Assignment 4
CIS 453/553 Data Mining, Spring 2017
due 11:59 pm, Friday May 12th

1. Given a training table with T tuples and n attributes, show that the worst-case complexity of growing a decision tree is $n \times T \times log(T)$. (Hint: first show that the maxim depth of the tree is $log(T)$.)

2. A bank database has five attributes for customers.

credit-ranking	age	gender	year-income	count
Excellent	< 30	Male	60k - 100k	4
Excellent	< 30	Female	> 100k	16
Excellent	> 60	Male	> 100k	16
Excellent	> 60	Female	60k - 100k	4
Good	30 - 60	Male	< 60k	15
Good	30 - 60	Female	60k - 100k	5
Good	< 30	Male	> 100k	15
Good	< 30	Female	60k - 100k	5
Fair	> 60	Male	> 100k	18
Fair	30 - 60	Female	60k - 100k	18
Fair	< 30	Male	60k - 100k	2
Fair	< 30	Female	< 60k	2

Let *credit-ranking* be the class label.

(a) How would you modify the ID3 algorithm to take into consideration the *count* of each tuple?

(b) Build a decision tree based on your algorithm.

(c) Given a customer information: age is "< 30," gender is "Male," and year-income is "> 100k," what would a naive Bayesian classification of the *credit-ranking* for the customer be? Is it the same result if using your decision tree from (b)?

(d) If we want to use the given data to train the a neural network, such as Figure 9.2, what modification we need to do for the data tuples, considering there are only three input nodes and the neural network normally accepts numerical values?

3. Based on your understanding, please compare SVM with decision tree learning about their advantages and disadvantages.

4. Based on your understanding and reading, explain why deep learning models get very high accuracy (e.g., better than SVM in hand writing recognition).

5. The questions are based on the Excel dataset German.xls: `http://www.cs.uoregon.edu/classes/15S/cis453/data/German.xls`. The German Credit data set contains observations on 30 variables for 1000 past applicants

for credit. Each applicant was rated as "good" (700 cases fulfilled terms of credit agreement) or "bad" (300 cases defaulted on loan payments). New applicants for credit can also be evaluated on the 30 "predictor" variables. We want to develop a classification model that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. All the variables are explained in the codelist table and the data are in the data table of German.xls.

a) Use Weka to find the decision tree. Please choose one attribute as the class label based on your understanding of the problem. If the attribute is numerical or binary, you may use a way to change it to nominal. Report the general accuracy of the tree. The tree could be large, please report the path as a classification rule which represents the largest number of instances and report its accuracy. Please report the path as a classification rule for "good" applicants with the highest accuracy but represents at least 20 instances. Please report the path as a classification rule for "bad" applicants with the highest accuracy but represents at least 20 instances.

b) Please select at least around 11 attributes based on the results of a). Then apply Naive Bayes classifier in Weka to find the class label for an applicant with CHK_ACCT = 1, DURATION = 20, HISTORY = 3, EMPLOYMENT = 2 REAL_ESTATE = 0.

c) Based on the five attributes (CHK_ACCT, DURATION, HISTORY, EMPLOYMENT, REAL_ESTATE) and the class label attribute to build a neural network with Weka. Report the neural network in a graph with weights and thresholds (2 digits after point is fine). Calculate the output if the input is the same applicant with CHK_ACCT = 1, DURATION = 20, HISTORY = 3, EMPLOYMENT = 2 REAL_ESTATE = 0.

To turn in by paper version: Ask Cheri to put your answers to Prof. Dejing Dou's mailbox.

To turn in by emails: Send email to **dou@cs.uoregon.edu**. We prefer that you send in a pdf file. If you are using Word, you should be able to convert your word file to a pdf file.