ML Lab Report08

Name: Raman Kumar
Reg NO: BL.EN.P2DSC22009
Course: Mtech Data Science

Q.1 Compare the accuracies of NB classifier with other classifiers obtained on your project data so far. State if you'd use NB classifier for your project data or NOT. Justify your answer.

Classifier	Accuracy(%)	Time Taken(sec)
SVM	78.70	170 sec
MLP	69.89	0.53 sec
KNN	68.11	0.22 sec
NBC	79.84	0.61 sec

The majority of applications for naive Bayes algorithms include sentiment analysis, spam filtering, recommendation systems, etc. NB Classifiers are quick and simple to use, their major drawback is the need predictors to be independent.

Naive Bayes implements the Naive Bayes algorithm for multinomial distributed data and is one of two classic variants of Naive Bayes used in text classification (where data is typically described as word vector counts). This algorithm is a special case of the general Naive Bayes algorithm and is specially used for prediction and classification tasks with more than two classes.

KNN vs MLP vs SVM vs NB Classifier:

- kNN: For this model, it calculated an accuracy of 70.24% upon using the Euclidean distance measure and the nearest neighbor value used for it was k=5 and the time taken for the model was 0.22 seconds.

- MLP: Earlier the model calculated an accuracy of 69.89% when using the SGD algorithm with the ReLU activation function and learning rate of 0.05. The time taken by the classifier to classify the labels is 0.53 seconds which is not better than SVM.

- SVM: For this model, it calculated an accuracy of 69.89% when regularised parameter C was taken as 1 and kernel function opted as linearSVC. The time calculated for this model was most of all i.e 170 sec.

- NB Classifier: For this model, it calculated an accuracy of 79% when Gaussian naive bayes is used. The time calculated for it was 0.61 seconds.

The dataset was the reviews about products from the various customers. So text classification was done on the given dataset. Therefore, NB classifier is good model to classify the good, average and bad words.

Q.2 What is the assumption made for usage of NB classifier? If a given dataset doesn't meet the condition, what approach may be taken to use NB classifier?

Conditions for NB Classifier:

- Check if the dataset contains independent features?
- If not.
  If the features are somewhat dependent on each other because in real data this is the case. Then it will decrease the accuracy or might not give the best model.
- For that, we can use the "dimensionality reduction" technique to reduce rank and by finding the RREF of dataset we can remove redundant features.
- Gram-Schmidt process: As it gives the orthogonal set of vectors and orthogonality gives linear independence of features.
- One more problem is size of the dataset.
- If there will be large number of features, then we need large observations to estimate prob.
- To overcome this situation, we can add extra samples to population.
- What is the type of data you have ?
- Continuous (or) Categorical
- Categorical is must needed. We can use binning (or) discretization to convert it into nominal or ordinal.
- If we encounter the words not present in the test data for any class like bad, average or good then it will be turn out to zero probs.
- Solution to this situation is Laplace smoothing or m-estimate to make it normally distributed and remove outliers.