**Assignment 3:**

(Group 10)

**ML Method evaluated:**

**A short description of the ML methods evaluated and justification of why these methods were selected**

The models chosen are LDA, K-nearest neighbor and Random forest. The reason to choose these models are simple, we wanted pick three classifiers with three different properties so that we can explore between these differences and select the best one.

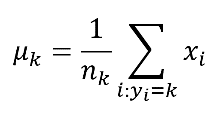
**Linear Discriminant Analysis:**

LDA was selected primarily because we wanted to use a linear model for evaluation. LDA seemed to be more stable than other linear methods available while testing for different performance metrics.

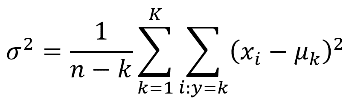
Linear Discriminant Analysis was chosen to be evaluated because LDA performs better for binary classification, provided the training sets are smaller. LDA models the distribution of features separately in each of the response classes. It then applies Bayes theorem to estimate the probability.

LDA makes following approximations:

Average of all training examples from k classes

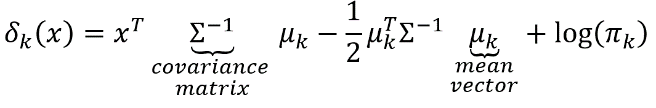


Weighted average of individual covariance matrices.



When there are more than 1 features for LDA, it is assumed that all the classes are drawn from a multivariate gaussian distribution, with class specific mean vector and a common covariance matrix.

Discriminant equation:



*Parameters and its range:*

**Solver**: The method used to compute the covariance matrix.

**n\_components**: The number of components for dimensionality reduction

**store\_covariance**: It is used to compute covariance matrix if ‘svd’ method of solver is used, since it does not compute the covariance matrix.

Range of values given for the parameters:

solver: 'svd', 'lsqr', 'eigen'

n\_components: [0, 1],

store\_covariance: [True, False]

**K-nearest neighbour:**

KNN is a simple and powerful classifier with nonparametric, lazy learning approach. As it doesn’t make any assumptions on the test data, it is suitable for classification. Here the classification is made based on the distance between the test data and the K neighbours, where K is a parameter. Lesser the distance more likelihood that it belongs to that particular class. If the K is very less the model tends to overfit and for the higher values of K the model underfits. Thus, we need to select the best K value which is done by performing cross-validation.

While feature selection (number of features with less correlation) we observed that KNN was performing equally well for the variation of number of features, thus it has less variance.

*Parameters and its range:*

While fitting the model with the KNN we choose number of nearest neighbours, different metrics to calculate the distance and weight function which is used in prediction as our parameters, with the below range of values in each.

Number of Neighbours in range (5,15)

metrics = ['minkowski','euclidean','manhattan']

weights = ['uniform','distance']

**Random forest:**

As we have selected two methods one linear and one non-parametric, we chose a tree-based classifier, random forest. It is a decision tree which is built on bagging and to improve the same. Random forest builds de-corelated trees and then merges them to get the accurate prediction. Which provides a good estimate with features as it tries with different features.

*Parameters and its range:*

Initially we selected min\_sampels\_split, min\_sample\_leaf, n\_estimators, max\_depth. While tuning the parameters min\_sampels\_split, min\_sample\_leaf (with below range of values) tend to cause issues to find the best model as it was varying a lot.

min\_samples\_split: [1.0, 10],

min\_samples\_leaf: [0.5, 5],

n\_estimators: [1000, 1589, 2000],

max\_depth: [11,15,20]

We then tried to use only n\_estimators, max\_depth only for which we were able to get comparable models with range of features selected.

n\_estimators: [1000, 1589, 2000],

max\_depth: [11,15,20]

**Explanation of data pre-processing step**

As a part of data preprocessing, we have performed unsupervised filtering on features. We have filtered out the features with **higher correlation** (with a correlation threshold between 0.05 and 0.15) and obtained a final feature size around 6-9. We made use of Spearman Correlation which is a non-parametric measure of the monotonicity of the relationship between two data sets. Spearman correlation does not assume the datasets are normally distributed. The correlation threshold range varies between -1 to +1 with 0 implying no correlation.

After filtering out correlated features, they were added back to the training and test dataset and the classifier model was fit with the training dataset.

**How the best model was selected with a clear definition of the specific measure or statistic used to select the model,**

The best model was selected after passing the classifier, range of parameters, number of folds, performance scoring measure to the GridSearch Cross Validation algorithm (using sklearn library). The Grid Search Cross Validation performs an exhaustive search over the range of specified parameter values for a particular classifier. The grid search internally performs K-Fold cross validation along with parameter tuning and determines the hyper parameters for a particular classifier.

We determine the hyperparameters as well as the average precision and recall score for every fold using the grid search. The model chosen for the type of classifier generates a precision recall-curve that is reasonable, hence chosen as the best model.

Performance measure chosen:

Precision: It is the measure of relevant instances among all the instances. Also known a positive predictive rate which is True Positives / (True Positives + False Positives)

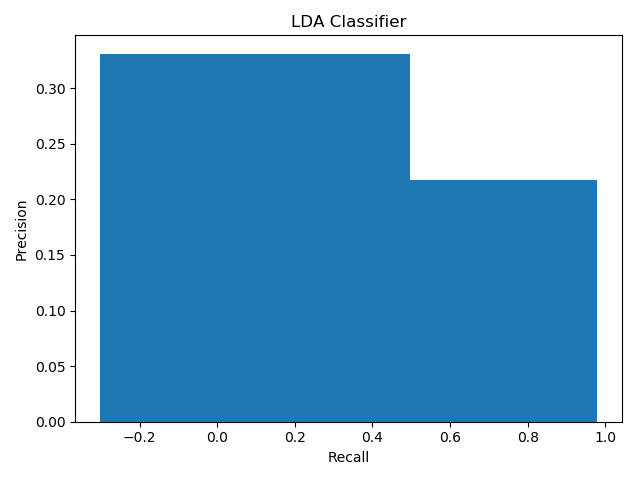
Recall: It is the fraction of relevant instances retrieved over all the instances. Also known as True positive rate which is True Positives / (True Positives + False Negatives)

LDA Model selected:

Number of features: 4

The parameters combination that would give best accuracy is : ***{'n\_components': 0, 'solver': 'svd', 'store\_covariance': True}***

**a figure showing a graphical representation of the cross-validation performance of the best models (one model per classifier evaluated) with mean and standard deviation. This description has to be submitted as a PDF file.**



The following model and ML method were not selected because it was having very high variance. When the size of feature space was increased, it was not generating a reasonable PR curve.