Take Home Final for CSE 490i, Winter 2002

Note: this exam is open book and open Web, but please do not discuss any question with any mammal besides Adam or Dan.

Due by 12:20pm sharp, Thurs 3/21. To hand in, either email a word or ASCII plain text document with your answers to both Adam and Dan or (if you have to) deliver a hardcopy to one of us in person (Dan will be in Sieg 422 on Thurs from 10:30am to noon; Adam will be in Sieg between noon and 12:20) or put it under Dan's office door. Good luck.

1) You are building a datamining system to analyze weblogs and you've decided to use a decision tree learning system. You write a perl script to process the log files and cull out the key attributes, A_i, e.g., features like whether they have followed a specific link and whether they have registered by logging in etc. Now comes the time to learn a tree that predicts whether they will make a purchase. Consider the following set of examples (+ indicates an example where the customer made a purchase):

Instance	Classification	A_1	A_2	A_3
1	+	T	T	F
2	-	T	F	T
3	+	T	T	F
4	-	T	F	T
5	+	F	F	T
6	-	F	T	F
7	-	F	T	F
8	-	F	T	F
9	+	T	F	F
10	+	F	T	T

A) What is the entropy of this collection of training example with respect to the target function classification?

B) What is the information gain of A_1, A_2 and A_3 relative to the training examples?

C) Draw the decision tree which would be found using greedy (hill climbing) search (maximizing information gain), which continued expanding the tree until no choice provided information gain. If the metric doesn't discriminate between attributes, break ties by choosing the lower numbered attribute.

2) Does SSL provide user authentication? How? If not, why not and what needs to be added?

3) What is the difference between a hash function and a cryptographic hash? What's an example of the latter?

4) What is the vector space angle between each pair of the following sentences. (Ignoring stopwords: an, a, out) Show all intermediate work.

A) Fruit flies like a banana.

B) Tent flies keep out Tsetse flies.

C) Time flies like an arrow.

5) We studied two techniques for organizing the results returned by search engines: clustering and classification.

A) List two differences between the approaches. Which did you find more compelling (i.e. if you were going to add one to your commercial search engine, which would you implement? And why?)

B) If you were forced to use the classification approach, what is an advantage of using k binary classifiers instead of one k-ary classifier?

6) Suppose you are hired as a consultant by a somehow-still-alive-dot-com that wants to better understand the behavior of their customers. You wisely decide to initiate a datamining project that will process the server logs to build models of customer behavior. You recall that Markov models are a good technique, but should you use first-order or second-order Markov models? They each have advantages and disadvantages.

A) Give a reason to prefer first-order

B) And a reason to prefer second-order