Homework 2 **CSCE 633**
Due: 11.59pm on October 7, 2019

Instructions for homework submission
Please submit on eCampus a single zip file containing your pdf solutions and your code. a) Submit PDF and .ipynb file (or similar code) of your solution Jupyter Notebook on eCampus, zipped into one file. b) Please write a brief report for the experimental problems. c) Please start early.

Implementation
For each programming assignment, please read the documentation of the libraries you are using and explain whether you need to change any default arguments.

Question 1: SVM (Ex 5.5 Mohri)
Please use the associated Satimage training and testing datasets for this problem.

(a) Normalize the data. Please note we have already split the data into training and testing vectors for you.

(b) (Hyperparameter Tuning) Consider the binary classification that consists of distinguishing class 6 from the rest of the data points. Use SVMs combined with polynomial kernels to solve this classification problem. For each value of the polynomial degree, d = 1, 2, 3, 4, plot the average 10-fold cross-validation error plus or minus one standard deviation as a function of C (let the other parameters of the polynomial kernels be equal to their default values) **ON THE TRAINING DATA**. Report the best value of the trade-off constant C measured on the training internal cross-validation.

(c) (Model Training and Testing) Let (C^*, d^*) be the best pair found previously in the 10-fold internal cross validation. Build a model for each pair on the full training data. Then plot the test errors for each model, as a function of d.

(d) (Results Evaluation) Plot the average number of support vectors obtained as a function of d

(e) (Results Evaluation) How many of the support vectors lie on the margin hyperplanes?

(f) (Conceptual) Explain how the parameter d influences the model fit (margin size and # number of support vectors).

(g) (Conceptual) Assume you were using an RBF kernel instead of polynomial kernel, what would the parameter γ influence in terms of the model fit (margin size and # number of support vectors).

Question 2: Ridge and Lasso
It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Assume n are the number of training samples, p the number of dimensions, x in the input and y the output.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, and $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Write out the lasso optimization problem in this setting.

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique – in other words, there are many possible solutions to the optimization problem. Describe these solutions.

Question 3: Random Forest

Classifying benign vs malignant tumors: We would like to classify if a tumor is benign or malign based on its attributes. We use data from the Breast Cancer Wisconsin Data Set of the UCI Machine Learning Repository:
`https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`.
The input file is named "hw2_ question3.csv" for our experiments. The rows of these files refer to the data samples, while the columns denote the features (columns 1-9) and the outcome variable (column 10), as described bellow:

1. Clump Thickness: discrete values $\{1, 10\}$
2. Uniformity of Cell Size: discrete values $\{1, 10\}$
3. Uniformity of Cell Shape: discrete values $\{1, 10\}$
4. Marginal Adhesion: discrete values $\{1, 10\}$
5. Single Epithelial Cell Size: discrete values $\{1, 10\}$
6. Bare Nuclei: discrete values $\{1, 10\}$
7. Bland Chromatin: discrete values $\{1, 10\}$
8. Normal Nucleoli: discrete values $\{1, 10\}$
9. Mitoses: discrete values $\{1, 10\}$
10. Class: 2 for benign, 4 for malignant (this is the **outcome** variable)

(a) Compute the number of samples belonging to the benign and the number of samples belonging to the malignant case. What do you observe? Are the two classes equally represented in the data? Separate the data into a train (2/3 of the data) and a test (1/3 of the data) set. Make sure that both classes are represented with the same proportion in both sets.

(b) *Implement* two decision trees using the training samples. The splitting criterion for the first one should be the entropy, while for the second one should be the gini index. Plot the

10-fold stratified cross validated accuracy on the train and test data while the maximum depth in the tree increases for both splitting criteria. Do you observe any differences in practice?

(c) *Feature Importance* Run a 10-fold stratified cross-validation for modeling this problem with a random forest model. You can choose to do a secondary 10-fold cross-validation on the training set in each fold to find the optimal number of trees and depth of each tree if you choose. Is this model more accurate (averaged over all 10 folds) than the decision trees?. As ranked by the feature importance, please provide a ranking of the features as their average position (standard deviation) of the rank order. Is this ranking it by Gini, Mean Decrease in Accuracy, or something else? What should the final model's list of features be? Please justify your answer.