SPRING 2010
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, February 1, 2010
3:00 – 7:00 p.m.

GENERAL INSTRUCTIONS:

(a) This exam has **14** numbered pages.

(b) Answer each question in a separate book.

(c) 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.*

(d) 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 Bxxx, where these two questions need *not* come from the same section, *and*
- both questions in the section labeled A760 or A766, again corresponding to your chosen 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 B760 if you are a machine-learning student or both in Section B766 if you are a computer-vision student. In addition, answer two more “B” questions (from any section).

B731 – ADVANCED ARTIFICIAL INTELLIGENCE: BASIC QUESTIONS

B731-1. Bayesian Network Parameter Learning

Consider the task of learning the parameters (conditional probabilities in the CPTs) of a Bayesian network with known structure. Assume that each row of each CPT is independent of the others, and you have Dirichlet-distributed priors on each row. Some variables of the Bayes net may be entirely unobserved and values for other variables in the data may be missing at random.

(a) Discuss two advantages of Expectation-Maximization (EM) over Gibbs sampling for this task. Show concrete Bayes nets that illustrate your answer.

(b) Discuss two advantages of Gibbs sampling over EM for this task. (You might want to read part (c) before choosing.) Show concrete Bayes nets that illustrate your answer.

(c) Choosing one of the advantages of Gibbs you listed in (b), is there a way to modify EM to at least partially negate this advantage? If so, how? If not, why not?

B731-2. First-Order Logic vs. Markov Networks for Knowledge Representation

You have been asked to build an expert system to select refreshments (food and drinks) for meetings. Your expert system should take into account time of day, whether attendees are vegetarians, and three additional properties of your choice. For simplicity, assume your expert system must recommend exactly one food item and one drink.

(a) Write such an expert system in Prolog.

(b) Write another such expert system as a Markov network (include the potentials).

(c) Discuss one advantage of each approach relative to the other for this particular task.

B760 – MACHINE LEARNING: BASIC QUESTIONS

B760-1. Inductive Logic Programming vs. Support Vector Machines

Suppose you are given access to a (de-identified) clinical database that contains the following tables:

Demographics:	birth date and gender of each patient
Diagnoses:	date and diagnosis of the patient at each doctor visit
Drugs:	date, drug, and dosage for each prescription for a patient
Labs:	date, lab test, and result for each lab test for a patient

You are asked to build a model to predict which patients are likely to have a stroke. You may assume you have data on 1,000 patients who suffered at least one stroke and a million patients who have not.   These patients have each visited the clinic from 1 to 10 times.

(a) Discuss how you would apply Inductive Logic Programming (ILP) to this task.

(b) Discuss how you would apply Support Vector Machines (SVMs) to this task

(c) Describe one advantage of ILP over SVMs for this task.

(d) Describe one advantage of SVMs over ILP for this task.

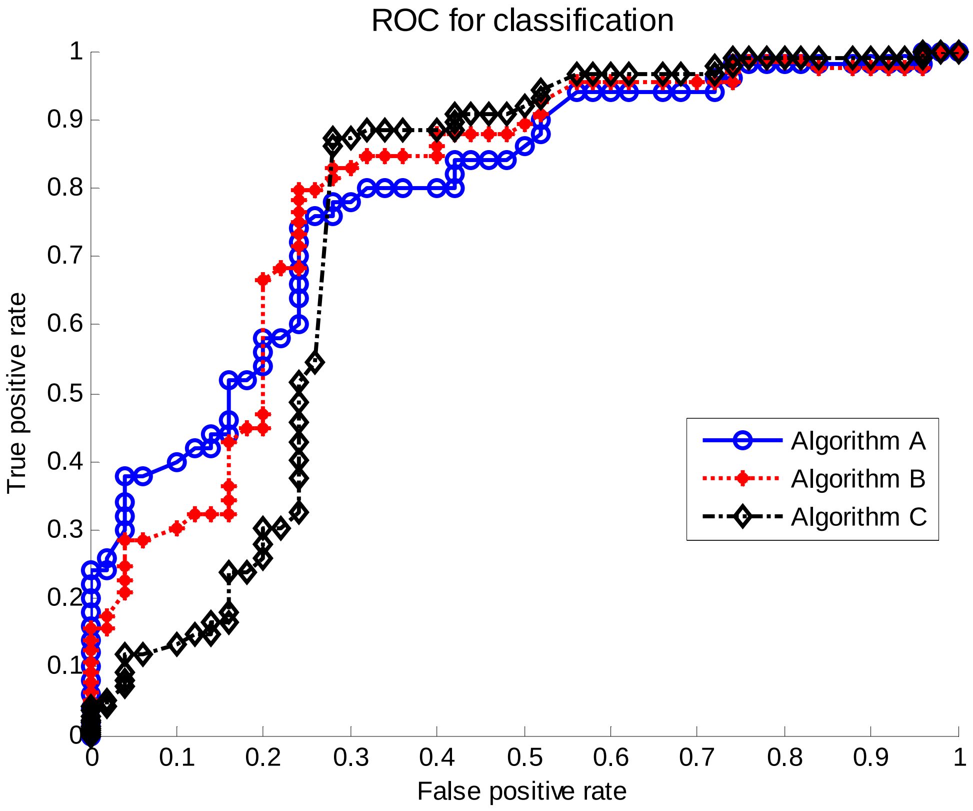
B760-2. Experimental Methodology

You are developing a new classification algorithm for your thesis. You decide to evaluate it using a holdout set consisting of 20% of your labeled data picked at random. To measure progress as you modify the algorithm, you decide to keep using the same holdout set for testing, to check if classification accuracy is improving or not.

(a) Is this a reasonable idea? Briefly explain why or why not.

(b) Instead, suppose you decide to use *k*-fold cross validation. However, every time you evaluate a new version of your algorithm, you use the same set of training-testing partitions, in order to see if accuracy is improving against a constant benchmark. How would this compare to using the holdout set described above? Does your answer depend on your choice of *k*? Briefly explain.

(c) You decide to compare your three favorite algorithms on this dataset. You plot Receiver Operating Characteristic (ROC) curves for each of them. Does your ROC analysis indicate which algorithm is best? Briefly justify your answer.



B766 – COMPUTER VISION: BASIC QUESTIONS

B766-1. Feature Detection

(a) Describe the procedures for constructing the (i) Gaussian Pyramid and (ii) Laplacian Pyramid representations of an image.

(b) How is the Gaussian Pyramid used in SIFT feature detection?

(c) Why is the Laplacian pyramid a useful representation for image compression?

(d) Explain how a Gaussian Pyramid is used in binocular stereo matching. There are multiple answers to this question. You can just pick any one of them.

B766-2. Object Recognition using Shape Contexts

(a) Briefly state the key steps involved in solving a shape matching problem using the Shape Contexts (SC) algorithm. Be sure to include the descriptor construction and the matching problem.

(b) Some feature descriptors (e.g., SIFT) generate keypoints as a first step. (i) Comment on whether such a step is necessary (or redundant) for the SC algorithm. (ii) Explain intuitively what the feature/shape descriptor in the SC method encodes.

(c) Briefly explain whether or not this framework is robust to outliers.

(d) Suppose you want to use the SC algorithm to compute distances between shapes for use in recognizing (or classifying) handwritten digits. Imagine that the classifier requires these distances between shapes to be a **metric** (i.e., non-negative, symmetric, $d(x, y) = 0$ if and only if $x = y$, and triangle inequality). Briefly comment on whether or not Shape Context satisfies the metric property.

B769 -- ADVANCED NATURAL LANGUAGE PROCESSING: BASIC QUESTIONS

B769-1. Language Model Size Reduction

A language model is usually very large. To fit it into small devices, it might be necessary to reduce the size of an existing language model, without overly compromising its quality. The following questions consider the simple case of unigram language models.

(a) Let P be the set of probability distributions over a given vocabulary. Let $Q \subset P$ be a subset of P. For example, one such subset contains distributions where the probabilities on words "apple" and "orange" are constrained to be the same. Let $\theta \in P \backslash Q$, i.e., in P but not in Q. Let KL( ) be the Kullback-Leibler divergence. Which of the following two is a better approximation of θ? Justify your answer.

$$\underset{q \in Q}{\operatorname{argmin}} \, KL(\theta || q)$$

$$\underset{q \in Q}{\operatorname{argmin}} \, KL(q || \theta)$$

(b) You have a unigram language model over a 100,000-word vocabulary. You want to reduce its size so it only uses 10,000 numbers, yet is still defined over the same vocabulary. Briefly explain how you might perform such a reduction, and what you are trying to optimize. You may assume that data structures used in your small language model do not count against your space.

B769-2. Incremental EM Algorithm

You want to classify text documents into C categories using a Naïve-Bayes classifier.

(a) Given n labeled documents and m unlabeled documents, describe (in pseudo code) the Expectation-Maximization (EM) algorithm that uses these $n + m$ documents to train the Naïve-Bayes classifier.

(b) What objective function is being optimized in (a)?

(c) Now instead of having those $n + m$ documents all available to you at once, they arrive sequentially one at a time. In addition, because of memory limits you cannot store them all. In fact, you can only store a constant number of things while $n + m \to \infty$. How would you adapt your algorithm so that it learns incrementally? Give pseudo code.

B776 -- ADVANCED BIOINFORMATICS: BASIC QUESTIONS

B776-1. Hidden Markov Model Duration Modeling

Suppose we are designing a hidden Markov model for inferring the locations of a certain class of genomic elements (e.g., genes, regulatory motifs, etc.) within the genome. We assume that the elements can be of varying length and that the emission probabilities are the same at each position within an element.

(a) What is the simplest (in terms of the number of states) (sub)-model for this element that allows it to have any length ≥ 1? Draw the topology for this sub-model and give the length distribution it models.

(b) Suppose we wish to have more control over the length distribution of this element. Draw the topology for a sub-model that can model any length distribution for elements of length 1 to 5 (and disallows elements of length ≥ 5).

(c) What is a disadvantage of using the model in (b), other than the fact that it cannot model elements of length ≥ 5?

(d) Suppose we obtain a few positive examples of these elements through experimental validation. These examples give us enough data to estimate the frequencies of the bases in these elements, but not enough to estimate the distribution of lengths of the elements. Given a genome sequence and these positive examples, describe the methodology you would use to choose between the sub-models of parts (a) and (b).

B776-2. Motif Finding with Hidden Markov Models

Consider using a hidden Markov model (HMM) to discover and characterize a motif in a given set of DNA sequences.

(a) Describe an HMM for this task that satisfies the following requirements:

- The motif is three bases wide.
- The sequences may be of arbitrary length.
- Each sequence has exactly one occurrence of the motif in it.
- The motif is equally likely to occur in any position in a given sequence.

Show all states and transitions in the HMM. Describe any parameters that would be shared (tied) in the model.

(b) Briefly describe how this HMM could be used to discover a motif in the sequences.

(c) Briefly describe how this HMM could be used to determine the most likely location of the motif in each sequence.

Answer both of the questions in the Section A760 if you are a machine-learning student or both in Section A766 if you are a computer-vision student.

A760 – MACHINE LEARNING: ADVANCED QUESTIONS

A760-1. Co-Training

Co-training is a semi-supervised learning technique based on two "views" of each instance.

(a) Briefly describe the co-training algorithm. Be sure to describe the inputs to the method.

(b) Discuss the key assumptions that underlie co-training.

(c) Now consider applying co-training to a dataset for which you do not know the two views a priori. Assume that each instance is represented as a fixed-length feature vector, and that there is a moderate sized set of labeled instances (say 100) available for training. Describe an approach that partitions the features into two views that are suitable for co-training, assuming such a pair exists.

(d) Briefly discuss the most significant limitation of your approach.

A760-2. Using Domain Expertise

Consider using machine learning to create a decision aid for a task where it is not possible to write by hand a correct and complete program for making these decisions. Specifically, consider recommending financial investments given one's age, income, education, work experience, health, family, etc. Assume you have a collection of prior cases (i.e., examples), marked as to whether or not in hindsight the chosen investment was a good idea. Imagine you are collaborating with three people who are experts in this task (assume all three experts agree with the labels on your training set).

(a) Assume you choose to employ a *decision tree* learning algorithm to produce a good method for recommending investments. What would be the best way to utilize the three experts? Justify your answer.

(b) Now consider using *Markov Logic Networks* (*MLNs*). What would be the best way to utilize these domain experts in this case (do not reuse your answer to part (a))? Again, justify your answer.

(c) If some critic said you should use an Inductive Logic Programming (ILP) method instead of MLNs for this task because trained MLN models are harder to understand than sets of Horn clauses, what would be your most powerful counter argument?

(d) Imagine someone asked you: "How many examples was each of your three experts worth?" Explain one good experimental methodology that you could use to empirically estimate an answer to this question. Explain how you interpreted this question and how your experiment will produce numeric answers to it.

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