Spring 2011
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, January 31, 2011

GENERAL INSTRUCTIONS:

(a) This exam has **13** 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 exactly 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.

731 Advanced Artificial Intelligence: Basic Questions

B731-1. Expectation Maximization

Let's consider a Bayesian Network A→B→C. Suppose we have the following data table, with entry ***x*** missing (at random):

A	B	C
F	T	T
T	***x***	F
T	F	F
T	T	T

(a) Use these data to estimate initial parameters for this network. Use maximum likelihood estimation for simplicity.

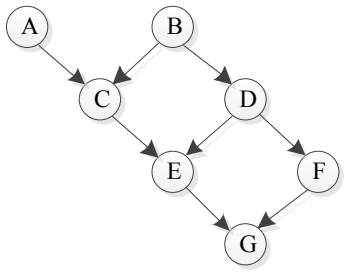
(b) Apply the Expectation Maximization (EM) algorithm by hand to estimate the values of the missing data, reestimate the parameters, etc., until convergence. Show all calculations, clearly identifying the E and the M steps.

(c) How many iterations does EM take to converge in step (b)? Will this always be the case? Explain.

(d) Name and justify an alternate technique for finding missing values in a Bayesian Network. When would it be appropriate to use this alternative technique instead of EM?

B731-2. Variable Elimination

Consider the following Bayesian network:



(a) Consider the query P(G) and the elimination ordering *A, B, C, D, E, F*, where we are eliminating *A* first, not last. Show what the variable elimination algorithm will do on this graph by creating and filling in the table like the following in your solution book. Note, your generated potentials (also known as "factors") should be written in the form $f_i(X_1,\ldots,X_k)$ for some set of variables $X_1,\ldots,X_k$. The first row has been filled out for you.

Variable eliminated	Potentials used	Potential generated
A	$f_1(A)\, f_2(A,B,C)$	$f_3(A,B,C)$

(b) Consider that the complexity of the inference process is highly sensitive to the order of elimination. Sketch a simple Bayes network over n (**>** 5) random variables, where there is one elimination ordering for some query P(X) that achieves performance that is linear in n and another elimination ordering for the same query whose performance is exponential in n. Demonstrate the linear case in a table and convincingly (but briefly) argue why the other example is exponential.

(c) Do we always know the optimum way to eliminate variables for a given query? Does it depend on the query, assuming we are considering a single Bayes net? Briefly justify your answers.

760 Machine Learning: Basic Questions

B760-1. Naïve Bayes

Consider using naïve Bayes to learn models for a two-class, text classification problem. Suppose we are interested in comparing naïve Bayes when used with two different representations for documents. In the *set-of-words* representation, a document is represented by a set of random variables, one per word in the vocabulary, each of which indicates the presence or absence of the corresponding word. In the *bag-of-words* representation, a document is represented by a set of random variables, one per position in the document, each of which indicates the word occurring at that position.

(a) Using an equation, define the general naïve Bayes classification procedure.

(b) What is the key assumption made by naïve Bayes?

(c) Referring to your equation in (a), describe how the set-of-words and bag-of-words variants of naïve Bayes differ. Use a specific example, such as the document "the cat in the hat," to describe how the calculations differ for the two representations.

(d) Discuss how and why you would use a regularization approach in estimating the parameters of your models.

B760-2. ROC and Precision-Recall Curves

Suppose we use a decision-tree induction algorithm and bagging to learn a classifier that is an ensemble of 20 decision trees, for a two-class problem.

(a) Describe a procedure for computing a Precision-Recall (PR) curve for this classifier, given a held-aside set of labeled examples.

(b) Describe a procedure for computing an ROC curve for this classifier, given a held-aside set of labeled examples.

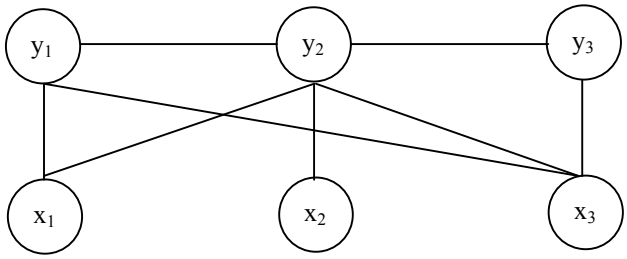
(c) For both types of curves, describe how you would informally determine if the decision-tree ensemble has any predictive value?

(d) Suppose we modify our task to be a three-class problem. Could we still use PR and ROC curves? If so, how could we compute them?

769 Advanced Natural Language Processing: Basic Questions

B769-1. Conditional Random Fields

Let x_1, x_2, x_3 be observed words and y_1, y_2, y_3 be their unobserved Part of Speech labels. The following undirected graphical model defines a Conditional Random Field (CRF) on the y's.



(a) Write down the formula of this CRF, using feature functions that match the edges. You do not need to define these feature functions.

(b) Note the cross edges between x and y (such as x_1-y_2). Is this a linear chain CRF? Briefly justify your answer.

(c) From this CRF, is it possible to compute $P(y_1, y_2 \mid x_1, x_2, x_3)$? If yes, briefly explain how; if not, briefly explain why.

(d) From this CRF, is it possible to compute $P(y_1, y_2, y_3 \mid x_1, x_2)$? If yes, briefly explain how; if not, briefly explain why.

(e) From this CRF, is it possible to compute $P(y_1, y_2 \mid x_1, x_2)$? If yes, briefly explain how; if not, briefly explain why.

B769-2. Latent Dirichlet Allocation

Consider Latent Dirichlet Allocation with k topics. Once the k multinomial distributions ϕ_1, …, ϕ_k (each is over the vocabulary) have been sampled from their Dirichlet prior and fixed, one generates documents with the following (incomplete) pseudo code:

Line 1: For each document

    Line 2: Sample a k-dimensional multinomial θ from _____

    Line 3: For each word position in the document

        Line 4: Sample _____

        Line 5: Sample _____

Questions:

(a) Copy the above pseudo code to your answer book and fill in the blanks.

(b) Let us change the algorithm as follows: move Line 2 above Line 1, and move Line 4 above Line 3. What does the resulting algorithm do? Be sure to contrast it with Latent Dirichlet Allocation.

Advanced Bioinformatics: Basic Questions

B776-1. Viterbi Training

In genomics applications, the *Viterbi training* algorithm is often used as an alternative to the Baum-Welch algorithm for training hidden Markov models (HMMs). Given a set of sequences $\{s_1,\ldots,s_n\}$ and initial parameter values, $\theta^{(1)}$, for an HMM, the Viterbi training algorithm involves alternating the following two steps until convergence:

> Step 1: Run the Viterbi algorithm on each sequence, s_i, to determine a most likely path, $\pi_i^{(t)}$, given the current parameters, $\theta^{(t)}$.
>
> Step 2: Update the parameters to maximize the joint probability of the sequences and their most likely paths: $\theta^{(t+1)} = \arg\max_{\theta} \prod_i P(s_i,\pi_i^{(t)} \mid \theta)$

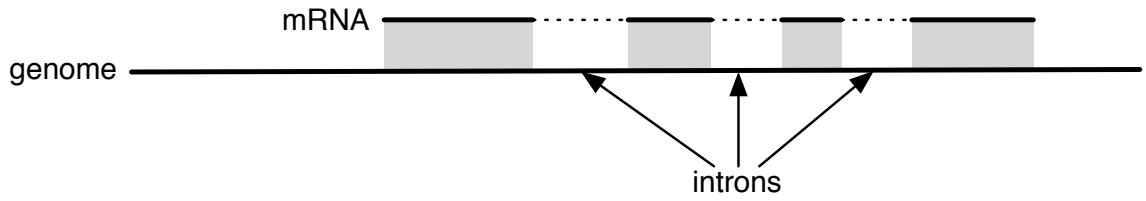
(a) Describe one potential *advantage* of Viterbi training over Baum-Welch training.

(b) Describe one potential *disadvantage* of Viterbi training relative to Baum-Welch training.

(c) The Viterbi training and Baum-Welch algorithms are related to two well-known techniques for the task of clustering. Briefly describe the correspondence between these HMM algorithms and their clustering counterparts.

B776-2. Gene Annotation with Pair Hidden Markov Models

One approach to finding the locations of genes within a genome is to obtain the sequences of messenger RNAs (mRNAs) and align each mRNA sequence against the full genome sequence. For eukaryotic species (e.g., human), this task is complicated by the fact that a gene sequence in the genome may contain introns, which are not present in the mRNA sequence for that gene due to the splicing process. Thus, an example mRNA alignment might look like:



(a) Draw the state transition diagram for a pair hidden Markov model that would be appropriate for aligning a eukaryotic mRNA sequence against a genome sequence.

(b) Describe how you would use your model to predict an alignment for each mRNA.

(c) Even with error-free sequences, there may be some uncertainty in the exact locations of the intron boundaries. Describe how you could use your model to give levels of confidence for each position in a predicted alignment.

766 Computer Vision: Basic Questions

B766-1. Image Segmentation

(a) The Normalized-Cut image segmentation algorithm computes the eigenvalues and eigenvectors of a normalized affinity matrix, $A(i, j)$, between two pixels, i and j. Describe the meaning of the *two* eigenvectors associated with the two smallest eigenvalues of the normalized matrix.

(b) Define an affinity matrix, $A(i, j)$, that would be appropriate for segmenting regions based on their texture differences. Briefly explain how your definition measures texture similarity or difference.

(c) The original Normalized-Cut method assumes segmentation into exactly two regions. Describe how to generalize the method so that it can segment an image into more than two regions. Include a stopping criterion for deciding how many regions are found.

(d) An alternative segmentation method is the k-means algorithm. Describe two major differences between the k-means and Normalized-Cut methods for image segmentation.

B766-2. Fundamental Matrix

Assume known coordinates for point sets in two views P = $\{p_1, \ldots, p_n\}$ and Q = $\{q_1, \ldots, q_n\}$ such that p_i *matches* q_i (i.e., they are corresponding points). We often calculate the *fundamental* matrix, **F**, which encodes information on the epipolar geometry of the imaging system using the eight-point algorithm (by Longuet-Higgins).

(a) First, briefly describe why eight point matches are needed in this method. That is, $|P|=|Q| \geq 8$.

(b) Now, assume that you are provided 20 matches in all, and you have successfully written out your objective in the form of a set of linear equations, $A\mathbf{f} = 0$, where **f** is a vector consisting of entries of the fundamental matrix. What can you say about the rank of A?

(c) Given that **f** can only be recovered up to an unknown scale, give a constraint that you can impose on **f** so that you do not get the trivial solution?

(d) Succinctly describe how you can calculate **f** using Singular Value Decomposition (SVD).

760 Machine Learning: Advanced Questions

A760-1. Variations on Supervised Learning

This question assumes no medical knowledge beyond your everyday experiences with doctors, hospitals, etc.

Imagine you are given access to an electronic version of 100 years worth of the medical records for a community of 100,000 people, one where 100% of them live in that community for their entire life. For simplicity, assume that each year 1% of the people die and an equal number are born. The records are complete in the sense that all doctor and hospital visits are in the data set you have (but do *not* assume that during each such visit all possible medical tests are performed, nor that all possible diagnoses are recorded).

Assume you wish to learn to predict whether or not a given patient *P* will get disease *D* within the next 10 years.

(a) Briefly describe how you would address this task via standard *supervised learning*; be sure to include how you would tune parameters and estimate future performance. It is fine (encouraged, even) to choose one established machine-learning algorithm and reference it in your answer. (In this part of this question discuss at most one specific learning algorithm.)

(b) For each of the following technical issues, discuss how it might arise in this setting and how you would address it. It is fine if a different machine-learning algorithm is used in each of your three answers, but for each answer focus on at most one specific algorithm.

(1) Medical treatments and laboratory tests change over time, so medical records from 100 years ago are not the same as from 10 years ago.

(2) The amount of information is not the same for each patient. For example, some people have undergone medical test *X* many times, while others never have had this test.

(3) There might be formalized domain knowledge about disease *D.* You may assume "formalized" means "written in some AI knowledge representation" and you are free to choose the particular formalism in which it is written.

A760-2. Kernel Methods and Model Comprehensibility

The Next Big Corporation has hired you to build a predictive model from some of its data; the data is in feature-vector format, with several thousand records and several thousand features, including a binary class variable they would like to predict. You run all the leading machine learning algorithms and find that a support vector machine (SVM) with a Gaussian (or RBF) kernel is significantly more accurate than all the other algorithms. But the company wants to understand the SVM-produced model in order to use it to inform future business decisions; they would not be satisfied with a black-box predictor.

(a) How do you explain to the company representatives how the learning algorithm and prediction algorithm work? Assume they are college-educated (business or liberal arts majors). Also, assume they only have time to read a couple of paragraphs and perhaps look at a picture or two if helpful.

(b) Describe <u>two</u> ways to obtain a comprehensible representation of the final predictive model (or approximation to it), which they can inspect and understand.

(c) Discuss <u>one</u> advantage of each of your answers in (b) relative to the other.