ML-CSE-543 – Machine Learning

Assignment 4 Analysis

By: Rajat Agarwal and Abhinav Saurabh Date: 18th Nov 2020

Q1

1. Given data of IRIS flower with attributes sepal length,sepal width,petal length,petal width in cm and class. Dataset contains 150 rows and 5 columns.
   The dataset has 4 features and 1 target variable.
   We use sklearn train_test_split to divide the data in the 70:30 train:test ratio.
   X - 'sepal length', 'sepal width', 'petal length', 'petal width'
   Y - 'class'

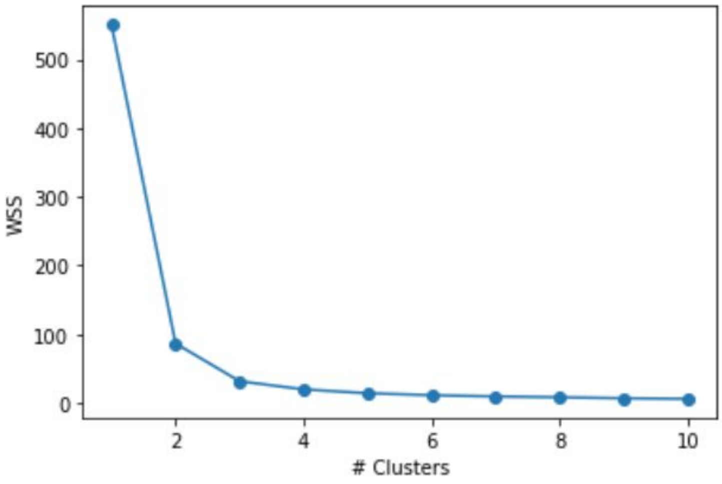
   X_train has 105 rows and 4 columns
   y_train has 105 rows and 1 column
   X_test has 45 rows and 4 column
   y_test has 45 rows and 1 column

2. Elbow method is a popular way to find out the optimal value of k i.e number of clusters for the given dataset or the problem.
   Here we use the concept of within cluster sum of squares.

   Each point is allocated to the nearby cluster,distance between them is calculated from cluster center.Then each cluster center will be updated as a means of observing the clusters.

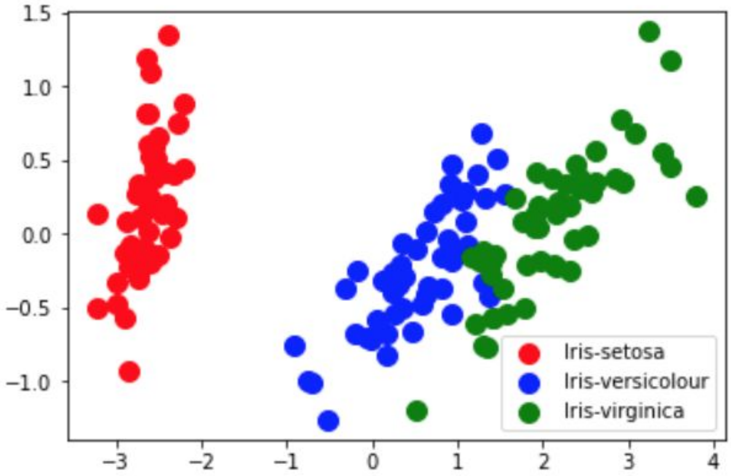
$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

where S_k is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.



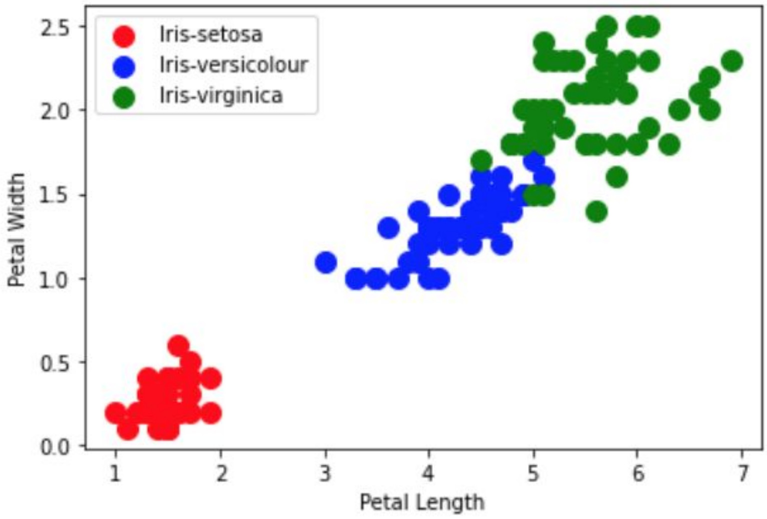
As we can see from the graph, the optimal number of clusters is 3 in this case.

3. **Scatter Plot with PCA.**
    - We have transformed the data from 4 dimensions into 2 dimensions using Principal component Analysis(PCA).
    - The target variable was given in string format for which we have used label encoder to convert them into numerical.
    - Then we have applied PCA with n_components =2.
    - Then we have used both the components of pca output to draw the scatter plot of three classes which was given in the target variable.

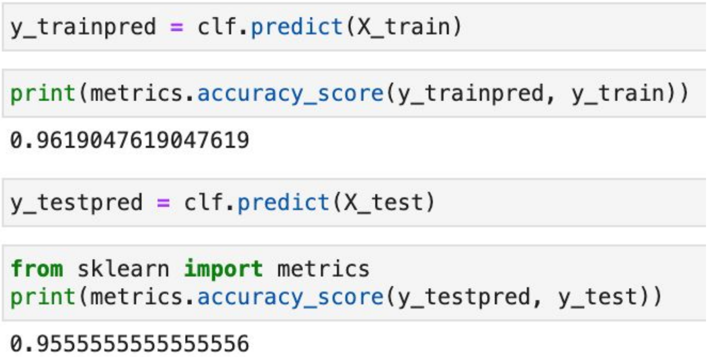


Scatter Plot (without PCA) with Petal Length and Petal Width

- Here we have used two featured i.e petal length and petal width to draw the scatter plot.
- These two features were found most useful in predicting the outcome were petal length and petal width. They have the highest correlation with the class variable.
- Then we used the two to draw a scatter plot using x-axis as petal length and y-axis as petal width.



4. **Classification using K Means , K=3**
    - We are using an optimal number of clusters i.e 3 to classify and report training and Validation accuracy.
    - We converted the target value into a number using a label encoder.
    - We have obtained 96.19 % accuracy in training data.
    - We have obtained 95.55% accuracy in testing data.
    - This model can result in huge errors when the target value doesn't match the cluster number. The cluster number in training period could be assigned differently as there is no target value to train k-means which results in loss of accuracy.



```
y_trainpred = clf.predict(X_train)
```

```
print(metrics.accuracy_score(y_trainpred, y_train))
```

```
0.9619047619047619
```

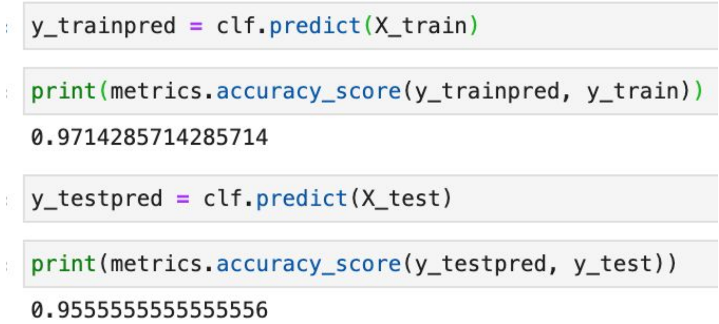
```
y_testpred = clf.predict(X_test)
```

```
from sklearn import metrics
print(metrics.accuracy_score(y_testpred, y_test))
```

```
0.9555555555555556
```

Classification using logistic Regression

- Used logistic regression as Kmeans was not always reliable.
- Converted target value into number using label encoder.
- We have obtained 97.14 % accuracy in training data.
- We have obtained 95.55% accuracy in testing data.



```
y_trainpred = clf.predict(X_train)
```

```
print(metrics.accuracy_score(y_trainpred, y_train))
```

```
0.9714285714285714
```

```
y_testpred = clf.predict(X_test)
```

```
print(metrics.accuracy_score(y_testpred, y_test))
```

```
0.9555555555555556
```

Q2

(1) Dataset consists of 1000 sentences, 500 each belonging to class 0 and class 1. After 70:30 split 700 and 300 are divided into training and validation set

(2) Preprocessing is applied on both training and validation data.

(3) The vocabulary size is 1561, Feature matrix size for training and validation 700x1561 and 300x1561 respectively.

(4) Done

(5) Training accuracy: 94.58%
Validation accuracy: 79.33%

Sample misclassified sentences are:

1. want sandwich go firehouse
2. first time going think quickly become regular
3. considering two us left full happy 20 cant go wrong
4. cant beat
5. made drive way north scottsdale one bit disappointed
6. cant go wrong food
7. miss wish one philadelphia
8. Thing
9. bit sweet really spicy enough lacked flavor
10. never anything complain
11. appetite instantly gone
12. warm feeling service felt like guest special treat
13. ive better atmosphere
14. checked place couple years ago impressed
15. weird vibe owners

The reason for misclassification is due to the fact that Naïve Bayes considers features as independent, with respect to text all words in a sentence are seen independently and there is no ordering considered. For this dataset likelihood is dependent on the word frequency in each class. As a result, the class which is favoured is the one with high frequency of the word.

This is illustrated with the stats given below:
Words in class 0 (training set) = 3740
Words in class 1 (training set) = 3719
Length of vocab is = 1561

Freq of word ‘go’ class 0 = 23
Freq of word ‘go’ class 1 = 9
log likelihood (‘go’ | 1) = -6.26
log likelihood (‘go’ | 0) = -5.39

log prior for each class is -0.69

Therefore, the likelihood of a word is dependent on its frequency per class.
‘go’ appears more frequently in class 0 than 1 and for testing sentence class 0 will dominate over 1 for ‘go’.

For a sentence having words such that words are more frequent in wrong class, is more likely getting assigned wrong class and misclassification happens