CIS 490 MACHINE LEARNING HOMEWORK 3

(Full Score: **35** + 7 bonus points)

Due: Nov 29th, 2017 Fall

Group number:

Group leader:

Group members and student IDs:

(Please mark “G” after names for Graduate Students in your group)

INSTRUCTOR: Dr. Julia Hua Fang

# Overview of Homework 3 Score Distribution (35 points + 7 Bonus points)

1. CART, Bagging, Random Forest, and Boosting: 8 points.
2. Kmeans: 7 points
3. Hierarchical Clusteirng: 7 points
4. Gaussian Mixture Models: 6 points
5. Principal components Analyses: 7 points

Submission instruction:

1. Submit one **Word** file (e.g., \*.docx) to address all questions below and add an “Appendix” section at the end of your file to attach all scripts and outputs
2. Submit all your zipped folders that contain all your original data, scripts and output as instructed for specific questions in this homework.
3. (8 points) CART, Bagging, Random Forest, and Boosting, monte Carlo:
   1. What does CART stand for? (0.25 points)
   2. If you have a continuous outcome, what kind of trees would you use? (0.25 points)

A. Regression Tree; B. Classification Tree

* 1. If you have a discrete outcome, what kind of trees would you use? (0.25 points)

A. Regression Tree; B. Classification Tree

* 1. What algorithm does CART typically use? (0.25 points)
  2. What term is similar to “Regions” used in CART? (0.25 points)
  3. What is the goal of regression trees? (0.25 points)
  4. Why should we use cost complexity pruning (also called weakest link pruning)? (0.25 points)
  5. In the pruning procedure, what technique do we use to select the tuning/penalty parameter *which* controls a trade-off between the subtree’s complexity and its fit to the training data? (0.25 points)
  6. What are the options to be used as a criterion for making the binary splits in the classification trees? (0.25 points)

A. Classification error

B. Gini index (a measure of purity; select the split with the smallest Gini index)

C. Cross entropy/deviance/information gain

* 1. What is the tree size? (0.25 points)
  2. what computational approach does the tree-building process use? What is the weakness of a greedy approach? (0.25 points)
  3. Cross-validation (CV) can be used for (Hint: more than one option) (0.25 points)

1. choosing the optimal tuning parameter in regularized regression
2. choosing the optimal number of k for nearest neighbor classification
3. choosing the optimal tuning parameter in the pruning process of CART
   1. Write out decision tree algorithms including pruning. (2 points)
   2. Please use heart disease data, posted under Homework 3 on myCourses, to implement classification decision trees and replicate Figure 1, 2 and 3 shown on our lecture slides and below. (required for both undergraduates and graduates) (3 points)

Additionally, for graduates, please replicate figure 4 and 5 using bagging, random forest or boosting algorithms (2 points; for undergraduates, this will be counted as Bonus points)

**Instruction**:

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_Trees.R” (e.g., Group1\_Trees.R) or “Group#\_Trees.m” (e.g., “Group1\_Trees.m”).



Figure 1



Figure 2



Figure 3



Figure 4



Figure 5

1. Kmeans: 7 points
2. Describe Kmeans objective function (0.5 points)
3. Describe Kmeans algorithm (0.5 points)
4. What’s difference between forgy and random partition initialization? (0.5 points)
5. How to choose the number of K for the K-means algorithm? (0.5 points)
6. How to deal with random initialization issues in K-means? (0.5 points)
7. What algorithm can be used to deal with outliers, if k-means is sensitive to outliers? (0.5 points)
8. What are the assumptions for K-means? How to handle empty clusters in k-means? (0.5 points)
9. How can we prevent local minima resulting from K-means (0.5 points)
10. Implement Kmeans on Iris data. Generate plots as illustrated below in a) and b): (3points)
11. (1.5 points) the actual “scatter” plot when the number of clusters is chosen as 3, E.g,



1. (1.5 points) the plot where x = number of clusters, y = J, which is the cost value computed from

, similar to the graph below



(c) Implement Kmeans++ on IRIS and report the findings you find in comparison to Kmeans. (1.5 Bonus points for both Graduates and Undergraduates)

**Instruction**:

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_Kmeans.R” (e.g., Group1\_Kmeans.R) or “Group#\_Kmeans.m” (e.g., “Group1\_Kmeans.m”).

1. Hierarchical clustering (7 points)
2. what are the two hierarchical clustering methods/approaches? What are the differences between the two approaches? (0.5 points)
3. What are the 3 dissimilarity measures? What are the differences among these three measures? (0.5 points)
4. What is the dendrogram? What does the vertical axis or the height indicate? Do we draw conclusions about the similarity of two observations based on their proximity along the horizontal axis or the location on the vertical axis? (0.5 points)
5. Would different linkage affect the hierarchical clustering behavior? Would scaling change the clustering results? (0.5 points)
6. Describe Agglomerative Clustering Algorithm (1 point)
7. What criteria can we use to select the optimal number of clusters in hierarchical clustering? (0.5 points)
8. How many clusters do we have if we cut the height at 9 based on Fig (a) below? How about when cutting at a height of 5 in Fig (b) (0.5 points)



Fig (a) Fig (b)

1. Implement Agglomerative hierarchical clustering on IRIS data, using three linkage methods, respectively, and output dendrograms for each method. (3 points)

**Instruction**:

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_Hierarchical.R” (e.g., Group1\_ Hierarchical.R) or “Group#\_ Hierarchical.m” (e.g., “Group1\_ Hierarchical.m”).

1. Gaussian Mixture Models(GMM) (6 points)
2. When do we use GMM? What is the difference between GMM and Kmeans? (1.5 points)
3. What is responsibility in the context of GMM? (0.5 points)
4. Describe EM algorithm. (3 points)
5. Why do we use log-likelihood instead of likelihood in EM (1 points)
6. Based on IRIS data, simulate/sample data from the GMM using the GMM sampling procedure. Please use the seed 999 to simulate data and plot your simulated data, e.g., a faked plot as below. (2 Bonus points for undergraduates; required for graduates)



**Instruction**:

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_GMM.R” (e.g., Group1\_ GMM.R) or “Group#\_ GMM.m” (e.g., “Group1\_ GMM.m”).

1. Principal Components Analysis (PCA) (7 points)
2. What is the difference between PCA and clustering in the context of unsupervised learning? (1 point)
3. PCA looks at means or variance in the high-dimensional data? By saying principal component are assumed orthogonal to each other, what does orthogonal mean? (1 point)
4. Describe SVD based PCA algorithm (1.5 bonus points for both undergraduates and graduates)
5. What are the loading, eigenvalues, eigenvector in the context of PCA? (1.5 point)
6. Is the total variance of a dataset equal to the variance explained by components identified in PCA? (0.5 point)
7. What closeness measure does PCA use to find the hyperplane closest to the observations? (0.5 point)
8. Based on the loading matrix from the USarrests data, which variables will be counted into PC1 and which one will be counted into PC2? (0.5 point)



8)

(a) For the USarrest data, what do the blue state names represent on this bi-plot shown below ? What does the orange arrows indicate? (1 point)



(b) For the USarrest data, what does the scree plot (Fig 1) depict? What does Fig 2 depict? (1 point)



Fig 1. Fig 2.