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

Artificial Intelligence

Monday, September 20, 2010

GENERAL INSTRUCTIONS:

(a) This exam has **16** 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 <u>both</u> of the questions in the Section B760 if you are a machine-learning student or <u>both</u> 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 QUESITONS

B731-1. Bayesian Network Inference

Consider the standard Bayesian network inference problem: answering queries of the form Pr(X|**e**), where X is some variable in the Bayesian network and **e** is evidence that specifies the values of some subset of the remaining variables.

(a) Define *exact* and *approximate* inference in Bayesian networks.

(b) When would you use approximate inference instead of exact inference.

(c) Briefly describe the fundamental approach that Markov chain Monte Carlo (MCMC) takes to approximate inference.

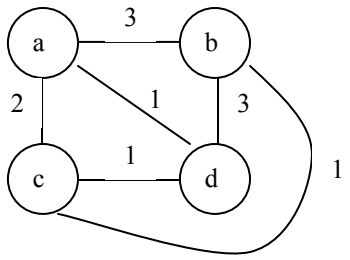
(d) Briefly describe the fundamental approach that loopy belief propagation takes to exact inference.

B731-2. First-Order Logic and Prolog

Consider the task of representing and reasoning about undirected graphs in definite clause logic, specifically Prolog.

(a) Represent the following graph *g* in Prolog.

g:



(b) Write a Prolog program to compute the maximum weight spanning tree of a graph represented in this fashion. The top-level predicate of your program should be a binary predicate *mwst(X,Y),* where on any call to this program the first predicate is bound to a constant naming the graph. Your program should bind *Y* to the maximum weight spanning tree, represented as a list of edges, where each edge is a list consisting of its two endpoints and its weight. For example, a call to the top-level predicate looks like:

mwst(g,Y)

and the answer to the query might look like:

mwst(g,[[a,b,3], [a,c,2], [b,d,3]])

B760 – MACHINE LEARNING: BASIC QUESTIONS

B760-1 Computational Learning Theory

This question compares the Mistake Bound and Probably Approximately Correct (PAC) models of learning. We will assume for this problem that the class of target concepts C

is contained in our hypothesis space H (i.e., $C \subseteq H$).

(a) Does learnability of a concept class C under the PAC-learning model imply learnability of C under the Mistake Bound model? If not, give a counterexample (a concept class that is PAC-learnable but not Mistake Bound learnable).

(b) What does *agnostic* learning mean? The Find-S algorithm for a concept class C always returns the most specific hypothesis in C that is consistent with the data. Is this algorithm agnostic?

Recall, the PAC learning theorem states that with probability 1 - δ, a hypothesis from set H found to be consistent with N training examples will have true error rate at most ε whenever:

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \log\frac{1}{\delta}\right).$$

Now, let's show that an algorithm that learns concept class C in the mistake bound model also learns it in the PAC-learning model. To do this, we'll use a variant of the Find-S algorithm: namely, **it only modifies the current hypothesis when it makes a mistake**; otherwise, it doesn't change anything. We'll call this algorithm **Find-S'**.

Suppose we run **Find-S'** until we locate a hypothesis h that is not eliminated after $\frac{1}{\varepsilon}\ln(\frac{M}{\delta})$ examples, where M is our given mistake bound.

(c) What is the probability this algorithm accepts hypothesis h with an error greater than ε? Show this implies that hypothesis h is PAC learnable. (Hint: you'll need to make use of the familiar inequality: $(1-x) \leq e^{-x}$ for suitable x.)

B760-2. Reinforcement Learning

In some reinforcement-learning (RL) tasks, one is given a 'world model' of how actions impact the environment. For instance, for the game of chess such a model would say how a legal move changes the current board configuration into the next board configuration.

(a) Explain one important way that having an accurate world model can change how an RL task is computationally addressed (compared to performing RL without such a model). Clarify your answer. Be sure to justify why you think this change will improve learning.

(b) Describe an experimental methodology that will help answer the question of whether or not a given world model (which may be imperfect) improves learning on some given testbed.

(c) Assume that in some RL task you are given *N* Boolean-valued sensors and *M* possible actions, but no world model. You are assigned the task of using some machine-learning method to learn a non-linear world model. Which learning method would you use? Justify your choice. Describe how you would collect training examples for your chosen world-model learning algorithm.

(d) As a reinforcement learner improves its performance in a given testbed over time, it is likely to visit 'bad' world states and perform 'bad' actions less and less frequently. What impact, in any, does this phenomenon have on the choice of algorithms for learning world models?

B766 – COMPUTER VISION: BASIC QUESTIONS

B766-1. Edge and Feature Detection

You want to identify edges in noisy images acquired from a security camera. Let the given image be denoted as *I*.

(a) (**Commutativity**). Assume G stands for a Gaussian filter kernel, and L stands for a Laplacian filter kernel. Consider the following two filtering operations on the image, *I*, where • stands for convolution.
(1) Gaussian of Laplacian (GOL): G • (L • *I*)
(2) Laplacian of Gaussian (LOG): L • (G • *I*)
Briefly discuss whether these two operations are equivalent or different.

(b) (**Associativity**). Does (1) *first* convolving the image with a Gaussian and *then* applying the Laplacian on the resultant image, i.e.,

L • (G • *I*)

the same as (2) *first* computing the Laplacian on the Gaussian and *then* applying it to *I*, i.e.,

(L • G) • *I*?

Explain which alternative is often used in practice in terms of quality of solution and efficiency of code.

(c) Suppose you had access to a set of pre-computed Gaussian masks with different (but known) standard deviations. Can you suggest a way you could use this input to approximate the LOG filter from (a) above? Give <u>one</u> potential advantage of this strategy.

B766-2. Face Detection

You have been asked to build a surveillance system that can detect people's faces in real-time as they walk past a video camera located at a building's entrance.

(a) Assuming you use the Viola-Jones detector for face detection, explain how and why "integral images" are used for computing image features.

(b) Is each feature in the Viola-Jones method invariant to (i) scale, (ii) rotation, and (iii) illumination? Briefly explain why or why not for each invariant. For a face detection system using the Viola-Jones method, how could the system be built so as to be reasonably robust to changes in scale, rotation, and illumination?

(c) After using the AdaBoost algorithm to select a set of weak classifiers, how are they used to achieve real-time face detection performance in the Viola-Jones detector?

B769 -- ADVANCED NATURAL LANGUAGE PROCESSING: BASIC QUESTIONS

B769-1. Information Theory

Consider document categorization using a feature vector representation of the document. The label can be viewed as a random variable Y. Each feature can be viewed as a random variable X too.

(a) Define the entropy $H(Y)$, and the conditional entropy $H(Y|X)$. Be sure to clearly state any assumptions you made on the values these random variables take.

(b) Prove the information inequality $H(Y) \geq H(Y|X)$. You may use the properties of mutual information.

(c) The inequality you just proved states that no matter what the feature is, knowing it will not increase the uncertainty in the label. Therefore, there is no harm in including an arbitrary feature. On the other hand, the practice of *feature selection* excludes certain features and often achieves improve label prediction. Explain this paradox. Your answer should not be about computational speed.

B769-2. Language Modeling

Your new boss gave you a training corpus and a test document in language X. This language has words separated by white spaces. The training corpus is written on a single, very long line. The test document is written on a second very long line. You built a unigram language model (LM) and a bigram LM on the training corpus, and computed the *perplexity* on the test document. (Note: please read all parts of the question before beginning to write your answer.)

(a) Define bigram LM.

(b) Define perplexity.

(c) Thinking your job is done, you discarded the training corpus and the test document (but kept the vocabulary and the LMs). Just before submitting your LMs, you realized that language *X* is written from right to left, while you had assumed left-to-right. Assume the test document is sufficiently long. How far off is your perplexity on the test document? Explain.

(d) Can you correct your bigram LM without asking your boss for the training corpus again? Explain.

B776 -- ADVANCED BIOINFORMATICS: BASIC QUESTIONS

B776-1. Clustering with Stochastic Context Free Grammars

Suppose we are given a stochastic context free grammar (SCFG) that characterizes various types of RNA secondary structures, and a set of RNA sequences that can be parsed using this grammar. Assume that both the productions and the probability parameters for the grammar are given.

(a) Describe how you would use this grammar to cluster the sequences such that sequences containing similar secondary structures fall into the same cluster.

(b) Discuss one significant limitation of your approach.

B776-2. Higher-order and Inhomogeneous Markov Chains

Consider first-order homogenous Markov chains and their more complex counterparts.

(a) In comparison to a first-order homogeneous Markov chain, describe one similarity between an inhomogeneous Markov chain and a kth-order ($k > 1$) homogenous Markov chain.

(b) Suppose we are modeling sequences over an alphabet of size n. How many free parameters are required for

i. a kth-order homogenous Markov chain
ii. a kth-order inhomogeneous Markov chain where the transition probabilities at time t depend on (t mod w)?

(c) A biologist approaches you with an interesting set of DNA sequences and wants to model these sequences with a Markov chain. The biologist believes that a first-order homogeneous Markov chain will not be sufficient, but is having difficulty expressing what type of more complex Markov model will be necessary. Describe the computational tests you would perform on the biologist's set of sequences in order to determine whether

i. a higher-order model is appropriate
ii. an inhomogeneous model is appropriate.

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

A760 – MACHINE LEARNING: ADVANCED QUESTIONS

A760-1. Classification with Complex Class Structure

Consider a task in which we want to learn models that classify companies using text from their web sites. We want to classify each company according to both its age (e.g. Microsoft is 35 years old) and its economic sector (e.g. Microsoft is in the *software* sector). Assume that we are using a discretized representation of company ages; e.g. companies fall into one of the following classes: 0-5 years, 5-20 years, 20-80 years, > 80 years.

Describe a learning approach for this task that takes into account the following aspects of the problem. Provide specifics about how you plan to handle each of these issues.

(a) Although the age variable is discretized, some classes are more similar than others (e.g. 0-5 years is more similar to 5-20 years than to 20-80 years). When the learned model makes classification mistakes, we would prefer mistakes that confuse similar classes over those that confuse distant classes.

(b) The classes of the sector variable are organized into a hierarchy. For example, the *software* sector is part of the more general *computers* sector, which is part of the *technology* sector.

(c) The age and the sector variables are correlated. For example, most software companies are relatively young. We would like to exploit this information to make more accurate classifications for both classes.

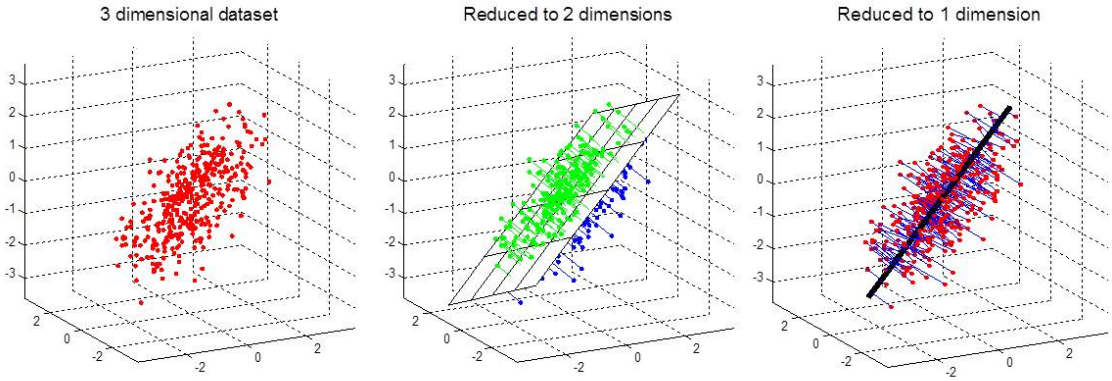
(d) At test time, we may encounter companies that do not belong to any of the defined sector classes. We would like a classification procedure that is able to predict when a given instance represents a new class.

A760-2 Principal Components Analysis

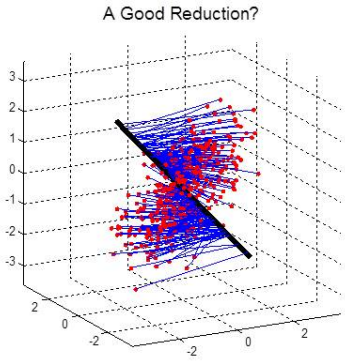
This question examines some basic ideas behind Principal Components Analysis (PCA), which is among the most popular algorithms for reducing the dimensionality of a dataset.

(a) Briefly explain the idea behind dimensional reduction. Specifically, what can make it possible to eliminate dimensions in a given data set? Is it always reasonable to do this?

PCA works by preserving orthogonal (i.e., perpendicular) dimensions with the highest variance, in a greedy fashion. For example, consider the following three-dimensional data set, shown below on the left. In the middle, we reduce it to a two dimensional plane, showing how the original points are projected onto it. (Points above the plane are green and points below it are blue.) On the right, we reduce it to one dimension, namely a single line, again illustrating the projections of the original data.



(b) Given the general description of PCA above, do these reductions make sense? Briefly explain why the line shown below would not be a good one dimensional representation of the data.



(c) Granted we have methods such as support vector machines that work well in practice in high dimensional spaces, give three circumstances where you might want to use PCA (or dimensional reduction in general). Justify your answers.

A766 – COMPUTER VISION: ADVANCED QUESTIONS

A766-1. Mean Shift

This question is based on the Mean Shift algorithm.

(a) Explain (in plain language or more formally) what is the objective function of the Mean Shift method. Also explain what are the key steps of each iteration.

(b) Describe one significant difference between the Mean Shift algorithm and standard *k*-means clustering. Ignore that Mean Shift is more computationally expensive in practice.

(c) With Mean Shift, you may still be required to calibrate the bandwidth parameter(s) in practical applications. Comment on adverse effects of high and low values of this parameter on the final clustering/segmentation solution. How could you adjust these value(s) if you wanted a segmentation with a **large** number of components?

(d) Will a sequence of successive iterations in Mean Shift eventually reach a fixed (i.e., stationary) point? Will this be the global optimum?

A766-2. Structure from Motion

Consider the task of inferring Structure from Motion.

(a) The Tomasi-Kanade factorization algorithm assumes that
$x=\mathbf{u}^T\mathbf{S}$,
$y=\mathbf{v}^T\mathbf{S}$
where x and y are the feature coordinates on the image and S is the 3D position of the feature. Explain the simplification made by this assumption. Write the matrix form of this assumption for multiple images and multiple feature points.

(b) If you take a sequence (say 40) of photos of our CS building while walking around the building in a full circle, using a camera in wide angle mode (short focal length), highlight two challenging issues when using the Tomasi-Kanade method to reconstruct the building, and give technical reasons why these two issues are challenging.

(c) For each of the issues you mentioned in (b), discuss one possible idea how the issue may be (partially) addressed.

(d) In photometric stereo, a pixel's intensity is $I = \mathbf{L}^T\mathbf{N}$, where **L** is the light and **N** is the surface normal. How can we apply the factorization idea to recover 3D object shape?

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