CSC 422/522: Automated Learning and Data Analysis
Homework 4: Due 4/16 @ 23:55

Student Name:
Student ID:
Section (422 or 522):

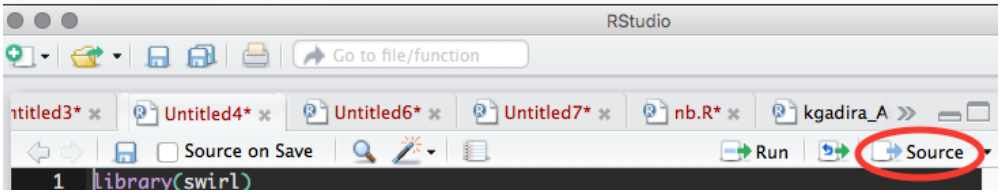
Important: Please follow submission instructions strictly; non-conforming submissions will not be evaluated.

Submission Instructions:

1. All files should be named appropriately as instructed.
2. Separate R script for each sub-question as required (using naming convention specified for each question)
3. Code should be executable and results should be reproducible
4. Include all required libraries in the code – Include a README document indicating which libraries need to be installed.
5. All plots should be included in a PDF document.
6. Submit your .R files, PDF file and the README document as a zipped folder with the name: student_lastname_id.fileExtension (For example: gadiraju_kgadira.zip)

Code Instructions:

1. At the top of your code use the following line:
   rm(list=ls(all=T))
   This will remove any existing global/hidden variables from your previous runs.
2. The TA will set the working directory accordingly on his local machine using setwd() function. **Don't set working directory in your code.**
3. Following up on the previous instruction, when you read the csv files, only use the file name, don't point it to any of your local paths.
   For eg: read.csv('dtr3.csv',header=F) - CORRECT
   read.csv('~/hogwarts/gryffindor/dtr3.csv', header=F) – INCORRECT
4. Don't print out any other results other than what the question asks for.
5. In addition, follow any instructions given specific to each question.
6. To run your code, if you are using RStudio, use the "Source" button at the top right corner.



Q1. Comparison of classification schemes (Weightage: 40 points CSC422; and 20 points for CSC522)

(a) Use the supplied data (Training: dtr3.csv and dtr8.csv; and Testing: dte3.csv, dte8.csv).

(b) Data description: data contains normalized handwritten digits of 3 and 8. Actual data is generated from scanning the handwritten digits, but processed (normalized) into 16x16 gray scale images.

(c) Training data contains 1200 samples (658+542=1200 rows), where each sample (row) contains label (e.g., 3 or 8) in the first column followed by 256 grayscale values (16x16=256).

(d) Test data contains 332 samples (166+166=332); same format as training data.

(e) As you learned from the class, (i) use training data to build a classification model, (ii) apply the model on test data to predict the label for each sample, and (iii) estimate accuracy (or error rate) by comparing predicted label vs. the expert given label (first column in the test data).

Answer the following question (a)

(a) Compare and contrast the following classifiers in terms of test accuracy

(i) Naïve Bayes Classifier. Name your file: nb.R
(ii) KNN Classifier (with K=3). Name your file: knn3.R
(iii) Linear SVM. Name your file: lsvm.R
(v) Decision Tree (use all three impurity measures) Name your files: dtEntropy.R, dtGini.R, dtError.R. Show the decision tree plots (add them to your PDF).
(vi) MLP. Name your file mlp.R

(1) Please note that if a classifier requires tuning parameters (e.g., number of hidden layers in MLP), experiment with at least 3 variations. If an algorithm requires more than one tuning parameters; experiment with most important parameter (that is, that parameter that impacts accuracy) by keeping others constant (with suitable constant).

(2) Your answer should consist of the following elements:

a. Confusion matrix
b. Individual class accuracy and overall class accuracy
c. Interpretation of your results, which method performed best and worst, and why?

For answer (2) a, b above in tabular format as shown below:

Method	Confusion Matrix	Individual Accuracy	Overall Accuracy

Tip: Since code for reading data and accuracy calculations are same for all the classifiers, you can write them as functions and then use the same function in all your files.

Q2. **Implement** maximum likelihood classification (MLC) in R. Name your file mlc.R (Weightage: 20 points for CSC 522. For CSC 422, this will be considered as bonus of 20 points)

(a) Implement model building (that is, computing mean and covariance matrix; you can use standard library functions for computing mean and covariance).
(b) Implement prediction (that is, implementation of discriminant function and assigning label) – this should be your own code (however you can use standard library functions for mean, covariance, inverse, transpose, etc.). The discriminant function is computed using the formula:

$$g_i(\bar{x}) = -\frac{1}{2}(\bar{x}-\bar{\mu_i})^T\Sigma_i^{-1}(\bar{x}-\bar{\mu_i}) + \ln P(w_i) - \frac{n}{2}\, ln2\Pi - \frac{1}{2}\ln|\Sigma_i|$$

where $g_i(\bar{x})$ is the discriminant value obtained from log-likelihood function for vector $\bar{x}$ for class i.

- $\bar{\mu_i}$ is the mean for class i for all three features (R,G,B) (it will be a vector of length 3)
- Σ_i is the covariance matrix for class i for all three features (R,G,B) (it will be a 3 by 3 matrix)
- $P(w_i)$ is the apriori probability and n is total number of samples.

(c) Apply your implementation on the data provided, and generate confusion matrix, individual class accuracy, and overall accuracy.
(d) Apply Naïve Bayes (NB) (use existing) on the same data; generate confusion matrix, individual class accuracy, and overall accuracy.
(e) Compare and contrast MLC and NB.
(f) Generate 2-d scatter plot using Red and Green features. Also plot mean and covariance for each class. Please note covariance matrix is plotted as an **ellipsoid**. Include these plots into your solution. Based on ellipsoids, comment on different classes and their separation.

Dataset description:

(a) Training and test samples are collected from a satellite image containing 3 features, Red, Green, and Blue.
(b) Both training and test data has header record (R,G,B,Class), and contain four columns, and 720 rows for training data (**img-train.txt)** and 720 rows for test data (**img-test.txt**.)

"**Compare and Contrast**" – describe key differences, advantages and disadvantages of the methods.