

**ASSIGNMENT 1**

**K-NEIGHBOR CLASSIFIER**

**Student No: R00183153**

Author BLESSY JOHN GEORGE based on the lecture of prof **TED SCULLY**

**Details of Real data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CATEGORY | TOTAL DATA FOR EACH CLASS | | | | TOTAL DATA |
| 0. | | 1. | 2. |
| TRAIN DATA | 1317 | 1341 | | 1342 | 4000 |
| TEST DATA | 344 | 336 | | 320 | **1000** |

**Code snippet**

np.unique(test\_label,return\_counts=True)

np.unique(train\_label,return\_counts=True)

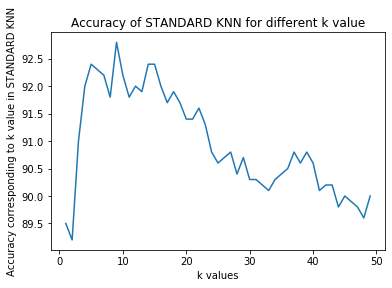
This function returns the count of different classes

**K NEIGHBORS**

Selection of value of K is very crucial in the prediction of query instance in K-Neighbor Classifier. A very low value of k means that noise will have higher influence on the result and a large value of k is computationally expensive hence it’s necessary to find a optimum value of k for better performance.

Lets investigate the effect of value of k in standard KNN :

In standard KNN the majority class within the specified neighborhood k is voted as the class of the query instance:

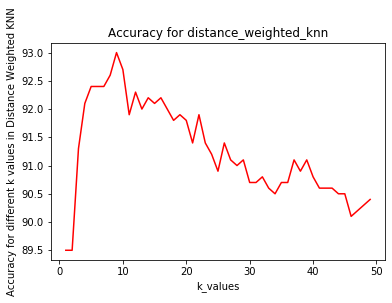


The graph above demonstrates the accuracy of standard KNN for different k (upto 50 neighbors). We can see that the accuracy is maximum for k=9 . As the k value raises from 1 to 10 there was a steep increase in the accuracy after which there is a gradual drop in the accuracy values.

In distance weighted knn ,we calculate the weights based on the distance between data from each data class and find the average weight for each class and the class with highest weight is voted as the class of query instance

Lets Investigate the various k values in DISTANCE WEIGHTED KNN:

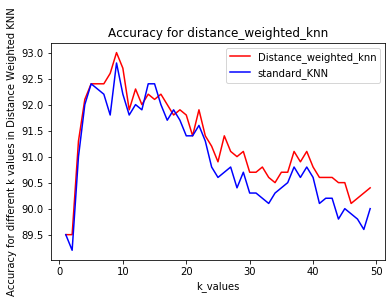
Graph below demonstrates the accuracy value of distance weighted knn for different k values upto 50 and there is a increase in accuracy till k=10



**WEIGHTS**

Standard knn uses uniform weights where each of the class is given equal weightage whereas in distance weights knn uses distance weights points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away. Here closer neighbor influence the prediction more.

Lets compare the accuracy values of standard knn and distance weighted knn .

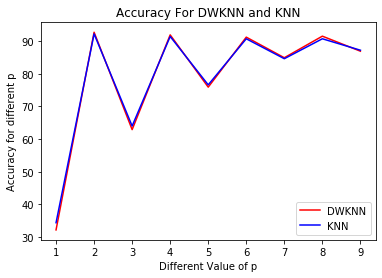


From the graph its clear that the accuracy of the prediction is higher for distance as weights rather than the standard knn which uses uniform weights.

**Varying the power parameter p in minkowski distance**

Lets vary the parameter p which is the power parameter for minkowski metric.

The distance parameter is of significant importance since the instance based learning build on distance metric.



DWKNN-Distance weighted knn

KNN – Standard knn

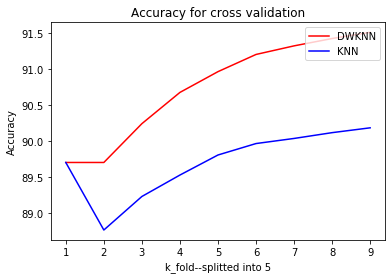
Its found that the accuracy is maximum when we use the eucledian distance.

* Accuarcy increases when the power of minkowski distance is even values (2,4,8 etc)
* A drop in accuracy is observed when the p is odd.
* So by using p even values we can improve the accuracy

**CROSS VALIDATION**

Selecting the optimal value of k is important to avoid the effect of noise.But we used the test set as the training set for this and we are actually forcing our model to fit the test set in the best possible way and this results in underfitting. So to avoid this we can go cross validation , where the data is split into k and each k can act as the test set once and as the training set k-1 times.

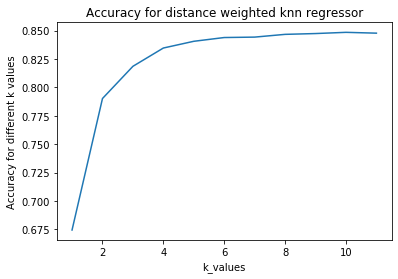
Lets look at the effect of cross validation in distance weighted knn and the standard knn:



* Dataset is splitted to K\_folds of equal size
* Each fold acts as the **testing set** 1 time, and acts as the **training set** K-1 times
* Mean of the accuracy is calculated for each value of k and is plotted

**PART3**

**The accuracy for distance weighted regressor different value of k is given below:**



ii)

Since distance between each of the feature in the query is computed with each of the feature in the training set, Each feature contribute equally to the prediction. So the irrelevant features has an adverse effect on our model.

So for distance weighted regression lets exclude each of the feature and evaluate the accuracy(r2 score).

By excluding each feature we computed the r2 score , lets now tabulate the result

|  |
| --- |
| R2 score by excluding the 0 feature from training set is 0.72487567603691 |
| R2 score by excluding the 1 feature from training set is 0.7297670816544641 |
| R2 score by excluding the 2 feature from training set is 0.7390598194674175 |
| R2 score by excluding the 3 feature from training set is 0.7333982658556217 |
| R2 score by excluding the 4 feature from training set is 0.7338342923304684 |
| R2 score by excluding the 5 feature from training set is 0.7375013980279185 |
| R2 score by excluding the 6 feature from training set is 0.08724635621459953 |
| R2 score by excluding the 7 feature from training set is 0.0896514779618226 |
| R2 score by excluding the 8 feature from training set is 0.6706100428083945 |
| R2 score by excluding the 9 feature from training set is 0.19346583749946644 |
| R2 score by excluding the 10 feature from training set is0.7429518168926097 |

* So by evaluating the table it is found that the r2 score is high on excluding feature 1,2,3,4,5 so we can conclude that these features are less relevant.
* But by excluding the feature 6,7 and 9 the r2 score reduced tremendously, So these features are important for the prediction of query instance.

Lets now remove the irrelevant features from data and compute the r2 score:

Features 1:5 is removed and the r2 score is computed and the output is:

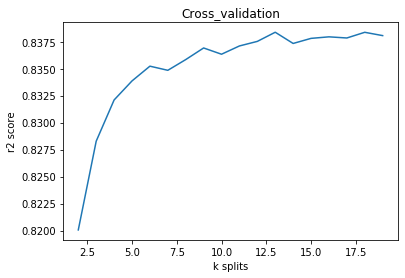
R2 score by removing the irrelevant features in distance weighted regressor is 0.9274813905596385

We can see that by removing the irrelevant features the R2 score increased to 0.927481. This proves that on removing the irrelevant features the r2score of the model increases.

2) Feature selection avoids overfitting of data.

One method of achieving this is **Cross validation**

* + - Dataset is splitted to K\_folds of equal size
    - Each fold acts as the **testing set** 1 time, and acts as the **training set** K-1 times
    - Mean of the accuracy is calculated for each value of k and is plotted



When Cross validation is performed on the regressin data we get the above curve .It is found that for k splits the r2 score increases.

3) We can also perform **GRID SEARCH CV** to find the important pararmeters in the data and the best score.

**OUTPUT**

The Best score of the model is 0.8450809609022755

The best parameters for the model is {'n\_neighbors': 8}