# ASSIGNMENT 3

An analysis of various classifiers with selected parameter values, verified against a given set of verification methods for each classifier

Report Submission
Hsien-Ming Lee
Souradeep Sinha

# DATA: BANK NOTES

Dataset: reduced.banknote.data.arff
Performance Report: A3PerformanceResults(banknote).csv
Comparison Report: comp(banknote).csv

## CLASSIFIER: IBK (K NEAREST NEIGHBOR CLASSIFICATION)

**k = 1**
Statistics over all verification methods:
Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.001
**k = 3**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.003, Mean of average generalization errors: 0.002
**k=5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.005, Mean of average generalization errors: 0.003
**k=10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.01, Mean of average generalization errors: 0.004

## CLASSIFIER: J48 (PRUNED C4.5 DECISION TREE CLASSIFICATION)

**Minimum leaf size = 2**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.006, Mean of average generalization errors: 0.014
**Minimum leaf size = 5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.015, Mean of average generalization errors: 0.015
**Minimum leaf size = 10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.03, Mean of average generalization errors: 0.02
**Minimum leaf size = 30**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.06, Mean of average generalization errors: 0.04

## CLASSIFIER: NAÏVE BAYES CLASSIFICATION

Statistics over all verification methods:
Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.05

## CLASSIFIER: SMO (SUPPORT VECTOR MACHINE CLASSIFICATION)

**Kernel: Polynomial, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.02, Mean of average generalization errors: 0.006
**Kernel: Polynomial, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.01, Mean of average generalization errors: 0.006

**Kernel: Radial Basis Function, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.22, Mean of average generalization errors: 0.15
**Kernel: Radial Basis Function, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.035

## ANALYSIS

*All cases in which the difference between Generalization error and Resubstitution error is obtained negative, we consider them to be 0. This is done because a negative difference means the following:*

- *i) a very simple decision tree/model was built,*
- *ii) data points in training set occurs in sparse clusters with plenty outliers in a vector space, whereas data points in testing set (miraculously) lie in proximity to the training clusters, or*
- *iii) discrepancy in data values.*

*For simplicity of understanding, we treat all such differences as zero. But this consideration is neglected during the selection of best classification of an algorithm and only considered while selecting a verification method, by virtue of the suspicion that it might be caused by data discrepancies.*

*All decisions are based on the basis of difference between the average error rates. For similar differences, we go ahead and check the standard deviation of generalization error, and if it's still a tie, we go ahead and look up the paired T-test results. (This box applies to all the datasets.)*

Based on the means of the average resubstitution and generalization errors, we believe that overall, the k Nearest Neighbor algorithm runs better on this dataset, with low error rates per verification per run.

For kNN = 1, the best performance is exhibited by Cross Validation bearing minimum standard deviation and minimum average error difference for 10 runs.
For kNN = 3, the best performance is exhibited by Cross Validation bearing minimum standard deviation and minimum average error difference for 10 runs.
For kNN = 5, the best performance is exhibited by Cross Validation bearing minimum standard deviation and minimum average error difference for 10 runs.
For kNN = 10, the best performance is exhibited by LOOCV bearing minimum standard deviation for generalization error and minimum average error difference for 10 runs.

Among other algorithms and verification methods, best promise is shown by Support Vector Machine with a Polynomial Function kernel with C = 5.0 and C = 1.0, C4.5 Tree with minimum leaf size of 5, and Naïve Bayesian Classifier; all with the LOOCV verification method.

# DATA: CENSUS

Dataset: reduced.census.data.arff
Performance Report: A3PerformanceResults(census).csv
Comparison Report: comp(census).csv

## CLASSIFIER: IBK (K NEAREST NEIGHBOR CLASSIFICATION)

**k = 1**

Statistics over all verification methods:
Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.07
**k = 3**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.07, Mean of average generalization errors: 0.07
**k=5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.09, Mean of average generalization errors: 0.07
**k=10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.09, Mean of average generalization errors: 0.06

## CLASSIFIER: J48 (PRUNED C4.5 DECISION TREE CLASSIFICATION)

**Minimum leaf size = 2**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.08, Mean of average generalization errors: 0.065
**Minimum leaf size = 5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.07
**Minimum leaf size = 10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.065
**Minimum leaf size = 30**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.13, Mean of average generalization errors: 0.07

## CLASSIFIER: NAÏVE BAYES CLASSIFICATION

Statistics over all verification methods:
Mean of average resubstitution errors: 0.09, Mean of average generalization errors: 0.05

## CLASSIFIER: SMO (SUPPORT VECTOR MACHINE CLASSIFICATION)

**Kernel: Polynomial, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.07, Mean of average generalization errors: 0.07
**Kernel: Polynomial, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.06, Mean of average generalization errors: 0.07
**Kernel: Radial Basis Function, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.15, Mean of average generalization errors: 0.09
**Kernel: Radial Basis Function, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.11, Mean of average generalization errors: 0.064

# ANALYSIS

Based on the data generated by the Driver.java, Support Vector Machine (SMO) classifier, it performs better than the others. It has lower average of resubstitution and generalization error compared with the rest of the classifiers' results. Although K-nearest neighbor (IBK) has similar performance but IBK's results have higher standard deviations. It means that the results of the 10 runs are more spread out compare with SMO.

Another reason to pick SMO is that it is less likely to be influenced by outliers and noises.
The best parameter for SMO is to use complexity setting 5.0 and polynomial kernel. From the result of SMO, we can conclude those settings perform better and results in a lower average resubstitution and generalization errors.

# DATA: WINE QUALITY

Dataset: reduced.winequality.data.arff
Performance Report: A3PerformanceResults(winequality).csv
Comparison Report: comp(winequality).csv

## CLASSIFIER: IBK (K NEAREST NEIGHBOR CLASSIFICATION)

**k = 1**
Statistics over all verification methods:
Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.17
**k = 3**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.19, Mean of average generalization errors: 0.19
**k=5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.23, Mean of average generalization errors: 0.18
**k=10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.012, Mean of average generalization errors: 0.003

## CLASSIFIER: J48 (PRUNED C4.5 DECISION TREE CLASSIFICATION)

**Minimum leaf size = 2**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.09, Mean of average generalization errors: 0.18
**Minimum leaf size = 5**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.17, Mean of average generalization errors: 0.18
**Minimum leaf size = 10**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.23, Mean of average generalization errors: 0.18
**Minimum leaf size = 30**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.29, Mean of average generalization errors: 0.19

## CLASSIFIER: NAÏVE BAYES CLASSIFICATION

Statistics over all verification methods:

Mean of average resubstitution errors: 0.33, Mean of average generalization errors: 0.22

## CLASSIFIER: SMO (SUPPORT VECTOR MACHINE CLASSIFICATION)

**Kernel: Polynomial, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.29, Mean of average generalization errors: 0.18
**Kernel: Polynomial, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.28, Mean of average generalization errors: 0.17
**Kernel: Radial Basis Function, C: 1.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.36, Mean of average generalization errors: 0.2
**Kernel: Radial Basis Function, C: 5.0**
Statistics over all verification methods:
Mean of average resubstitution errors: 0.34, Mean of average generalization errors: 0.2

# ANALYSIS

Given the performance results obtained from Driver.java, we believe that C 4.5 decision tree algorithm would produce the best classification, based on the fact that it has comparatively much smaller average resubstitution and generalization errors. kNN performs a close second best with at par error rates, and standard deviations, but the paired t-test matriz being heavier towards the J48 region, meaning the similar statistical measures were unlikely caused by chance, and that the J48 would have a better chance of classifying more accurately.

We find that the best performance of C4.5 is exhibited with the least size of leave nodes being 30 and works best against the Resampling method of verification.