Viveksinh Solanki – 10441787
BIA_660_C – Assignment_5_analysis

Q2:

1. How does the sample size affect each classifier’s performance?

   Multinomial Naïve Bayes: We can observe from plots, as sample size increases f1 score and auc score increases rapidly upto certain limit. After 4000 samples, adding more samples doesn’t help much. Because after this limit, f1 score is almost same and auc score is increasing but super slowly.

   SVM: For this dataset, SVM is also performing almost similar to Naïve Bayes, but it has higher f1 and auc scores than Naïve Bayes for every round.

   Hence, we can say that for given classifiers, increasing sample size helps upto certain number of samples (4000 in this case), but then adding more samples is meaningless.

2. How many samples do you think would be needed for each model for good performance?

   For given dataset, both models have best performance at **sample size = 4000**.

3. How is performance of SVM classifier compared with Naïve Bayes classifier, as the sample size increases?

   For given dataset, regardless of sample size, SVM is performing better than Naïve Bayes from the start. And as sample size increases, SVM has higher scores than Naïve Bayes

   Overall, SVM is better classification model than Naïve Bayes. As it uses largest margin boundaries for separating classes, it can generalize very well on test set as compared to Naïve Bayes, which classifies based on probabilities only.