FALL 2008
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 15, 2008
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, corresponding to your chosen focus area, *and*
- any two additional questions in the sections Bxxx, where these two questions need *not* come from the same section, *and*
- both questions in the section labeled A760.

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 B760. Also answer any two additional questions in any of the other B sections (these two questions need NOT occur in the same section).

B760 – MACHINE LEARNING: BASIC QUESTIONS

B760-1. Supervised Learning Methods

Design algorithms for supervised machine learning that combine each of the following pairs of traditional approaches. Be sure to justify each design, by explaining how the combination addresses weaknesses of the individual approaches. (Do not say "combine using an ensemble method" for any of your answers.)

(a) Decision-tree induction and instance-based learning

(b) Genetic algorithms and neural networks

(c) Naïve Bayes and an inductive-logic learner

B760-2. PAC Learning

(a) Define the *VC dimension* of a concept class C over a domain of possible examples X.

(b) Explain the importance of VC dimension for PAC-learnability.

(c) Discuss the relationship of VC dimension to overfitting. Illustrate your answer with a concrete example involving at least two concept classes with different VC dimensions.

B766 – COMPUTER VISION: BASIC QUESTIONS

B766-1. Feature Point Detection and Description

(a) The SIFT feature point detector consists of four main steps:

i. scale-space analysis
ii. detection of a sparse set of feature points,
iii. orientation estimation at each feature point, and
iv. feature point description.

Briefly describe how each of these steps is performed.

(b) Describe two invariants that SIFT feature descriptors exhibit, and two invariants that SIFT feature descriptors do *not* have. Briefly explain each of your four answers.

(c) Given two images that are known to partially overlap and a set of extracted SIFT feature points and their descriptors for each image, how can this information be used to find a set of point correspondences between the two images?

B766-2. Stereo

(a) State the main advantage of using epipolar geometry in stereo matching.

(b) Consider a pair of stereo images. Assume the fundamental matrix for this pair is **F**. For one point **x** in one image, give the equation of the epipolar line in the other image.

(c) The fundamental matrix is a 3x3 matrix. To estimate this matrix, we need to solve for 9 unknowns. Why are 8 pairs of point correspondences enough for estimating the fundamental matrix?

(d) In the presence of errors in point correspondences, describe a robust approach to compute the fundamental matrix.

(e) Usually estimating a fundamental matrix from point correspondences is ill-conditioned numerically. What procedure can we perform to make the estimation well-conditioned?

B769 -- ADVANCED NATURAL LANGUAGE PROCESSING: BASIC QUESTIONS

B769-1. Conditional Random Fields and Part-of-Speech Tagging

A linear chain Conditional Random Field (CRF) is a probabilistic distribution of the form

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}, t) \right\},$$

where $\boldsymbol{x}$ is a sequence of input items, $\boldsymbol{y}$ is a sequence of labels, K is the number of features, and T is the length of the sequence.

(a) Define, in English, the feature functions $f_k()$. Be sure to explain the meaning of its arguments and its output.

(b) Explain the role of λ_k. What does the sign of λ_k mean, assuming f_k is non-negative?

(c) Describe, in English, the NLP task of Part-of-Speech (POS) tagging.

(d) Formulate POS tagging as a CRF. Be sure to explain what $\boldsymbol{x}, \boldsymbol{y}, K, T$ are for this task, and define <u>two</u> concrete feature functions that you think will work for POS tagging.

B769-2. The PageRank Algorithm

Given a set of Web pages and the hyperlinks among them, the PageRank algorithm constructs a directed graph and computes the "page rank" of each Web page.

(a) Define the directed graph using matrix notation.

(b) Define a single step of random walk on the graph, including teleporting.

(c) Explain (in English) the relation between the random walk on the graph and PageRank.

(d) Intuitively, what kind of Web pages receives high PageRank? What is a common method that spammers utilize to try to boost the PageRank of their Web sites?

B776 -- ADVANCED BIOINFORMATICS: BASIC QUESTIONS

B776-1. Classification of Biological Sequences

A biologist is generating a large amount of DNA sequence data from an environmental sample using one of the latest high throughput sequencers. The biologist wishes to identify which sequence reads contain part or all of a protein-coding gene. Reads classified as containing a gene will be examined in more detail later on. You are given the task of developing a classifier for labeling each read as either GENE or NOT_GENE. As a training set, you are given a set of sequences annotated with the positions of protein-coding genes.

(a) Describe how you would classify sequence reads using a *generative* probabilistic model (i.e., one which gives the joint distribution p(x,y) of the data, x, and the unobserved variables, y).

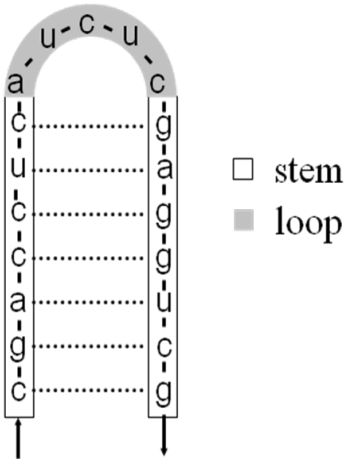
(b) Describe how you would classify sequence reads with a *discriminative* model (i.e., one that does not model the joint distribution p(x,y)).

(c) Which model would you expect to have better predictive accuracy? Briefly justify your answer.

B776-2. Stochastic Context-Free Grammars (SCFGs

)

Consider a probabilistic grammar for modeling RNA sequences that could fold into a "stem-loop" structure like the one shown below.



(a) Show the productions for a context-free grammar that characterizes such sequences. You should assume the following:
- the stem consists of at least three base pairs, but may be arbitrarily long,
- only canonical base pairings (i.e. a paired with u, and c paired with g) are present in the stem,
- the loop consists of one or more bases and can be arbitrarily large.

(b) Is your grammar *ambiguous*?

(c) Suppose you are given a set of RNA sequences that we believe contain stem-loop structures, and we want to train an SCFG to model this class of sequences. Describe how you would learn the probability parameters in a stochastic version of your grammar?

(d) Before training, we might initialize the probabilities to account for prior knowledge we have about the task (e.g. c-g pairings might be preferred over a-u pairings in the stem). Would this initialization affect the final result of the learning procedure for your grammar? Explain your answer.

Answer both of the questions in the section A760.

A760 – MACHINE LEARNING: ADVANCED QUESTIONS

A760-1. The Exploration-Exploitation Tradeoff in Reinforcement Learning

A central issue in reinforcement learning (RL) is the *exploration-exploitation* tradeoff.

(a) First define *exploration* and *exploitation*, and then explain why there is a tradeoff between the two.

(b) Describe and justify one approach for addressing the need for both exploration and exploitation.

(c) Assume you decide to employ *function approximation* (to learn Q functions, say), leaving all other aspects of your RL system unchanged. Does this increase, decrease, or not impact exploration? How about exploitation? Be sure to justify your answers.

(d) Does the exploration-exploitation tradeoff arise when using genetic algorithms for *supervised* machine learning? Explain your answer.

A760-2. Inductive Logic Programming

(a) Define *top-down* and *bottom-up* induction within inductive logic programming.

(b) Discuss one advantage of top-down induction over bottom-up induction.

(c) Discuss one advantage of bottom-up induction over top-down induction.

(d) Propose one hybrid induction method that can at least partially maintain the advantages of each. (Your hybrid method should be different in some way from the bottom-clause approach of Progol or Aleph, as well as different from a simple use of random clause generation coupled with local search.)

(e) Describe a type of learning task, or give a specific learning task, where your induction methods might outperform either pure top-down or bottom-up induction, and say why you believe it might do so.

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