ASSIGNMENT 6

Assignment 6 is described in the next few pages. Before that, here are some ground rules that everyone needs to follow:

- **Do not ask questions through email**. There is a Piazza discussion board at https://piazza.com/, where you can post questions, and the instructor and TA or even other students can answer them. This will prevent us from answering the same questions individually.
Again - absolutely no emails to the instructor or TA about this homework! All questions are answered on discussion board only.
You will receive an email from Piazza with details on how to join.

- The deadline is **fixed and final.** There will be no extensions for any reason. You should plan your time accordingly. In the past, I have received many requests at the last hour for extensions through email. From now on, these will simply be ignored. If you can't finish on time, you can submit whatever you have done.

- There are two parts to this assignment – Part I involves analysis using Weka and Part II has questions about Bayes net and d-separation.

- Part I of this assignment will use Weka. There is no programming needed, you will use the GUI to do all the steps. Weka can be downloaded from:
http://www.cs.waikato.ac.nz/ml/weka/downloading.html
You will find many good online tutorials for Weka online.

- It's fine to make any reasonable assumptions. State them in the README file.

- What to submit for Part I:
    - The tables as indicated in each of the sections. They can in a PDF or Word document.
    - The output files (from Weka) as indicated in each section.
    - README file that states any assumptions you made.

- For Part II, you should answer all questions and submit explanations where necessary.

- Part I is worth 150 points and Part II is 100 points.

ASSIGNMENT 6 PART I

This assignment will help you gain a better understanding of clustering as a pre-processing tool and its subsequent use for classification. In the second part, you will also get a chance to perform PCA on the data and find the most useful features.

Background:

The dataset in this assignment has been taken from:
"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring"[1], T. R. Golub et al., Science, Volume 286, 1999

The dataset consists of a training set in the file golub-data-train.csv available for download here:
www.utdallas.edu/~axn112530/cs6375/assignment6/golub-data-train.csv.

It consists of gene expression profiles for 38 bone marrow samples from acute leukemia patients, with each profile consisting of about 7000 gene expression levels. This means that there are 38 patients (instances) and 7000 genes (features) have been examined to see the type of cancer.

Here is some background information – there are two major classes in the dataset - ALL (acute lymphoid leukemia) and AML (acute myeloid leukemia), which are two clinically distinct types of leukemia. The ALL type samples can further be divided into T-lineage ALL (ALL-T) and B-lineage ALL (ALL-B) (see paper for details). So, there could also be 3 classes in the dataset with two classes hierarchically grouped into one bigger class. The true labels for each instance are available in the prefix of the first column –"Gene Accession Number".

I. Clustering without PCA: (50 points)

1. Your first task is to run a clustering algorithm to predict the labels for each instance. This means you do not look at the true label, and run the clustering algorithm on the dataset to obtain the predicted class.
You are to run at least two clustering algorithms → k-means and hierarchical clustering. You should run your experiments with k=2 and k=3 both. Weka also has a seed parameter for clustering. You can change that to obtain best clustering results.

Note: Weka will output the results as cluster0, cluster1, and so on. You will have to look at the true labels to find correlation with the classes described above. To clarify, if the cluster 0 contains 90% instances of class AML, then you can assume cluster 0 corresponds to AML.

[1] available for download at: http://www.cs.columbia.edu/~cleslie/cs4761/golub.pdf

Deliverable of this step:

- You should output the cluster assignment file for any one of your clustering experiments. It can be generated by right clicking result list and then clicking on "Visualize Cluster Assignments" and then clicking save.
- You should output results as follows:

1. k-means with k = 2: SSE = _____

Confusion matrix (Predicted vs Actual)

…

2. k-means with k = 3: SSE = _____

Confusion matrix (Predicted vs Actual)

Similarly for hierarchical clustering..

II. Clustering after PCA: (20 points)

In the dataset, there are over 7000 attributes (referred to as gene expression levels). In the second part, you will first perform a PCA analysis to find the most important attributes.

Note: PCA can be run in Weka by clicking on the "Select Attributes" tab and then selecting "Principal Components" after clicking the "Attribute Evaluator" button.
** This is a costly process and will take time. You should also find the best parameters that can finish this in a reasonable time **

Run the clustering in Step I again after PCA and see if you get an improvement in SSE and confusion matrix values. Turn in the same items as indicated in Part I.

III. Classification (80 points)

In the third part, you will use Weka to build a classification model of the training data. You can load the data, go to classify tab and then build the model using at least four different algorithms (preferably ensemble methods). You can read more about how to classify using ensemble methods here:

http://machinelearningmastery.com/improve-machine-learning-results-with-boosting-bagging-and-blending-ensemble-methods-in-weka/

Deliverable:

You should construct a table similar to the one below:

Algorithm	Parameters	% Training Accuracy
Random Forest	…	…
Boosting		
J48		
…		

You should also include any one of your models along with your submission.

Finally, you should test your model on the test dataset, which is available here:
http://www.utdallas.edu/~axn112530/cs6375/assignment6/golub-data-test.csv

You can see the steps for predicting using Weka here:
http://www.answermysearches.com/how-to-actually-use-a-saved-model-in-weka/199/

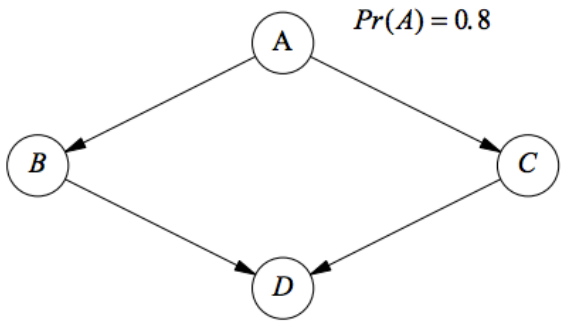
Deliverable:

Algorithm	Parameters	% Test Accuracy
Random Forest	…	…
Boosting		
J48		
…		

ASSIGNMENT 6 – PART II

1. (5 points)

Given below is a Bayes Belief Network with 4 nodes representing Boolean random variables A, B, C, and D. Compute $Pr(C \wedge D)$. Here you should first show the formula for $Pr(C \wedge D)$ and then use the Bayes Network semantics to simplify the formula, and then plug in the probability values to derive the final result.



$Pr(A) = 0.8$

$Pr(B|A) = 0.4$

$Pr(B|\neg A) = 0.9$

$Pr(C|A) = 0.6$

$Pr(C|\neg A) = 0.2$

$Pr(D|B, C) = 0.8$

$Pr(D|B, \neg C) = 0.5$

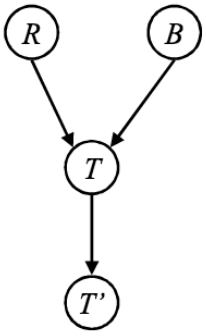
$Pr(D|\neg B, C) = 0.6$

$Pr(D|\neg B, \neg C) = 0.3$

2. It has been estimated that .05 percent of the US population has HIV. There is a test for HIV: If a person has HIV, the test has a 98% chance of being positive. If a person does not have HIV, the test has a 3% chance of being positive. Tom has tested positive. Assuming these values, what is the probability that he has HIV? Show your work.

(10 points)

3. Given the following Bayes Net (10 points)



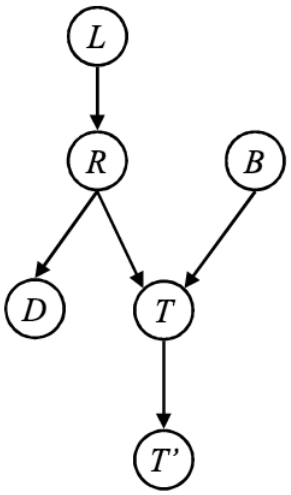
determine and explain whether the following are conditionally independent?

$R \perp\!\!\!\perp B$

$R \perp\!\!\!\perp B|T$

$R \perp\!\!\!\perp B|T'$

4. Given the Bayes net below, (10 points)



explain whether the following are conditionally independent?

$L \perp\!\!\!\perp T' | T$

$L \perp\!\!\!\perp B$

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$

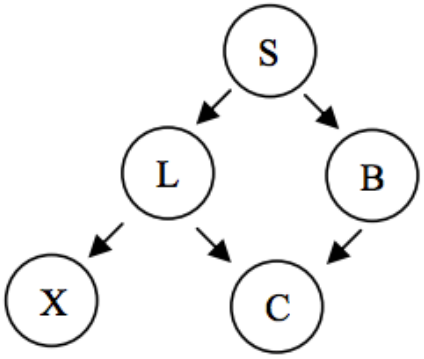
5 (10 points each = 20 points)
5a. Suppose you have a set consisting on n discrete random variables, which can each take m distinct values. Assuming they are totally independent, indicate how many entries will need to be there in the probability table to specify the full joint distribution for the following:

a. n = 20 and m = 2
b. n = 20 and m = 5
c. n = 500 and m = 10

5 b. Now suppose the n variables form a Bayes net such that the network has one root node, one node with a single parent, two nodes with two parents, and the remaining all have 3 parents each. For the following values on n and m, indicate the total size of the conditional probability table? You can assume that we are storing the full information for each node i.e. for each variable we are storing $P(X \mid Pa_x)$ as well as $P(\overline{X} \mid Pa_x)$.

a. n = 20 and m = 2
b. n = 20 and m = 5
c. n = 500 and m = 10

6. Consider the following Bayes net: (10 points)



Identify all pairs of nodes that are conditionally independent (CI), when you have following evidence:

a. S i.e. you know value of S, find all pairs that are CI
b. L
c. {L, B}

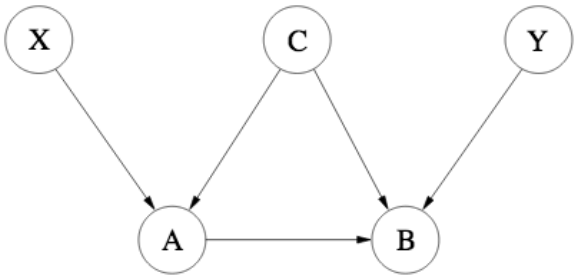
7. (20 points)

"A Lecturer's Life." *Dr. Ann Nicholson spends 60% of her work time in her office. The rest of her work time is spent elsewhere. When Ann is in her office, half the time her light is off (when she is trying to hide from students and get research done). When she is not in her office, she leaves her light on only 5% of the time. 80% of the time she is in her office, Ann is logged onto the computer. Because she sometimes logs onto the computer from home, 10% of the time she is not in her office, she is still logged onto the computer.*

1. Construct a Bayesian network to represent the "Lecturer's Life" scenario just described.
2. Suppose a student checks Dr. Nicholson's login status and sees that she is logged on. What effect does this have on the student's belief that Dr. Nicholson's light is on?

8. (15 points)

Consider the following graph.



1. Find all the sets of nodes that d-separate X and Y (not including either X or Y in such sets).