Fall 2016
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 19, 2016

GENERAL INSTRUCTIONS

1. This exam has 10 numbered pages.
2. Answer each question in a separate book.
3. 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.
4. Return all answer books in the folder provided. Additional answer books are available if needed.

SPECIFIC INSTRUCTIONS

You should answer:

1. both questions in the section labeled 760 – MACHINE LEARNING
2. 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 Supervised Learning

Suppose you wish to perform supervised learning (assume a binary class variable). Your data set has an order of magnitude more features than examples; if it helps you answer the question, you may assume the data set has 1000 examples and 10,000 features.

(a) Describe a set of conditions under which you can learn an accurate classifier – one that will perform well on a new set of 1000 examples drawn randomly and independently from the same probability distribution as the training set and labeled according to the same target function.

(b) Name one supervised learning algorithm you believe will perform well on this task under your conditions from part (a), and explain why you believe it will perform well.

(c) Name one supervised learning algorithm you believe will not perform well under those same conditions, and explain why not.

(d) Describe an evaluation methodology you would employ to test your hypothesis that the algorithm you named in part (b) will outperform the algorithm in your answer to part (c) on this data set. Explain why this is an appropriate evaluation methodology.

760-2 Understanding Learned Models

In some machine-learning applications, we are interested in understanding what a learned model tells us about the application domain when the model is sufficiently accurate. For example, a learned model in a medical application may inform us about previously unknown risk factors for a disease of interest. One approach to characterizing a learned model is to ask which features the model has deemed to be the most important.

(a) Describe a procedure for ranking features according to their importance in a learned naive Bayes classifier.

(b) Describe another procedure for determining which features are most important in a complex "black box" model such as a deep neural network classifier.

(c) Suppose we also want to assess the *stability* of feature importance measures. That is, we want to know how senstive the importance of each feature is to small changes in the training set. Describe an approach for doing this that could be used with both of the methods you described in parts (a) and (b).

(d) Discuss one limitation of feature importance as an approach to understanding a learned model.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Inferring Academic Genealogy

An academic genealogy organizes a family tree of scholars according to mentoring relationships. For example, part of the math genealogy is:

$$\begin{array}{ccccc} & & \text{Euler} & & \\ & & \downarrow & & \\ \text{Poisson} & \leftarrow & \text{Lagrange} & \rightarrow & \text{Fourier} \end{array}$$

For this question we take a broader view on genealogy: a directed edge between two scholars $A \rightarrow B$ means that A's work greatly influenced B.

You are given a set of scholars. For each scholar you only have

- A summary of their life-long work. Specifically, this is a bag-of-words vector that represents the count of words in that scholar's work. Here "work" means the collection of all papers authored by the scholar in his/her lifetime, and only includes the main paper content. Unfortunately, it does not include co-authorship, acknowledgments, or citations. You also do not have information on individual papers.
- The year when the scholar first published, and the year when the scholar last published.

Your task is to infer the genealogy among the set of scholars. We ask you to take a probabilistic approach and formulate the problem as statistical inference. Since this is an open-ended question, your answers below should be mathematically rigorous and contain sufficient details.

(a) List the major assumptions in your formulation.

(b) Define a probabilistic model for the genealogy problem. Be sure to define and explain your major variables.

(c) Define and discuss an inference procedure for your model.

(d) Discuss one way in which your model could be improved if you were given some new kind of data.

761-2 Debugging a Training Set

It's election time and every data analytic company is trying to build a binary classifier to classify every person x (in some very rich feature representation) in the US into their political leaning y, denoted L or R, so that advertisers can sell targeted ads. Your company is doing that, too, but there is a big problem. Your company acquired its large training set

$$S = \{(x_i, y_i)\}_{i=1:n}$$

through some murky black market before you joined the company. Everyone is highly suspicious of the quality of S: rumor has it that portions of the class labels were annotated using unreliable, under-paid crowdsourcing workers. However, n is just too big to exhaustively check the training items. On the other hand, your manager has total faith in the DeeeePNET training software used in the company so that has to be the training algorithm, even though nobody seems to have any idea how it works.

Under this circumstance, you and colleagues spent considerable money and effort to collect a new, trusted data set

$$T = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1:m}.$$

Every item in T is verified and correct. Due to resource constraints, $m \ll n$. Your hope is to use T to identify *potential* mislabels in S so someone can take a closer look.

(a) A simple approach is to find the same x in S and T with different labels. Given the rich feature representation x, though, very few x will appear in both S and T. Describe one approach that specifically addresses this issue without involving DeeeePNET. Discuss the major assumptions behind your approach.

(b) Your job just got harder: your manager, citing privacy concerns, decided to hide the feature vectors x from you. They are replaced by entry IDs. That is, S now looks like $(1, y_1), (2, y_2), \ldots, (n, y_n)$ to you, and T looks like $(n+1, \tilde{y}_1), \ldots, (n+m, \tilde{y}_m)$ to you. You can, however, make copies of the datasets, change any y labels there, and ask your manager to train DeeeePNET and classify items for you. Describe another approach to identify potential mislabels in S. Your approach should be mathematically rigorously defined, for example, as an optimization problem.

766 – COMPUTER VISION QUESTIONS

766-1 Perspective Projection and Optical Flow

Consider a pinhole camera with perspective projection.

(a) Given a circular disk that lies anywhere on a plane *parallel* to the image plane, what is the shape of the image of the disk? (You don't need to give a formal mathematical derivation; a qualitative description supported by correct reasoning and perhaps a figure should be sufficient.)

(b) Now, replace the disk with a sphere. What is the shape of the image of the sphere? (As above, you don't need to give a formal mathematical derivation; a qualitative description supported by correct reasoning and perhaps a figure should be sufficient.)

(c) Suppose the origin of the world coordinate system is at the pinhole. Let the z-axis be along the optical axis of the camera (perpendicular to the image plane), and the x and y axes are parallel to the image plane. Let the distance of the image plane from the pinhole be f.

Now, consider a point in the scene moving with velocity (u, v, w) from the starting point (x_0, y_0, z_0). Write the equation for the image coordinates (p, q) at time t. Show that the image point moves along a straight line as t increases.

(d) Recall the optical flow constraint equation from the paper "Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods".

This constraint states that the optical flow estimates (u, v) at a point (x, y) in image space lie on a straight line whose coefficients are the derivatives of image brightness in the x, y (spatial), and t (temporal) dimensions. Clearly, this constraint does not yield a unique (u, v) solution for each point (x, y). Now consider a structured environment (say, industrial) where illumination can be controlled (changed) at a speed much faster than the motion of objects in the scene. Using the optical flow constraint equation, show how two (perhaps unknown) different illuminations can be used to obtain a unique solution for (u, v) at each image point. (Hint: Two illuminations provide two constraints rather than one.)

766-2 Loopy Belief Propagation (LBP) for Early Vision

The underlying inference algorithms for many early vision problems such as stereo and image restoration are based on loopy belief propagation (LBP). This question relates to various modeling and efficiency considerations of the Felzenszwalb/Huttenlocher algorithm described in the paper "Efficient Belief Propagation for Early Vision", as used in computer vision.

Let $\mathcal{P}$ be the set of pixels in an image and $\mathcal{L}$ correspond to the set of labels. Generally, the labels correspond to the parameter (or quantity) we want to estimate at a pixel (e.g., depth/disparity, partition, intensity and so on). The objective corresponds to the sum of costs of assigning pixel $p \in \mathcal{P}$ to label f_p and assigning a pair of labels f_p and f_q to neighboring pixels p and q expressed as

$$\sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{(p,q) \in \mathcal{N}} W(f_p, f_q)$$

where $\mathcal{N}$ denotes the full set of neighborhoods (p, q).

(a) Assume that we choose the widely used quadratic cost of assigning pixels to labels for the function $D_\cdot(\cdot)$. To calculate the iteration-wise message passing cost update, Felzenszwalb/Huttenlocher suggest solving for a lower envelope of a collection of functions (i.e., pixel to label assignment costs). At a high level, describe a scheme which computes this lower envelope for a set of $k \ll |\mathcal{P}|$ such functions. (It is fine if the running time of your algorithm is worse than the best possible).

(b) Why do many proposed methods use a truncated quadratic function for $W(\cdot, \cdot)$ within LBP? In other words, what is the potential advantage of "truncation" for applications in low-level vision?

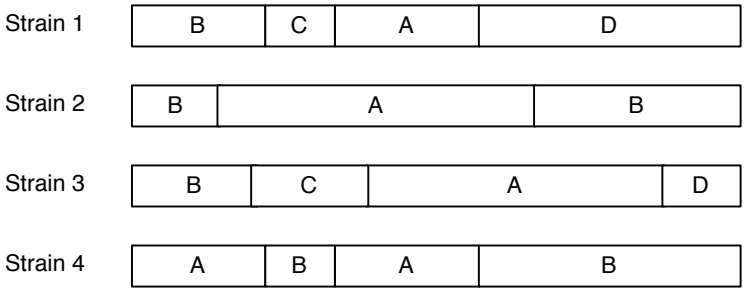
(c) Your LBP implementation is taking very long to run and someone suggests a simple "solution" that involves **(i)** dividing up the graph into equal pieces, running LBP on each piece separately and then concatenating the results together. Or **(ii)** changing the graph in a way to reduce its resolution by merging nodes in order to remove some of the loops. Describe one potential disadvantage of each of these two heuristics.

(d) Based on how Loopy Belief Propagation is used for Stereo, briefly describe how you could modify the procedure to identify SIFT feature matches between a pair of images. The module that you create may be used as a pre-processing step to fundamental matrix/epipolar geometry determination.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Viral Recombinant Quantification

Viruses often have many strains, each strain having a genome that is slightly different from the rest. Some viruses, such as HIV, have the ability to form "recombinant" or "hybrid" strains from two or more "parental" strains. The genome of such a recombinant strain is a mosaic of its parental strains, with different segments copied from different parents. Suppose that in a sampled viral population (e.g., from an HIV patient), there exist k strains, all of which are recombinants of four well-known parental strains (A, B, C, and D). The figure below gives the layout of each genome (in terms of the parental origin of each its positions), for an example scenario with $k = 4$.



Further, suppose that the viral DNA from this sample is randomly fragmented and sequenced, giving rise to N sequence reads, each with length L (where L is generally 100 or less). You are given this sequencing data and the full genome sequences of the k strains present within this sample and are asked to estimate the relative frequencies of these k strains. A key difficulty in this task is the fact that each sequence read is not guaranteed to uniquely identify a single strain. You should make the following assumptions:

- The genome sequences of all strains have exactly the same length and only differ in terms of substitutions.
- The genomes are much longer than the read length (e.g., the HIV genome is $\approx 10,000$ bases long).
- The reads may contain sequencing errors in the form of substitutions, but not insertions or deletions.

Given the defined task and these assumptions:

(a) Briefly describe how you will efficiently determine the subset of the k given strains that are most compatible with each read in terms of sequence similarity.

(b) Describe a method for estimating the relative frequencies of the strains using only those reads that are most compatible with a single strain.

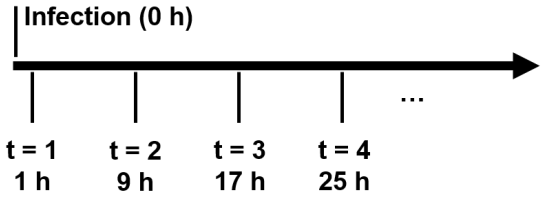
(c) Describe a method for estimating the relative frequencies of the strains using all reads.

(d) Describe the relative strengths and weaknesses of your methods for parts (b) and (c).

776-2 Immune Response Biomarkers

Suppose that you are studying the gene expression response of resistant and susceptible individuals to influenza infection. 1000 resistant and 1000 susceptible human subjects have volunteered to participate in the following experiment:

- The subject is infected with a non-lethal influenza virus. After the infection, an assay measures gene expression abundance, which is summarized into a single discrete state. Formally, the observed expression state at time t is a categorical variable with six possible values: $e_t \in \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F}\}$. For example, **A** may represent *high inflammation*, **B** is *metabolic shift*, etc.
- The expression data are collected every eight hours for two weeks post infection generating a time series of 42 expression states e_1 through e_{42}.



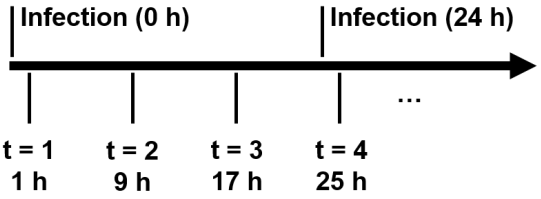
Assume that the resistant and susceptible subjects exhibit different expression responses. Your task is to model the expression data to predict whether a new individual is resistant or susceptible to infections.

(a) Your predictive algorithm will require computing $P(e_1, \ldots, e_{42}|\text{resistant})$ and $P(e_1, \ldots, e_{42}|\text{susceptible})$, the likelihood of the expression state data. Describe a probabilistic model that can be used to calculate these likelihoods. Your model should account for the dependency between the current expression state and the previous expression state.

(b) You suspect that the previous expression state e_{t-1} may not sufficiently capture the dependencies between the expression history and the current expression state e_t. Describe a statistical test to assess whether to include longer expression history in the probabilistic model.

(c) Suppose the test you design suggests that it is necessary to condition on the 6 previous time points when predicting e_t. What is one advantage and one disadvantage of using the longer expression history?

(d) Suppose the experimental design is modified such that the subjects are infected at the start of each day.



Describe how to modify your probabilistic model from part (a) to account for the new experimental design when computing the likelihood of the current expression state e_t.

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