

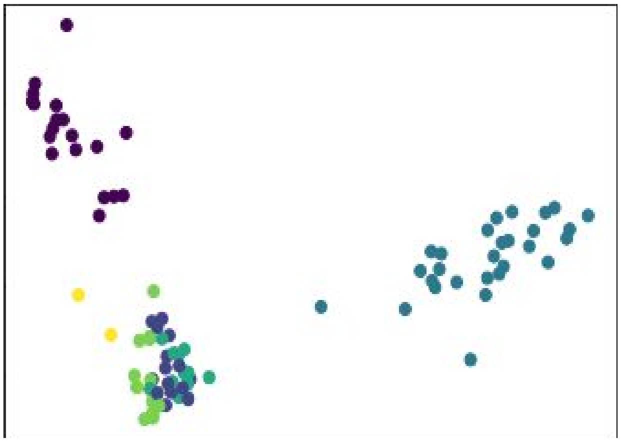
This database contains 34 attributes, 33 of which are linear valued and one of them is nominal. The names of the 34 attributes are following respectively : 'erythema', 'scaling ', 'definite borders', 'itching ', 'koebner phenomenon', 'polygonal papules', 'follicular papules', 'oral mucosal involvement', 'knee and elbow involvement', 'scalp involvement', 'family history', 'melanin incontinence', 'eosinophils in the infiltrate', 'PNL infiltrate', 'fibrosis of the papillary dermis', 'exocytosis', 'acanthosis', 'hyperkeratosis', 'parakeratosis', 'clubbing of the rete ridges', 'elongation of the rete ridges', 'thinning of the suprapapillary epidermis', 'spongiform pustule', 'munro microabscess', 'focal hypergranulosis', 'disappearance of the granular layer', 'vacuolation and damage of basal layer', 'spongiosis', 'saw-tooth appearance of retes', 'follicular horn plug', 'perifollicular parakeratosis', 'inflammatory mononuclear infiltrate', 'band-like infiltrate', 'Age', 'Disease' .

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

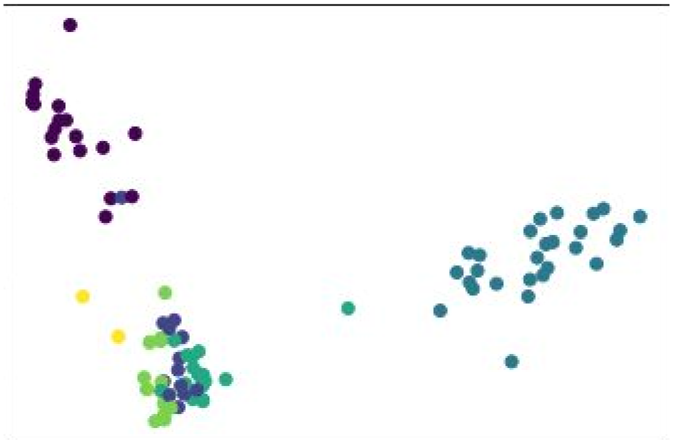
In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The names and id numbers of the patients were recently removed from the database. Number of Instances: 366 Number of Attributes: 34

Performance :

Looking at the Visualization before trying to predict it , the 34 column data has been dimensional reduced to 2 with PCA just for Visualization purpose . in case of training its trained on whole 34 columns as they were (Xtest , Ytest)



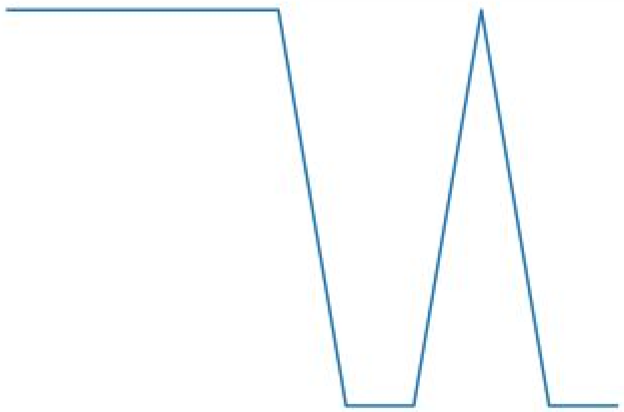
The Visualization between (Xtest , Ypred) , as you can see we've done more mistakes towards the cluster over dark green , light green and purple



Finding out the best K values between 0 and 10 , considering finding best metric is Accuracy

```
K :  1  Accuracy :  0.9560439560439561
K :  2  Accuracy :  0.9560439560439561
K :  3  Accuracy :  0.9560439560439561
K :  4  Accuracy :  0.9560439560439561
K :  5  Accuracy :  0.9560439560439561
K :  6  Accuracy :  0.945054945054945
K :  7  Accuracy :  0.945054945054945
K :  8  Accuracy :  0.9560439560439561
K :  9  Accuracy :  0.945054945054945
K :  10  Accuracy :  0.945054945054945
```

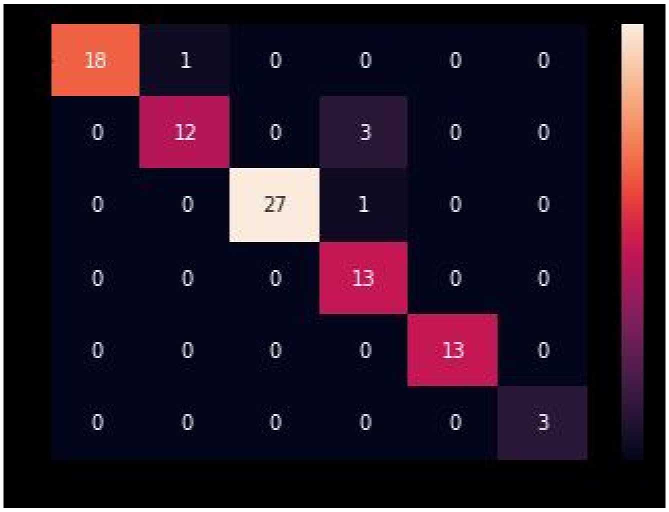
K - Accuracy Graph :



Initially the Accuracies were high , as K increased Accuracies were down by 1% , the best Accuracy were found at K = (1 , 2 , 3 , 4 , 5 , 8)

The task mentioned to find the accuracy at K = 6 , the accuracy came out to be 0.945%

Confusion matrix over Testing set



18	1	0	0	0	0
0	12	0	3	0	0
0	0	27	1	0	0
0	0	0	13	0	0
0	0	0	0	13	0
0	0	0	0	0	3

THANK YOU

LOVELY PROFESSIONAL UNIVERSITY,

SCHOOL OF COMPUTERS SCIENCE ENGINEERING .