SPRING 2014
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, February 3, 2014

GENERAL INSTRUCTIONS:

(a) This exam has 10 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:

- both questions in the section labeled 760 – MACHINE LEARNING

- 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 Linear separability, inductive bias and linear models

(a) Define the concept of *linear separability*.

(b) How does linear separability relate to the concept of *hypothesis space bias*?

(c) Briefly describe two machine-learning methods that learn linear decision boundaries. For a given training set, do these methods always learn the same model? If not, describe how they differ in their *preference biases*.

(d) Suppose we want to use gradient descent to train a neural network with one linear output unit (i.e.,. its output is the same as its net input) and no hidden units for a binary classification task. Suppose we want to use the following error function (which is known as cross-entropy error): $E(\boldsymbol{w}) = -\left[y \log o + (1-y)\log(1-o)\right]$, where $\boldsymbol{w}$ is the weight vector, y is the target value (0 or 1) and o is the output produced by our current network. What is the update rule we should use for adjusting our weights during learning with this error function?

Hint: the derivative of $\log(a)$ is $\dfrac{1}{a}$

(e) The neural network in (d) is capable of learning only linear functions. How could we extend it to learn nonlinear functions?

760-2 Learning with unbalanced data sets

Suppose you're developing a binary classifier for detecting a rare and dangerous disease. Let's say only one in a million people have it.

(a) Suppose you'd like to train a Support Vector Machine (SVM) with a Gaussian kernel for this purpose. What parameters would you have to specify? Briefly explain the meaning of each one.

(b) What do you expect the approximate accuracy of your classifier would be and why? (State any assumptions you make.)

(c) What other measure of quality might you use to evaluate your classifier? Briefly explain why you would do this. What factors could influence your decision?

(d) Do you need to introduce any new parameters for this evaluation? If not, briefly explain why this is. If you do, explain what they mean and how you would set their values.

(e) Does your answer to part (d) impact your selection for the value of the parameters from part (a)? Briefly explain your answer.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 1-Nearest-Neighbor (1NN)

Consider classification with 1-nearest-neighbor classifier (1NN). Let the labeled training data set contain n points, where the i^{th} point is (x_i, y_i). Each x_i is a d-dimensional real-valued feature vector, and y_i is a class label in one of C classes.

One major issue with 1NN is the choice of the distance function. The Euclidean distance may not be appropriate for a number of reasons. For instance, different features may be in different units whose scaling is arbitrary. One solution is to transform the features into a q-dimensional space by

$z_i = \mathrm{A}\, x_i$

where A is a $q{\times}d$ matrix, and then use the Euclidean distance between the z's.

(a) Write down the formula for the leave-one-out (LOO) error of 1NN under a given transformation A. Clearly define any functions and variables you introduce.

(b) Typically, A is not given and must itself be learned from the same data. One strategy is to learn A to minimize the LOO of 1NN. Write down an optimization problem to do so and explain why it is difficult to solve.

(c) Now consider a "stochastic 1NN" (s1NN): instead of finding the nearest neighbor of $\mathrm{x_i}$, it has some chance of selecting any other point x_j in the training set, but with probability related to some notion of distance between x_i and x_j. Design an s1NN algorithm, be sure to read part (d) below first, and clearly specify all relevant formulas.

(d) Propose an optimization problem to learn A using the LOO of s1NN. Your algorithm should at least partially address the difficulty in part (b). Be sure to sketch one way to solve your optimization problem, and explain why it is easier than in part (b).

761-2 Crowdsourcing

In crowdsourcing for real-valued estimates, a company recruits M human workers to evaluate N items. The i^{th} item (such as a stock) has an intrinsic value U_i (such as its fair price) that is fixed. For this question we assume U_i can take any real value, in particular, it can be negative, too.

Each worker j assigns a value V_{ij} to the items $i=1...N$. We assume that the value V_{ij} is a random variable with the following considerations:

- V_{ij} is based on the intrinsic value U_i;
- The j^{th} worker has a bias B_j which is the same on all items. In other words, if $B_j=10$ then the j^{th} worker on average prices any item 10 higher than the item's intrinsic value U;
- All workers have some zero-mean random fluctuation when they assign values.

(a) Propose a probabilistic model of V_{ij}. Clearly define all terms and variables.

(b) Given a collection of $\{V_{ij}\}$ for M workers on N items and nothing else (in particular, we do not know the B's), and assuming your model in part (a), is it possible to estimate $U_1, ..., U_N$? Be sure to explain your answer.

(c) Let us assume that the company already knew the intrinsic values $U_1, ..., U_k$ for the first $k<N$ items through other means. Given a collection of $\{V_{ij}\}$ for M workers on N items, propose a method to estimate $U_{(k+1)}, ..., U_N$. Be sure to clearly state your model assumption, present the algorithm (if any), and explain your solution.

766 – COMPUTER VISION QUESTIONS

766-1 Fundamental and essential matrices

(a) Let F be the fundamental matrix relating a pair of cameras, P and P'. If you want to calculate F, do you need information about the respective camera matrices for P and P'? Explain why or why not. If this information is not available, can you still solve for F? Explain why or why not.

(b) In the literature, it is common to think of the fundamental matrix F as a projective map. Comment on the invertibility of this mapping using any appropriate property of F.

(c) Consider decomposing the fundamental matrix into two parts, F_s and F_a , such that $\mathrm{F} = \mathrm{F}_s + \mathrm{F}_a$, where $\mathrm{F}_s = (\mathrm{F} + \mathrm{F}^T)/2$ and $\mathrm{F}_a = (\mathrm{F} - \mathrm{F}^T)/2$. Here, the geometric interpretation of F_s is that of a conic (i.e., the image of a curve passing through the two camera centers). Let e and e' correspond to epipoles in this setup. Prove that the epipoles lie on the conic. (Hints: For a skew-symmetric matrix, M, the quadratic form, $x^T\mathrm{M}x = 0$ for all x.)

(d) How is an essential matrix different from a fundamental matrix? Specifically, relative to a fundamental matrix calculation, do you need more or less information to solve for an essential matrix? Explain briefly.

766-2 Stereo matching

(a) The Lucas-Kanade algorithm is a widely used method for optical flow estimation. Describe the main steps if you were interested in adapting this method for stereo matching instead.

(b) The most common approaches for stereo matching rely on belief propagation and alpha expansion based ideas. Argue (using an example, if possible) why your Lucas-Kanade strategy may work better in some specific cases.

(c) Consider a situation where an application requires the use of piecewise affine warps and other arbitrary non-linear warps between images. Are there any bottlenecks (theoretical or practical) you can think of in terms of using some specific variants of the Lucas-Kanade algorithm to find the solution?

(d) Consider a situation where the two input images are perceptually similar but have significant illumination differences. Making direct use of least squares difference between intensities as an objective function is clearly problematic. Describe how you could extend or adapt the basic Lucas-Kanade algorithm to work in this situation. (Note: your procedure may not be elegant or fast but should have a reasonable chance of working.)

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Gene regulatory network learning with Bayesian networks

Consider the task of learning the structure of a gene regulatory network given gene expression data across many conditions. Suppose we wish to use Bayesian network models for this task.

(a) Describe a technique that generally prevents a directed acyclic graph with the maximum number of edges from scoring highest when searching over the space of possible networks.

(b) Briefly explain the advantages of using *module networks* over general Bayesian networks for this task.

(c) What key assumptions are made about gene regulatory networks when using the *module network* framework?

(d) Describe a technique for assessing the confidence in the presence of an edge in the regulatory network.

776-2 Gene function prediction

You are given gene expression measurements for *N* genes over *M* experimental conditions. Of these *N* genes, you know that *P* of these genes belong to a class representing biological function A and *Q* of these genes belong to a class representing a biological function B. Assume that every gene belongs to either A or B, but not both, and you know the identity of the *P* and *Q* genes. However *P*+*Q*<*N*, that is, there are *N*-*P*-*Q* genes for which you don't know whether they belong to class A or B.

(a) Describe a non-clustering based classification approach that makes use of the gene expression data to predict whether the *N*-*P*-*Q* genes with unknown function are members of class A or B.

(b) Describe a clustering approach that makes use of gene expression data to predict whether the *N*-*P*-*Q* genes with unknown function are members of class A or B.

(c) Give one strength and weakness of a classification versus clustering approach for the gene function prediction problem.

(d) Assume additionally that you are given a gene-gene interaction network. Describe a classification approach that uses the network and the gene expression data to infer the function of the *N*-*P*-*Q* genes.

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