CSE-454$||DataMining||Assignment4$

Nurettin Cem Dedetas-171044028

January 17, 2021

1 Algorithm

1.1 Dataset

The dataset I have chosen to work on is a well known testing data set called iris from Sklearn which includes 4 attributes of a given flower and the type of the flower.There are 3 types of flower.

1.2 K-Fold Cross Validation

The k-fold cross validation is done by hand in the function kfold(). The data enters and is deviden into n parts, which are then rotated as the test(size=dataset/n) and train(size=dataset-train) groups.

1.3 Naive Bayes Algorithm

From the previously split train and test groups, the train group is then used to make model . Mean and Standart deviation are calculated for all attributes for each type of flower.
Then we predict the types of the test group and compare the predictions to the actual values.

1.4 Accuracy,Precision, Recall and F1

1.4.1 Accuracy

The number of correct predictions are divided by the total predictions. Then taken the mean from all folds of K-fold Cross Validation.

1.4.2 Precision

The number of correct guesses for each type is divided by all the guesses for that type. Which is then avareged for all types. Then taken the mean again for all folds of K-fold Cross Validation.

1.4.3 Recall

The number of correct guesses for each type is divided by all of the elements of that type. Which is then avareged for all types. Then taken the mean again for all folds of K-fold Cross Validation.

1.4.4 F1 Score

F1 score is then calculated from recall and precision. Which is about 90% for raw data and 50% for PCA

1.5 PCA

In my case the pca algorithm actually droped my accuracy by about 50% which I suspect caused by the dataset which is already prepared, thus PCA made the wrong correlations.