CS536 Machine Learning
Spring 2007
Assignment 1
Due Date March 7th 2007

Submission instruction: submit your work in the TA's mailbox. Any questions related to Weka should be directed to the TA.

[Q1 – 10 points] Consider the following set of training examples

Instance	A1	A2	Classification
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

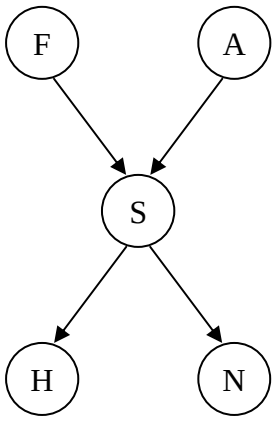
a. What is the entropy of this collection of training examples with respect to the target function?
b. What is the information gain of both A1 and A2 attributes relative to these training examples?

[Q2 – 10 points] As discussed in class, any joint probability distribution can be decomposed using a chain rule as follows:

$$P(x_1,x_2,\ldots,x_n)=P(x_1)\prod_{i=2}^{n} P(x_i|x_1,\ldots,x_{i-1})$$

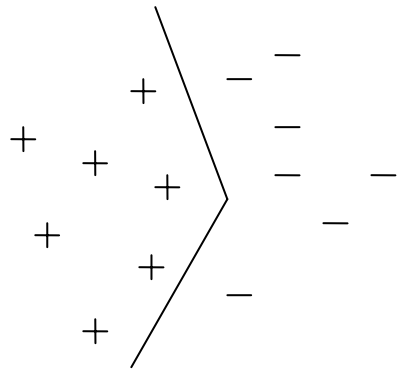
Using the chain rule and the independence and conditional independence assumptions made in the shown Bayesian network, prove that

$$P(F,A,S,H,N)=P(F)P(A)P(S|F,A)P(H|S)P(N|S)$$



[Q3 – 10 points] True or False: If a decision tree D2 is an elaboration of tree D1, then D1 is more-general-than D2. Assume D1 and D2 are decision trees representing arbitrary Boolean function, and that D2 is an elaboration of D1 if ID3 could extend D1 into D2. If true, give a proof; if false give a counterexample.

[Q4 – 10 points] Consider the space of linear "hinges" consisting of two line segments joined at a point. The drawing below shows a linear hinge separating positive and negative examples.



(a) What is the VC dimension of linear hinges in 2 dimensions? Explain why (with diagrams if you like).
(b) What is the minimum number of parameters needed to specify a general linear hinge in 2 dimensions? Explain your answer.

[Q5 40 Points USING WEKA] Learning decision trees.

For this question you need to submit only answers for parts 6 and 8.

1. Download the two datasets, trndata and tstdata from http://www.cs.rutgers.edu/~elgammal/classes/cs536/HW1data/ . We want to build a decision tree such that given someone's age, work class, occupation etc., we will be able to tell whether he earns more than 50k per year or not. This is a binary classification where class={>50, <=50}. The attr.txt file describes the variables used in these datasets.
2. Generate ARFF for both of the files. Please read the tutorial that comes with weka to learn how to convert txt/xls files to arff. Name the dataset (i.e. Relation) as income.
3. Open trndata.arff in the explorer window of weka. Skip the preprocessing for now and choose the classify tab. Choose J48 tree classifier and set tstdata.arff as test set. Press the start button and observe the output.
4. In this step, we will generate random subsets of the training data.
    a. Go to preprocess section of explorer window and choose supervised instance based filter Resample.
    b. By right clicking on the filter text box, select sampleSizePercent 10.
    c. Click apply and save the data subset (i.e. the relation as trndatat_10_0.arff. Open this file in any text editor (e.g. Wordpad)and change the name of the relation as income.
    d. Click again on open file, and select trndata.arff.
    e. Repeat steps a~d for sample size percentage m=20,40,50,70,90 , save the result as trndata_n_0.arff and rename all datasets as 'income'.

5. Repeat 4, but this time set the biasToUniformClass parameter of the filter Resample to 1.0 (click on the text box which displays the name of the filter and then click More button on the dialog box that appears next to learn about this biased sampling). Save the resulting datasets as trndatat_m_1.arff where m=10, 20,40,50,70,90. Again, change names of all the relations to 'income'.
6. Use these 12 datasets as the training sets for the decision tree classifier J48 (select tstdata.arff as the test set). As we know six of them were generated by sampling that tried to retain the distribution of the 'class' variable. In the other 6 datasets, the 'class' distributions were tried to make a uniform one. Plot two curves for the percentage of correctly classified instances of these two types of training sets. Can you explain the graph? Hint: Look at the distribution of the variable class in the training and test sets. please explain why the two curves are same or different.

7. J48 tree classifier uses Rule-post pruning by default. Turn off this feature by clicking on the text box and selecting true for unpruned option. Then press start button to implement C4.5 without pruning.
8. Now set reducedErrorPruning to true and observe the output. Briefly state the difference between these two pruning using the trees generated (unpruned, pruning by rule-post scheme and by Reduced error pruning). Note: learn the decision tree on the original datasets, trndata and tstdata (NOT on the subsets created in steps 4~5).