SPRING 2009
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, February 2, 2009
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 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).

B760 – MACHINE LEARNING: BASIC QUESTIONS

B760-1. Supervised Learning

You have a supervised learning data set with missing feature values. Although each example is missing only a few values, every example and every feature has some missing values, so you cannot just delete problematic examples or features.

(a) Choose two different supervised learning algorithms, and for each algorithm present one (different) way to modify the algorithm to handle missing values. For this part, you may assume values are "missing at random," that is, "missing-ness" is not correlated with class or any feature values.

(b) Now assume values are not missing at random – specifically, negative examples tend to have more missing values than positive examples. Are your modifications in part (a) still appropriate? Why or why not?

B760-2. Reinforcement Learning

(a) Define the reinforcement learning (RL) task and discuss one key difference between it and supervised machine learning.

(b) Describe how the Q-learning algorithm converts the RL task into a supervised learning one.

(c) Imagine you are using RL to learn how to play a one-player board game in which there is no random aspect such as tossing dice or drawing cards. The game has five different board states (S_0 through S_4) and there are three possible actions (*A*, *B*, and *C*). Assume the only rewards are 1 for a win, -1 for a loss, and 0 for a tie game

Assume you initialize all the Q values to zero. Imagine that your RL agent wins the first game after only three moves, going from board state S_0 to S_1 via $Action_A$, then to state S_4 via $Action_C$, and finally to S_3 via $Action_A$.

Show the changes to the Q function after this game. Be sure to state any additional assumptions, including parameter settings, that you need to make.

B766 – COMPUTER VISION: BASIC QUESTIONS

B766-1. Feature Point Detection in Images

One important operation in computer vision is to extract distinctive feature points (key points) from an image.

(a) Consider a small image patch *A*. What is the criterion used by the Moravec algorithm to decide whether it represents a feature point? In other words, how should it be compared to its neighboring patches?

(b) The Harris corner detector is an improvement upon the Moravec algorithm. Explain the criterion for detecting corners.

(c) One critical parameter in feature detection is the patch size. Discuss the effect of this parameter; that is, describe <u>one</u> important issue if the patch is too small and <u>one</u> important issue if the patch is too large.

B766-2. Face Recognition using Eigenfaces

The Eigenfaces algorithm projects an input image to a point in a *k*-dimensional "face space."

(a) What does each dimension in face space correspond to?

(b) If there are *m* training images for each of *n* different faces to be recognized, how can all of these *nm* images be used to define a classifier for recognition using Eigenfaces?

(c) Is the Eigenfaces algorithm invariant under illumination direction change? Explain why or why not.

(d) What preprocessing could be done to an input image so as to make the algorithm invariant to illumination intensity change, where intensity change is defined as adding a constant value to each pixel?

(e) How could you modify the Eigenfaces algorithm so that it is used for face detection instead of face recognition?

B769 -- ADVANCED NATURAL LANGUAGE PROCESSING: BASIC QUESTIONS

B769-1. Zipf's and Mandelbrot's Laws

(a) What does it mean for a text corpus to exhibit Zipf's law? Be sure to give a concise mathematical definition, as well as explain it in English.

(b) What is the significance of Zipf's law for natural language processing?

(c) A simplified Mandelbrot's law is the following: $f = a\, r^{b}$
Discuss one way to estimate the parameters a and b.

B769-2. Factor Graphs

Consider a sentence with four word positions: *w1 w2 w3 w4*. Each word position can be one of two words: {A,B}. The probability of the sentence is determined by the following model:

$$P(w1\ w2\ w3\ w4) = (1/Z) \times \psi1(w1\ w2\ w4) \times \psi2(w2\ w3\ w4)$$

where Z is a normalization factor, and

$$\psi1(w1\ w2\ w4) = 1 / (\text{(the number of A's in } w1, w2, w4) + 1)$$

$\psi2(w2\ w3\ w4) = 1$ if there is exactly one A in *w2, w3, w4; 0.9* otherwise.

(a) Draw the factor graph that represents the distribution *P(w1 w2 w3 w4)*. Be sure to define all the nodes.

(b) Briefly describe the sum-product (i.e. Belief Propagation) algorithm in general. Clearly state what the algorithm computes.

(c) Is there a unique most likely sentence? If yes, what is it and why? If no, briefly explain why.

(d) Is there a unique least likely sentence? If yes, what is it and why? If no, briefly explain why.

B776 -- ADVANCED BIOINFORMATICS: BASIC QUESTIONS

B776-1. Clustering Alignments

Suppose we have two sets of sequences, $G = \{g_1 , \ldots, g_n\}$ and $H = \{h_1 , \ldots, h_n\}$ where each g_i has a corresponding h_i. For example, G and H might represent the genes in two genomes, with each g_i, h_i pair representing orthologous genes.

Describe an approach for clustering *alignments* of the g_i, h_i pairs. Your method should return k clusters such that the g_i, h_i pairs in each cluster share a similar alignment. Note that the g_i elements in each cluster may not be very similar to each other (and likewise for the h_i elements), but the g_i, h_i alignments in each cluster should be similar.

B776-2. Distance-based phylogeny

One approach for estimating the phylogenetic tree relating a set of DNA sequences is to first compute distances between each pair of sequences and then run a hierarchical clustering method (e.g., UPGMA or neighbor joining) with these distances as input.

(a) Briefly explain how to compute a distance between each pair from a multiple alignment of the set of sequences.

(b) Suppose you are *not* given a multiple alignment of the sequences. Describe an *alignment-free* method for computing the distance between each pair of sequences that uses a *pair hidden Markov model* or a *profile hidden Markov model*. By *alignment-free*, we mean that your method may not simply compute a single pairwise alignment between each pair.

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. Supervised Learning with Large Correlated Feature Sets

You are given a dataset with a million Boolean features, a Boolean class, and one thousand examples of each class. You are told that there is a high level of class noise, so you can expect that a learned concept will be no more than 75% accurate. Moreover, you are told that the target concept can be represented with a model using only 100 of the features, each having a small correlation with class. Nevertheless, each of these 100 features is highly correlated with five to ten additional features outside this set of 100.

(a) Design and justify a learning algorithm for this task.

(b) Would your change your learning algorithm if you were given a million examples of each class? Explain your answer.

A760-2. Classification with Feature and Misclassification Costs

Consider a classification task in which determining the value of *each* feature incurs some cost and each type of misclassification incurs some cost, expressed in the same units. For example, in a medical diagnosis domain, a particular feature value might be acquired by running a lab test that costs $300, and we might decide that a false-positive diagnosis represents a cost of $500, whereas a false-negative diagnosis represents a cost of $2000.

You should assume that:

- there is a fixed number of features,
- the feature values are known for all training examples,
- the costs are all specified.

At the start of the classification process for a test example, however, you should assume that no feature values are yet known, but any of them can be determined by incurring the associated cost.

Describe the following:

(a) what objective function your learning algorithm is trying to optimize,

(b) a learning algorithm for the task (be sure to describe the representation used by your learning algorithm), and

(c) a classification procedure that uses your learned model.

A766 – COMPUTER VISION: ADVANCED QUESTIONS

A766-1. Structure from Motion

From a set of images (or a sequence of video frames) of a single rigid object, we can recover its 3D structure. Such a technique is called Structure from Motion (SFM).

(a) Explain what "motion" means in this context.

(b) A simple case of SFM is affine SFM, which assumes a linear camera model. Explain what the "linear camera" means? How is it different from a general perspective camera model? Explain why general perspective SFM is more difficult than affine SFM.

(c) What is the minimum number of images that is required by affine SFM (i.e., Tomasi-Kanade algorithm)? And what is the required minimum number of feature points per image? Explain your answers.

(d) Given a sequence of images taken on an empty street (e.g., pictures like Google Street View but without moving cars or people), assume each image has enough features and you know the camera internal parameters (focal length etc). Describe a procedure that uses perspective SFM techniques to recover the 3D structure of the static objects on the street (buildings, traffic signs, etc). You do not need to explain details in every step; however, your answer should reflect that you know how the pipeline works. The street view pictures are used only to give you an example of input images; your solution should be generally applicable to arbitrary rigid scenes.

(e) Following question (d), and assuming that there are people and cars moving around, how might you automatically remove these dynamic objects (cars, people) to concentrate on reconstructing the static scene?

(f) A766-2. Image Segmentation using Graph Cuts and Mean Shift

(a) Given a 100×100 color image, define the main steps of the Normalized-Cut algorithm for image segmentation, including the quantity that is minimized. Include a reasonable way to define the affinity measure based on perceived color and texture. How large is the affinity matrix?

(b) 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 k regions, where k is unknown but no larger than 10. Indicate <u>one</u> main potential problem with your method for successfully segmenting the image.

(c) An alternative segmentation method is the Mean-Shift algorithm. Describe the influence of the parameter, r, that defines the radius of the spatial search window, on the resulting segmentations. How might you use multiple r values in a scale-space sense to obtain a good segmentation?

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