FALL 2007
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 17, 2007
3:00 – 7:00 p.m.

GENERAL INSTRUCTIONS:

(a) Answer each question in a separate book.

(b) Indicate on the cover of *each* book the area of the exam, your code number, and the question answered in that book. On *one* of your books, list the numbers of *all* the questions answered. *Do not write your name on any answer book.*

(c) Return all answer books in the folder provided. Additional answer books are available if needed.

SPECIFIC INSTRUCTIONS:

Answer:

- both questions in the section labeled B760 or B766, corresponding to your chosen focus area, *and*

- any two additional questions in the sections B760, B766, B776 and B838, where these two questions need *not* come from the same section, *and*

- both questions in the section labeled A760 or A766 that corresponds to your focus area.

Hence, you are to answer a total of six questions.

POLICY ON MISPRINTS AND AMBIGUITIES:

The Exam Committee tries to proofread the exam as carefully as possible. Nevertheless, the exam sometimes contains misprints and ambiguities. If you are convinced that a problem has been stated incorrectly, mention this to the proctor. If necessary, the proctor can contact a representative of the area to resolve problems during the *first hour* of the exam. In any case, you should indicate your interpretation of the problem in your written answer. Your interpretation should be such that the problem is nontrivial.

Answer both of the questions in the section labeled B7xx that corresponds to your chosen focus area. Also answer any two additional questions in any of the other sections (these two questions need NOT occur in the same section).

B760 – MACHINE LEARNING: BASIC QUESTIONS

B760-1 Supervised Learning

You are given a data set for a two-class supervised learning problem with 100 examples of each class and 10,000 Boolean-valued features.

(a) Without knowing anything else about the task, do you expect Naïve Bayes or a decision tree learner (say, C4.5) is likely to perform better on this data set? Why?

(b) Describe a scenario in which this expectation might be incorrect. Explain your answer.

(c) Does your answer to (a) change if you are permitted to use *bagging* with each approach? Why or why not?

(d) What procedure and statistical test can you employ to determine whether one of the algorithms in (a) is significantly better than the other on this data set?

B760-2 Reinforcement Learning

Consider a deterministic reinforcement-learning (RL) task, where

- the states are represented by three Boolean-valued features, f_1, f_2, and f_3,
- action 0 ends the episode, while action *i* (*i* > 0) has the effect of flipping the value of feature f_i,
- the reward for entering state *S*

  equals 2 when all feature values in *S* equal 1
  equals 1 when $f_1 = 1$ and at least one other feature equals 0
  equals 0 otherwise.

(a) If the discount rate equals 0.5, what is an optimal *policy* at each of the following states

$< f_1 = 0,\ f_2 = 0,\ f_3 = 0>$

$< f_1 = 1,\ f_2 = 1,\ f_3 = 1>$

Be sure to explain your answers (note that it is *not* necessary to compute the complete Q-function in order to answer this question).

(b) Assume you are employing two-step Q-learning, where initially the Q-function equals 0.1 for all possible inputs. Consider the episode where the RL agent starts in the state $< f_1 = 0,\ f_2 = 0,\ f_3 = 0>$, then takes action 2, followed by action 1, and ending with action 0.

If the Q-function is represented as a Q-*table*, which cells change? Show the calculations that produce their new values.

(c) Using the episode in (b), concretely explain the need to occasionally explore in RL, rather than always exploiting the current Q-function. Be sure to clearly explain what the terms *explore* and *exploit* mean in RL.

(d) Imagine that in some new RL task, states are represented by ten Boolean-valued features, but unbeknownst to the RL agent, only the first five features influence the effect of actions and the rewards received. Discuss one major weakness of Q-tables that this scenario highlights and briefly discuss how this weakness is typically addressed.

B766 – COMPUTER VISION: BASIC QUESTIONS

B766-1 Stereo Vision

(a) Describe the process of rectification that is commonly performed as part of a stereo algorithm, including what is its purpose.

(b) What parameters affect the accuracy of the depth estimate to a scene point **P** at actual depth z_P assuming the only source of noise is the localization of corresponding points in two images from two pinhole cameras? Give the parameters and how (qualitatively) varying each will affect the accuracy of the depth estimate.

(c) What is the fundamental matrix and describe <u>two</u> of its main properties.

(d) What is the *ordering* constraint (also known as the *monotonicity* constraint) that is sometimes used in solving the correspondence problem? Give <u>one</u> major advantage and <u>one</u> major disadvantage of using this constraint in a stereo system.

B766-2 Smoothing and Edge Detection

A grayscale image, $I(x,y)$, is to be smoothed by convolution with a discrete approximation of a 2D Gaussian kernel, $G_\sigma(x,y)$, of size $(2k+1) \times (2k+1)$ pixels.

(a) Give an expression for computing the intensity of a smoothed pixel, $S(x,y)$. Why does smoothing before edge detection help?

(b) Show how the convolution can be performed by two discrete 1D convolutions, and comment on the computational time savings this gives compared to using one 2D convolution.

(c) Describe and compare Burt and Adelson's Difference-of-Gaussian method for detecting edges with the Canny operator in terms of the types of edges they can and cannot detect, and their ability to correctly localize intensity edges.

(d) Give an expression for computing the directional derivative of I in the direction **n**, where **n** is a unit vector.

B776 -- ADVANCED BIOINFORMATICS: BASIC QUESTIONS

B776-1 Parsing in HMMs and SCFGs

Parsing with HMMs and SCFGs is typically done with the Viterbi and CYK algorithms, respectively, which find the *maximum a posteriori* (MAP) parse of a sequence. An alternative to MAP parsing, however, is *posterior decoding*.

(a) Define, for HMMs, what is computed by posterior decoding.

(b) Describe the calculations necessary to do posterior decoding for HMMs.

(c) Discuss one significant advantage of posterior decoding over Viterbi parsing.

(d) Discuss one significant limitation of posterior decoding.

(e) There are a number of possible definitions of posterior decoding for SCFGs. Give one such definition.

B776-2 K-Means Clustering

Suppose you wish to create groups of genes by clustering gene expression microarray data. You decide to use *k-means clustering*, but you do not know what value of *k* to use.

(a) Describe an approach for selecting a good value of *k*.

(b) Describe <u>one</u> strength and <u>one</u> weakness of your approach in (a).

(c) Can a similar issue (to choosing *k*) arise if you use *hierarchical clustering* instead of *k-means*? Explain your answer.

B838 -- ADVANCED NATURAL LANGUAGE PROCESSING: BASIC QUESTIONS

B838-1 Language Models

(a) Define an *n*-gram language model (LM) mathematically. Given a training corpus, define the maximum likelihood estimate of the *n*-gram LM.

(b) Given a test corpus, define one measure of LM quality.

(c) Consider a training corpus TRAIN and a test corpus TEST, both artificially generated from the same underlying *bigram* language model LMg. You train separately a unigram language model LM1, and a bigram language model LM2, using TRAIN. Both are maximum likelihood estimates (i.e., *not* smoothed). As the size of TRAIN and TEST approaches infinity, is LM1 or LM2 better on TEST? Briefly justify your answer using words.

(d) Same as (c), except that the underlying LMg is a *unigram* instead of a bigram.

(e) Same as (d) where LMg is a unigram, except that the size of TRAIN is *small* (the size of TEST still approaches infinity).

B838-2 Word Polarity

You have a dictionary with *N* words. Each word has a definition, for example

excellent	(very good; of the highest quality)
good	(having desirable or positive quality)
bad	(not achieving an adequate standard; poor)

You want to compute the polarity (positive or negative) of each word. Initially you only know that the word *good* is positive and *bad* is negative.

(a) Note that the word *good* appears in the definition of *excellent*. The intuition is that such word pairs tend to share the same polarity. Define a *feature function* that captures such an “in-definition” relationship.

(b) Discuss <u>one</u> significant weakness of this feature function.

(c) Despite (b), define a *graphical model* so that the inference problem attempts to compute the polarity of all words. Use the feature function defined in (a). Be sure to describe the structure and parameter(s) of your graphical model.

(d) Provide an expression for calculating the probability of the labeled data {(good, positive), (bad, negative)}. Describe how you can learn the model parameter(s) using the probability.

Answer both of the questions in the section labeled A760 or A766 that corresponds to your chosen focus area.

A760 – MACHINE LEARNING: ADVANCED QUESTIONS

A760-1 Spam Email Detection

You want to build a spam detector for a single email user. The detector takes a plain text email and outputs a binary spam / ham (not spam) label. The user has a training set in which each email is labeled. You may answer the following questions independent of each other.

(a) Instead of giving an explicit spam/ham training label, the user can either

- click "save to folder" which we take it to mean a ham
- click "report spam" button which is obviously a spam
- click "delete" which could mean either a spam or a ham (the user simply doesn't want to save it).

Give the pseudo code to use all three types of emails to train a spam/ham classifier. What assumptions does your method make?

(b) Suppose you have access to many other email users' training sets too, which are labeled with spam/ham. However, different users may have a different notion of spam. Describe one approach to utilize those data. Give one advantage and one disadvantage of your approach (it has to be smarter than simply pooling all the data together).

A760-2 Active Learning with Cost Information

Most approaches to *active learning* make the implicit assumption that there is a uniform cost to acquire the label of any given instance. Consider a situation in which you are using active learning for a classification task, and there is a large amount of variation in the cost of labeling individual instances. For example, your instances might be documents that vary significantly in their length, and the process of labeling may require reading the entirety of a given document.

(a) Describe one criterion for deciding which unlabeled instance is to be selected next for labeling in the *standard* active learning setting (i.e., in which labeling costs are not considered). You could describe a commonly used criterion or one of your own design.

(b) Suppose you can determine the *cost* of having any given instance labeled, before actually asking for it to be labeled. Discuss how you might use this cost information in conjunction with the criterion you described in (a).

(c) Now suppose that you do *not* have access to labeling costs before querying on instances. However, when a query is answered, the cost of answering it is also provided. How might you still take into account the differential costs of labeling instances in this situation?

(d) Describe how you would experimentally determine the effectiveness of the cost-sensitive methods you devised above.

This page intentionally left blank. You may use it for scratch paper. Please note that this page will NOT be considered during grading.