DV 2578 Assignment 2 Report

Name: Willford Zhan

# Introduction

In this assignment, Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) classifier are picked as the supervised learning classifier for algorithm comparison.

To exclude the effect of inappropriate dataset division, 10-fold stratified cross-validation is used to divide the dataset. After that classifiers are trained on each fold. F-1 measure and accuracy are used to measure the predictive performance, and training time consumption are used to measure the computational performance.

After measurements are done, significance of difference among the 3 classifiers on each measuring metric is conformed existing by Friedman test. Then through testing the classifier pair by pair with Nemenyi test, I found that NB classifier has significant difference with RF classifier on all of F-1 measure, accuracy, and time consumption, NB and DT, DT and RF don't have significant difference on all of F-1 measure, accuracy, and time consumption according to Nemenyi post-hoc, which should have in reality.

# Implementation

## Select Classifiers

I imported the Bayes, DT, and RF classifier provided by *sklearn* library. These 3 are common and widely-used classifiers and their performance has been checked by lots of previous researches, therefore I selected them. With manually set parameters, the classifier can be easily initialized.

## Stratified 10-fold cross validation

To carry out 10-fold cross validation, the dataset is firstly stratified by the class label: ham or spam. In the original dataset the proportion of spam and ham is 39.4 : 60.6. Then divided into 10 fold subsets with the same size with stratification via the class "*StratifiedKFold"* provided by *sklearn.*

# Initialized to set 10 fold, and shuffle the data before division   
skf = StratifiedKFold(n\_splits=10,shuffle=True)   
# Stratified via label value of Y, and split the data into 10 fold   
skf.split(X,Y)

Each of the subset has same proportion of spam and ham, which is also 39.4 : 60.6.

## Model Training

Use the "*fit()*"function provided by each *sklearn* classifier to train the models with divided train sets. For example:

# initialize a GaussianNB Bayes model  
clf\_bayes = GaussianNB()  
# train the model with train set  
clf\_bayes.fit(X\_train, Y\_train)

## Measurement

* Time consumption: record the time stamps before and after model training, then subtract them to get time
* Accuracy: "*score()*" function of *sklearn* is used to get model accuracy on the test set
* F-1 measure: "f1\_score()" function of *sklearn* is is used to get model f1 score on the test set

## Friedman Test

The "*friedmanchisquare()*" function from *scipy.stats* library is used to carry out Friedman test. It firstly ranks the performance result of each classifier on each fold, then calculates the average rank of each classirfier performance. In the end, *Q* value is calculated according to the average rank with the equation below and returned.

图片包含 物体, 时钟, 手表

描述已自动生成

After the *Q* value is calculated, I compared the *Q* with 7.82, which in this case is the value of k = 3, alpha = 0.05 on *x2* distribution. If *Q* is greater, then significant difference exists among the performance result of 3 classifiers on a specific measuring metric.

## Nemenyi Post-hoc

If the result of Friedman test is positive, then Nemenyi test is carried out. Average rank generated by Friedman test will be used as the input of The Nemenyi Test. The pairwised average rank difference is compared with the critical difference (CD) calculated with the equation below.

图片包含 物体

描述已自动生成

In this assignment, as we have 3 classifiers to compare and subset test result in 10 folds, therefore, *k* = 3, and *n* = 10. And with alpha = 0.05, we can get q𝛼 = 2.343, and CD = 1.047. if the average rank difference between two classifiers is greater than CD, significant difference exists, otherwise not.

# Test Results

## Performance Details

The detailed procedure results of Step 2,3,4 for all measuring metrics are presented in the latter part of my code (the running output of 'step\_2\_3\_4()' function call), formatted as instructed (Table 12.4 and Table 12.8).

## Friedman Test result

For all of F1, Accuracy and Time consumption, significance of difference is proved existing among the 3 classifiers.

## Nemenyi Post-hoc result

With Nemenyi post-hoc, it's proved that NB classifier has significant difference with RF classifier on all of F-1 measure, accuracy, and time consumption, NB and DT, DT and RF don't have significant difference on all of F-1 measure, accuracy, and time consumption according to Nemenyi post-hoc.