Fall 2013
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 16, 2013

GENERAL INSTRUCTIONS

(a) This exam has 11 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

You should answer:

(a) both questions in the section labeled 760 – MACHINE LEARNING

(b) two additional questions in another selected section, 7xx, where both questions *must* come from the same section.

Hence, you are to answer a total of four 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.

760 – MACHINE LEARNING: REQUIRED QUESTIONS

760-1 Empirical Evaluation Methodology

(a) Let tp, tn, fp, and fn denote counts of true positives, true negatives, false positives, and false negatives, respectively, on a test set of binary-labeled examples. Define each of the following terms. You may use terms you define earlier to help define later terms.

(i) Accuracy

(ii) Recall

(iii) Precision

(iv) True Positive Rate

(v) False Positive Rate

(b) How would you construct an ROC curve for a random forest learner? You may assume the ROC curve will be constructed for a single train-test split of the data set.

(c) Give one advantage of ROC curves over Precision-Recall curves and describe a task where, as a result of that advantage, you would prefer to use ROC curves.

(d) Give one advantage of Precision-Recall curves over ROC curves and describe a task where, as a result of that advantage, you would prefer to use Precision-Recall curves.

760-2 The *k* Nearest Neighbor (*k*-NN) and Weighted Majority Algorithms

Consider learning a model for a binary classification task with a large number of binary features, many of which are irrelevant. Suppose that the classes are not linearly separable. You are initially planning to apply two algorithms to the task: k-NN and Weighted Majority, where the prediction algorithms given to the latter are decision stumps for individual features.

(a) Describe one significant weakness of k-NN relative to Weighted Majority for this task.

(b) Discuss how you could adapt k-NN to mitigate this weakness.

(c) Describe one significant weakness of Weighted Majority relative to k-NN for this task.

(d) Discuss how you could adapt Weighted Majority to mitigate this weakness.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Dirichlet Process

(a) Recall $G \sim DP(\alpha, H)$ is a Dirichlet Process with concentration parameter α and base distribution H. Is a Dirichlet Process identical to a Chinese Restaurant Process? If so, explain how. If not, explain one major difference.

(b) Computational Ice Cream Academy (CIA) has thousands of ice cream flavors and is inventing hundreds more each year. CIA believes that each ice cream flavor x has a real-valued intrinsic tastiness score $f(x)$. To understand customers' tastes, CIA performs extensive user studies. The basic format is to present a human participant with two randomly selected ice cream flavors x and y, and ask the participant to select one of the two outcomes: x tastes better than y, or y better than x. This is known as a two-alternative forced-choice experiment. Note the user study is noisy by nature: regardless of the scores $f(x), f(y)$, one participant may think x is better than y while another participant may think otherwise. CIA hopes to estimate the function f from a large number of such experiments.

Formulate a statistical model that allows CIA to do so. Your answer should clearly explain the connection between f and the two-alternative forced-choice experiments. Make your modeling assumptions explicit. You should also briefly outline a parameter estimation procedure.

(c) In the previous question, there was no imposed structure on f. The director of CIA, however, believes that ice cream flavors form equivalence classes. Two flavors x, y are in the same equivalence class if $f(x) = f(y)$. It is not clear which ice cream flavors or how many of them belong to each equivalence class. It is also not clear how many equivalence classes there are. Of course, new equivalence classes may form as new ice cream flavors are invented.

Formulate a statistical model that allows the CIA director to impose such an equivalence class structure. Your answer should clearly explain how you handle a potentially unlimited number of equivalence classes. Make your modeling assumptions explicit.

761-2 Exponential Family

An exponential family distribution takes the form

$$p(x \mid \theta) = h(x) \exp\left(\theta^\top T(x) - A(\theta)\right), \tag{1}$$

where $T(x) \in \mathbb{R}^D$ is the D-dimensional sufficient statistics of x, $\theta \in \mathbb{R}^D$ is the natural parameter, $A(\theta)$ is the log partition function, and $h(x)$ modifies the base measure.

(a) For an *iid* set $\mathbb{D} = \{x_1, \ldots, x_n\}$, write down the likelihood function $p(\mathbb{D} \mid \theta)$ under the exponential family in a form similar to equation (1).

(b) The gradient of the log partition function $A(\theta)$ has a special property:

$$\nabla A(\theta) = \mathbb{E}_{x \sim p(x|\theta)}\left[T(x)\right].$$

Given a training set $\mathbb{D}$, write down the key equation that the maximum likelihood estimate (MLE) of θ must satisfy. Hint: your answer should include $\nabla A(\theta)$.

(c) For this question, for simplicity assume $h(x) = 1$. Further assume that given any training set $\mathbb{D}$ a machine learner will perform MLE as in the previous question. Now consider the *machine teaching* problem: there is a teacher who knows the target parameter θ^*. It also knows the machine learner's likelihood model in equation (1). The teacher wants to *construct a small training set* $\mathbb{D}$ such that, when the machine learner receives $\mathbb{D}$ the learner's MLE is approximately the target θ^*.

Propose a method for the teacher to construct such a $\mathbb{D}$. Explain your proposal. Hints:

- Your answer should be more sophisticated than "the teacher samples an *iid* training set from $p(x \mid \theta^*)$."
- $\mathbb{D}$ does not have to be *iid*.

766 – COMPUTER VISION QUESTIONS

766-1 Histogram of Oriented Gradients (HOG) Features

This question is based on the HOG detector proposed by Dalal and Triggs (CVPR 2005) and its various potential uses in computer vision problems.

(a) Assume you're designing an extension of HOG for the Kinect sensor. The key difference here is that rather than RGB images, you will instead operate on RGB-D images (where D corresponds to depth). Briefly describe the main modifications you must make to the procedure to leverage such additional information.

(b) Consider the task of designing a pedestrian detector for a video sequence acquired by a surveillance camera mounted on a *stationary* platform. The sequence will include frames acquired in bright sunlight as well as during the evening hours. Describe the main steps you will use to adapt your HOG descriptor for videos in the above application.

(c) Consider the task of using HOG for an athlete tracking application in sports videos. It is expected that your solution will work for various teams with distinctly different uniforms. You may choose to implement the histogram of either the signed or unsigned gradients. Which one will you prefer and why?

(d) Finally, imagine you want an extension of HOG to work for sequences acquired from moving cameras with dynamic background. Many available procedures only work when the camera and background are largely static. Describe *one* strategy you will adopt to make your detector more robust against the effects of camera motion in this application.

766-2 Mean Shift

Mean Shift is a popular method used in computer vision for problems such as image segmentation and object tracking.

(a) Define a feature vector representation for each pixel that is appropriate for use with Mean Shift in order to segment images into connected regions with roughly homogeneous perceptual color. Explain why you chose these features.

(b) Describe the main steps of how Mean Shift is used to perform image segmentation, including the parameter(s) that must be set in order to use Mean Shift.

(c) Comment on the sensitivity of the algorithm to its parameter value(s); i.e., what are the effects of setting the parameter(s) to values that are too big or too small.

(d) Instead of using Mean Shift to simultaneously extract all image regions, it is often valuable to create a hierarchical segmentation of regions in order to do a subsequent multi-scale analysis. Describe a variant of Mean Shift that could be used to create such a multi-scale hierarchy of image regions.

(e) An alternative image segmentation method is the Normalized-Cut algorithm by Shi and Malik.

  (i) Give one advantage of Mean Shift as a segmentation method compared to Normalized-Cut, and one advantage of using Normalized-Cut compared to Mean Shift.

  (ii) Instead of using the clustering process in Mean Shift segmentation for defining the final regions, how could the Normalized-Cut method be combined with Mean Shift to obtain better, more stable segmentations that are less sensitive to parameter values?

769 – ADVANCED NATURAL LANGUAGE PROCESSING QUESTIONS

This page intentionally left blank.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Network Clustering

A set of 10 proteins that are known to function together as a module are present in each of 100 different species. There is a known one-to-one correspondence between the proteins of each species. You are given the experimentally-determined protein-interaction network for this set of proteins in each of the species. Each network is represented as a set of undirected edges between pairs of the 10 proteins. You wish to cluster these networks to determine which species are most similar to each other in terms of the interactions between these proteins.

(a) Describe a K-means approach for generating a *flat* clustering of these networks.

(b) Describe a model-based approach (i.e., one in which you can compute the probability of the observed networks) for generating a *flat* clustering of these networks.

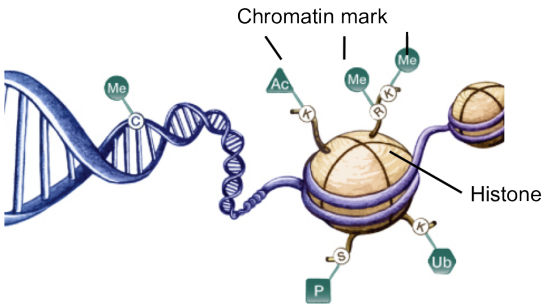
(c) Describe an approach for generating a *hierarchical* clustering of these networks.

(d) Of the approaches that you have described, which do you expect to generate the most useful clustering? Explain your reasoning.

776-2 Genome Segmentation

Recent advances have enabled us to generate genome-wide maps of chromatin state by measuring different chemical modifications called **chromatin marks** on histone proteins around which DNA is wrapped (Fig. 1A). There are multiple classes of chromatin marks, each of which can be measured separately. A genome-wide mark profile can be represented as a sequence of bits over the genome of an organism, each bit representing whether the mark is present or absent in non-overlapping bins of 100 base pairs (Fig. 1B). It turns out that different functional regions of the genome such as coding sequence and promoters have characteristic combinations of marks and such mark combinations can be informative for classifying regions of the genome in terms of their function.

A.



B.

	Bin 1	Bin 2	....	Bin m
Mark 1	0	1	....	1
Mark 2	1	1	....	1
⋮	⋮	⋮	....	⋮
Mark n	0	0		1

Figure 1: (A) DNA wrapped around histone protein complexes. A chromatin mark is a chemical modification on a histone protein. (B) A table representing n chromatin marks in m non-overlapping bins where each bin represents 100 base pairs of DNA, and 1 means the mark is present and 0 means the mark is absent.

(a) Given the genome-wide profiles of a small number of mark classes (e.g. n=5, in Fig. 1B), describe an unsupervised approach for using these data to segment the genome into m types of regions taking into account the sequential nature of genomic data. You can assume that the chromosomal locations of the bins are available to you and m is specified as input.

(b) Assuming that the organism for which you have this data has a well-annotated genome, and thus you know the locations of functional regions such as genes and promoters for this organism, how will you evaluate your results from part (a)?

(c) Now assume that these marks are measured in four species that are evolutionarily close. Describe how you would use these mark profiles from each species to identify regions in each species. A solution that independently applies your answer for part (a) to each genome separately is not sufficient.

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