Assignment #3

Q1 K-nearest neighbours

14.5 Course ratings. The Institute for Statistics Education at Statistics.com asks students to rate a variety of aspects of a course as soon as the student completes it. The Institute is contemplating instituting a recommendation system that would provide students with recommendations for additional courses as soon as they submit their rating for a completed course. Consider the excerpt from student ratings of online statistics courses shown in Table 14.16, and the problem of what to recommend to student E.N.

a. First consider a user-based collaborative filter. This requires computing correlations between all student pairs. For which students is it possible to compute correlations with E.N.? Compute them.

b. Based on the single nearest student to E.N., which single course should we recommend to E.N.? Explain why.

c. Use R (function *similarity()*) to compute the cosine similarity between users.

TABLE 14.16 RATINGS OF ONLINE STATISTICS COURSES: 4 = BEST, 1 = WORST, BLANK = NOT TAKEN

	SQL	Spatial	PA 1	DM in R	Python	Forecast	R Prog	Hadoop	Regression
L N	4				3	2	4		2
M H	3	4			4				
J H	2	2							
E N	**4**			**4**			**4**		**3**
D U	4	4							
F L		4							
G L		4							
A H		3							
S A			4						
R W			2					4	
B A			4						
M G			4			4			
A F			4						
K G			3						
D S	4			2			4		

d.Using the csv file for course ratings, apply item-based collaborative filtering to this dataset (using R) and based on the results, recommend a course to E.N.

Q2: Association rules

Identifying Course Combinations. The Institute for Statistics Education at Statistics.com offers online courses in statistics and analytics, and is seeking information that will help in packaging and sequencing courses. Consider the data in the file *CourseTopics.csv*, the first few rows of which are shown in Table 14.13. These data are for purchases of online statistics courses at Statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and bundling of courses. Use association rules to analyze these data, and interpret several of the resulting rules.

TABLE 14.13 DATA ON PURCHASES OF ONLINE STATISTICS COURSES

Intro	DataMining	Survey	CatData	Regression	Forecast	DOE	SW
1	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	0	1	1	0	0	1
1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	1	0	1	1	1
1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0

Q 3: A comparison between neural networks and decision trees

Carefully read and follow these steps:
Go to : http://aispace.org/dTree/index.shtml and in particular the two web pages for decision trees http://aispace.org/dTree/index.shtml
1. Use the provided tutorial to understand the details of how a decision tree learner works.
2. Now, run the provided applet for the decision tree algorithm by pressing the start applet box, http://www.aispace.org/dTree/dTree.jnlp , and open the electronics example.
File->load sample dataset->all electronics
This example uses the age, income, student/not a student, has a credit or not parameters to decide whether a customer will buy a computer. Build a decision tree following the tutorial instructions and click on different nodes to see their Gini scores.
3. Now you need to add two new parameters to this example: Married (yes/No) , # children (0, 1,2) such that :
Old parameters : AGE, INCOME, STUD, CRED, BUYS;
New parameters : AGE, Married, #children, INCOME, STUD, CRED, BUYS
To do this go to Edit ->view/edit text representation
4. Re-create the decision tree with the new data.
5. Double click on some node to check their Gini scores.
6. Change your data set again and see how it affects the Gini scores.
7. Repeat steps 3-6 using the neural network tool http://aispace.org/neural/. Replace the Gini scores with the weights.

Deliverables:
1. A set of rules extracted from the developed decision tree in step 4 (e.g., it could be that: students who are married with no kids will buy a computer)
2. Two snap shots (print screens) of the developed decision tree and the neural network after adding the two new parameters.

3. Answer the following questions:
A. Do you think the decision tree based data mining is comprehensible (to a human)? How about the neural network?
B. Can you create 5 new records for two new customers and compare the predictability of both the decision tree and the neural network?
C. What are the major advantages and problems of using each of these models?

Q4) K-Means Clustering

You are hereby provided with the Framingham data set. Using only the Sex and Age fields (ensure you standardize Age), complete the following:

a) Perform k-means clustering on the selected attributes, specifying k = 4 clusters and plot.

b) Apply the elbow method to determine the best k and plot.

c) Evaluate the quality of the clusters using the Silhouette Coefficient method.

2) Hierarchical Clustering

Complete this problem without the use of a computer to make sure that you understand the details of the clustering algorithms. Consider the following "data" to be clustered as described below.

10 20 40 80 85 121 160 168 195 For each part of the problem, assume that Euclidean distance will be used to measure the distance between the data points.

a) Use hierarchical agglomerative clustering with single linkage to cluster the data. Draw a dendrogram to illustrate your clustering and include a vertical axis with numerical labels indicating the height of each parental node in the dendrogram.

b) Repeat part (a) using hierarchical agglomerative clustering with complete linkage.

Part B: Model Evaluation & Performance Improvement

Customer churn is a huge problem for telecoms providers, considering an annual churn rate of 15-20% in some markets. To keep a low churn rate, telecoms providers need to predict which customers are likely to churn. You are provided with the customer_churn dataset from a telecoms company, complete the following:

a) Partition the data set using the holdout method, so that 67% of the records are included in the training data set and 33% are included in the test data set. Use a bar graph to confirm your proportions.

b) Identify the total number of records in the training data set and how many records in the training data set have a churn value of true (or 1). Calculate how many true churn records you need to resample in order to have 20% of the rebalanced data set have true churn values.

c) Perform the rebalancing described in (b) and confirm that 20% of the records in the rebalanced data set have true churn values.

d) Create a decision tree model that can predict Churn using the data set given. Use predictors you think are appropriate and obtain the predicted value.

e) Use an ensemble method (e.g., Random Forest, Adaboost) to obtain the predicted value of Churn. Tune the hyper-parameters (e.g., node size, max depth, max terminal nodes, etc.) of the ensemble model and compare against the initial model.

f) Using a confusion matrix, compare the evaluation measures from the ensemble method with the decision tree model based on the following criteria: Accuracy, Sensitivity and Specificity. Identify the model that performed best and worst according to each criterion.

g) Carry out a ROC analysis to compare the performance of the ensemble method with the decision tree technique. Plot the ROC graph of the models.