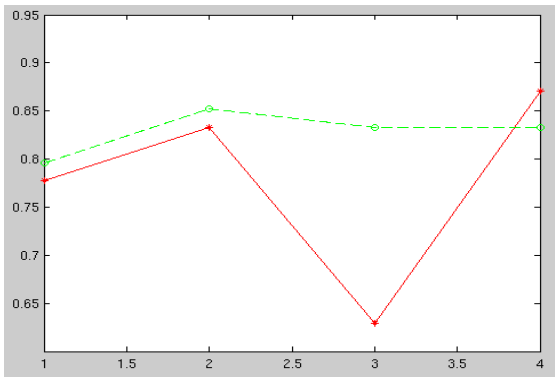
A demo of the comparison of Logistic Regression and Gaussian Nave Bayes for binary classification when supplied with different amount of training dataset

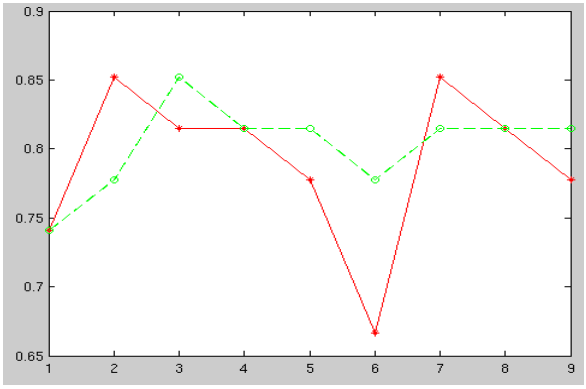
Lei Liu 9669373

First demo compares the accuracy of different learning models employing the Logistic Regression (LogReg) and Gaussian Naive Bayes (GNB) from MLOtools. The base algorithm proceeds as follows.

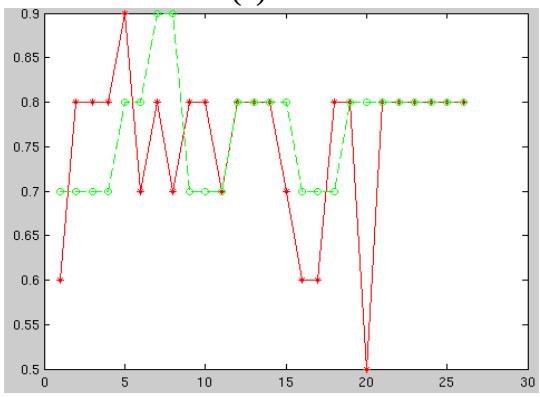
given data with corresponding labels and number of folds
shuffle examples and spilt dataset into folds
pick one fold (e.g. the last fold) as test dataset
repeat
    pick n folds of dataset as training data, n is range from 1 to the number of folds
    training LogReg and GNB models respectively
    output accuracy for each models
plot their accuracies with the increasing amount of training dataset



(a)



(b)



(c)

Figure1: Demo result of dataset “heart” from MLOtools. X-axis is the number of training folds. Y-axis is the accuracy. Red Cross values are the accuracies of LogReg model, while Green Circle values represent the accuracies of GNB model. Figure (a) spilt dataset into 5 folds, figure (b) split dataset into 10 folds, figure (c) split dataset into 27 folds

From this demo, the GNB performed better than LogReg when split into lower folds, however, when split into larger folds, the performance of two models are approximately similar. Thus, more experiments needs to be implemented in this topic.

For further experiments, three datasets with different types of features will be applied individually. Values of the features in dataset “breast” are continues, while “congress” has all discrete features. Features in “heart” dataset has both continues and discrete values.

All algorithms will be implemented and optimise the performance of models firstly before making comparison. The generalization performance for each model will be evaluated by ROC analysis, i.e. F-measure.