PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100-ft Ring Road, Bengaluru – 560 085, Karnataka, India

Report on

Machine learning Project

Aug - Dec 2019

DEPARTMENT OF CSE

PROGRAM B.TECH

Krithi D PES1201701789
Prathiba P PES1201700112
Nimisha Katyayani PES1201700083

Problem Statement :

- To estimate the accuracy of **Naïve Bayes algorithm** using 5 fold cross validation method on **House-votes-84** dataset.
- To estimate the precision, recall, accuracy, and F-measure.

Estimates and results are produced by replacing the missing values ('?' here by appropriate values)

ML Techniques used :

- We have used the Naïve Bayes Algorithm and had implemented a Naïve Bayes model for this project. This is one of the Supervised learning techniques.
- 5 fold cross validation method had also been used for the training and the testing sets. In general,
  - A k fold cross validation method is a procedure to estimate the skill of the model on new data.
  - The value of k can be chosen based on some methods with respect to the data.
  - No libraries have been used in this project .
  - There are generally commonly used variations on cross validation such as stratified and repeated cross validation using Scikit-learn .
- **Why cross validation method ?** The use of cross validation method was essential in our project because it is less biased estimate of the model than other methods with a simple training and testing split.
- These two main techniques constitute the working of the model in the project.

Results obtained:

- The accuracies for each fold dataset is obtained.
- Since cross validation method is used the training and the testing sets isn't the same any two times.
- Thus the accuracies obtained for each fold is efficient and may differ on each run of the model.
- The conclusion of the working model and the accuracy is calculated by taking the mean average of all accuracies obtained in the five fold cross validation method used here.
- A confusion matrix has been constructed using the values of true positive, false positive, true negative and false negative.
- **Performance measures**:
    - Now that the values of true positive, false positive ,true Negative and false negative, we calculate some of the performance such as recall , precision, and f-measure.
- Since the model is using an efficient and a simple method of cross validation, the accuracies of all the k folds are > 90 % and can extend up to 99% based on the split of the dataset. Now that the mean average of the accuracies are considered , the accuracy of the model is always >90%.
- A simple graph has been constructed to prove the above result based on all the accuracies.

Conclusion:

The model for estimating the accuracy of the House-Votes-84 dataset had been designed using the Naïve Bayes algorithm. The cross validation technique has been used to split the data and ensures unbiased splitting. Once the training and testing sets are obtained, the target classes are observed on the training set and the instances of the dataset have been separated according to the classes.

Now the model is implemented on the training set , gets the values from the statistics. Applying the results on to the testing set, the target class values are thus predicted using the best probability method and

the best class that it could fit in.
Having the actual and predicted values in hand, we thus find its accuracy and conclude the skill of the model with the confusion matrix.

References:

- https://www.saedsayad.com/naive_bayesian.htm
- https://machinelearningmastery.com/k-fold-cross-validation/
- Partners: Astrirg: http://astrirg.org/projects.html
- Cite as: Saha, S. et. al, Introduction to Machine Learning and AstroInformatics, pp 1-523, DOI: 10.13140/RG.2.2.10707.94242

Thank you.