FALL 2006
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 18, 2006
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 B731, B760, B766, and B776, where these two questions need *not* come from the same section, *and*
- both questions in the section labeled A7xx that corresponds to your 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 labeled B7xx that corresponds to your chosen focus area. Also answer any two additional questions in any of the other sections (these two questions need NOT occur in the same section).

B731 – ADVANCED AI BASIC QUESTIONS

B731-1. Bayesian Networks

You are building a Bayesian network system, and you decide to include two Markov Chain Monte-Carlo (MCMC) inference methods, one of which is Gibbs sampling. The other will be a different instance of Metropolis-Hastings. You also design your system to make a default selection between the two methods, based on analyzing a given Bayes net.

(a) Present the general Metropolis-Hastings algorithm.

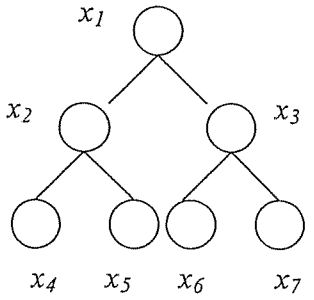
(b) Describe one or more properties of a Bayes net that should lead your system to **not** employ Gibbs sampling, but to use the other MCMC procedure instead. Illustrate your answer with a simple Bayes net.

(c) Describe your alternative MCMC search; more specifically, describe the proposal distribution that it will use.

(d) Are there any types of Bayes nets for which your new MCMC procedure will converge slowly or not at all? If not, why not?

B731-2. Undirected Graphical Models

Consider the undirected graphical model below. There are seven nodes $x_1, \ldots, x_7$. Let $Y=(y_1,\ldots,y_7)$ be a labeling of the graph, where $y_i \in \{\text{RED, GREEN}\}$ for $i=1\ldots7$.



We define a potential function ψ as follows:

$$\psi(i)=\begin{cases} a, & \text{if } y_i = y_j = y_k, \text{where } x_j \text{ is } x_i\text{'s left child and } x_k \text{ is } x_i\text{'s right child} \\ 1, & \text{if } y_i, y_j, y_k \text{ have different labels, or if } x_i \text{ has no child.} \end{cases}$$

We define a probability distribution over labelings as:

$$P(Y)=\frac{1}{Z}\prod_{i=1}^{7}\psi(i)$$

(a) Z is the normalization factor. Write down Z in terms of ψ. Discuss why there is no need for a normalization factor if the graph is a directed graphical model.

(b) When $a>1$, which labeling Y has the largest probability? Why? (Hint: do not use brute force and enumerate all 2^7 possible labelings.)

(c) Fix the labels on all nodes except for node x_i, and try either y_i=RED or y_i=GREEN. Write down the conditional likelihood ratio $\dfrac{P(y_i=\text{RED}\mid Y_{\setminus i})}{P(y_i=\text{GREEN}\mid Y_{\setminus i})}$ in terms of ψ, where $Y_{\setminus i}$ are all labels except y_i. Cancel as many terms as possible. You may use j, k to denote i's left and right child, and m for i's parent.

B760 – MACHINE LEARNING BASIC QUESTIONS

B760-1. Support Vector Machines (SVMs)

Standard linear SVMs for binary-classification tasks solve the following quadratic optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i(w' x_i + b) \geq 1 - \xi_i \ , \xi_i \geq 0 \quad , \forall i$$

where $w'x$ stands for the inner-production between w and x, and the ξ_i are *slack variables*.

(a) Explain the role of slack variables in SVMs for classification tasks?

(b) Consider the following one-dimensional dataset:

i	x_i	y_i
1	-1	-1
2	1	-1
3	2	-1
4	3	+1
5	4	+1

Write down a value for w and a value for b, such that $\xi_1 = \ldots = \xi_6 = 0$, and all the constraints are satisfied. You do not need to actually solve the SVM. But you need to briefly justify your answer. You also need to draw a sketch of the dataset and your decision boundary.

(c) Some people use an SVM formulation that does not include the bias b.

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i w' x_i \geq 1 - \xi_i \ , \xi_i \geq 0 \quad , \forall i$$

What restriction does this change put on the decision boundary? For the dataset above, can you find a value for w, such that all the slack variables ξ_i are zero and all the constraints are satisfied? If yes, write down w. If no, explain why.

(d) Briefly describe the role of kernels in SVMs and present a sample non-linear kernel function.

B760-2. Reinforcement Learning

Consider the case of performing reinforcement learning when no model of the environment (nor accurate simulator) currently exists. One idea is to *learn* a model of the world (i.e., environment).

(a) Using the vocabulary of reinforcement learning, what needs to be learned in order to produce a model of the environment (i.e., a world model)?

(b) Describe how you would collect training examples to learn the components of world models you described in Part (a).

(c) Describe one important advantage of learning world models.

(d) Describe one important disadvantage of learning world models.

(e) Discuss an experimental methodology to decide, in any given domain, whether or not to learn a world model.

B766 – COMPUTER VISION BASIC QUESTIONS

B766-1. RANSAC

(a) Given a set of measurements, some noisy and some irrelevant, and a hypothesized model, describe the main steps the RANSAC algorithm uses to estimate the parameters of the model. Include in your answer justification for how many measurement points should be used at each iteration of the algorithm.

(b) Say you want to detect roads in low-altitude aerial images by finding the edges of the road and their vanishing point. Describe how RANSAC could be used to detect the road from a set of line segment features detected in an input image.

B766-2. Epipolar Geometry

(a) Show with a figure the epipolar geometry between two images taken by two cameras. Include the epipoles, an epipolar plane, an epipolar line, the two cameras' optical centers, and a conjugate pair of points in the two images.

(b) If p is a point in one image and p' is its corresponding point in the other image, define the relation between p, p' and the fundamental matrix, $\mathbf{F}$, in the case where the two cameras are uncalibrated.

(c) What is the size of $\mathbf{F}$, the minimum number of point correspondences needed to compute $\mathbf{F}$, and what does a solution to $\mathbf{F}x = 0$ represent?

(d) Draw the epipolar lines and epipoles when the line through the two cameras' optical centers is parallel to both images' scanlines, the two optical axes form an angle of 90 degrees, and both cameras are equidistant from the intersection of the optical axes.

B776 – BIOINFORMATICS BASIC QUESTIONS

B776-1. Clustering

Suppose you wish to develop a variant of the *k*-means clustering algorithm that will operate on sequences rather than feature-vectors. While sequences may vary somewhat in length, assume your expectation is that length will not differ substantially from one cluster to another; you expect sequences in each of the k clusters to have an average length of, say, around 50 characters.

(a) Present your algorithm in commented pseudo code. Discuss how it differs from the standard *k*-means clustering algorithm.

(b) How would you change your algorithm into a *soft* version of *k*-means, that is, an algorithm that assigns each sequence a probability of being in each cluster?

B776-2. Time and Space Complexity of HMMs and SCFGs

Suppose we have a probabilistic grammar with N non-terminals, and we are given a sequence of length L to parse.

(a) What is the algorithm that we would use to find the most probable parse if the grammar can be represented as an HMM? What is the algorithm we would use if the grammar is an SCFG?

(b) What are the time complexity and space complexity of these algorithms?

(c) Suppose we have a grammar that is designed to model a relatively short sequence type, C, that is embedded in longer sequences. Further, suppose that C is best described using a context-free grammar but the rest of the sequence can be adequately characterized with a regular grammar. Describe how you would handle parsing more efficiently than the standard SCFG parsing method for this special case.

Answer both of the questions in the section labeled A7xx that corresponds to your chosen focus area.

A760 – MACHINE LEARNING ADVANCED QUESTIONS

A760-1. Active and Multiple-Instance Learning

Consider the situation in which a system is able to do active learning for a problem that is naturally framed as a multiple-instance task.

(a) Define the setting of multiple-instance learning.

(b) Consider the usual multiple-instance setting in which all bags in the training set are labeled. Suppose the learner can make queries that involve asking for the label of a particular instance in a positive bag. Describe a criterion you might use to select which instance in a positive bag would be labeled next for the learner.

(c) Would it make sense to query for the labels of instances in negative bags as well? Justify your answer.

(d) Consider the case in which some bags are unlabeled. Describe a criterion you might use to select which bag would be labeled next for the learner.

A760-2. Learning from Anonymized Data

Due to the wish of preserving privacy, the supervised-learning examples you are given have been anonymized into a file of what we will call *representative examples*. Assume that this has been done, to numeric data represented as fixed-length feature vectors, by

i. Clustering the original data into groups of size k (note that k is the size of each cluster, and is *not* the number of clusters), where all members of a cluster have the same output label.

ii. Creating one example to represent each cluster. For each feature, this representative example contains the minimum and maximum values of this feature for the k examples in this cluster (note that this means if the original data involves n numeric features, each representative example contains $2n$ numbers).

Notice that each representative example is a hyper-rectangle in feature space. You may assume that the clustering algorithm mentioned above strives to produce disjoint hyper-rectangles.

Present and motivate extensions, which *exploit* the structure of representative examples, to any two of the following supervised learning approaches. Aim to have two different extensions, rather than simply saying "One can also apply the same idea to Algorithm X."

- Decision-tree induction
- Support-vector machines
- Neural networks
- Genetic algorithms
- Nearest-neighbor methods
- Bayesian networks
- Any ensemble method (bagging, boosting, etc.)

You can assume that your extended algorithms are given the specific value for k used in a given anonymized dataset.

A766 – COMPUTER VISION ADVANCED QUESTIONS

A766-1. SIFT Features

(a) Name two kinds of image or scene changes that Lowe's SIFT features are generally invariant to, and one kind of image or scene change they are *not* invariant to. Briefly justify your answers.

(b) Describe a scene and associated image where object recognition using SIFT features will fail because of your answer in Part (a).

(c) Briefly describe how SIFT feature descriptors are constructed and represented given the output of the SIFT feature detector.

(d) Suppose you are given a range image in addition to a grayscale image. Define a modification of the SIFT algorithm that would use this additional data to compute a new invariant feature that is not possible to compute using standard SIFT features based on grayscale images.

A766-2. Object Recognition using Probabilistic Models

One approach to object recognition is to learn and use generative probabilistic models. One such method is Fergus et al.'s part-based approach in which each object category is represented by a collection of local part patches.

(a) Formulate the problem of deciding if a given image contains an object O or not given a set of observed parts using this Bayesian approach.

(b) Describe what is learned in this method and how learning is performed given a training set of images.

(c) How does this approach deal with the segmentation problem, i.e., determining which detected parts are part of the object and which are not (and therefore are part of the background)?

(d) Compare Fergus' method with the Eigenfaces method for object recognition in terms of any 3 of the following 4 issues:

i. number of parameters required
ii. number of training examples needed
iii. robustness with respect to variations in object shape and appearance
iv. invariance to translation

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