Fall 2015
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 21, 2015

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 Naïve Bayes, Linear Models and Ensembles

1. Explain why naïve Bayes is a linear model. What are the coefficients of the linear model?
2. Explain why naïve Bayes can be viewed as an ensemble of features.
3. Are linear models and ensembles the same? Why or why not?

760-2 Online Learning for Regression

Consider the task of learning a regression model from wearable sensor data in an online setting. For example, suppose we want a model that represents an individual's heart rate as a function of accelerometer measurements, temperature, altitude, time of day, etc. Assume that all of the variables, including heart rate, are observable and sampled at the same frequency. Even though heart rate is measured during training, we are interested in modeling it to gain biological insight and to be able to predict it when it is not directly measured.

1. Define the concept of *online learning*.
2. Describe how you would approach this as a supervised learning task. Specify the learning algorithm you would use and justify this choice.
3. If it was expected that there would be *concept drift* (i.e. the relationship between heart rate and the other variables changes) over time, how would you adjust your approach?
4. How could the bias/variance tradeoff be controlled when using the algorithm you described above?

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Labeling Features

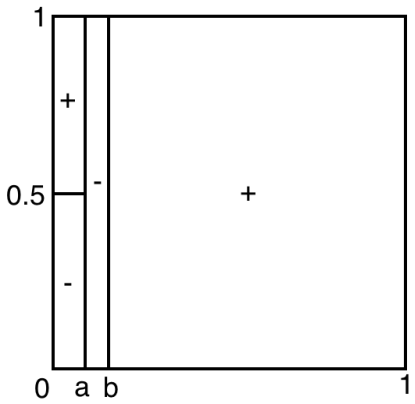
Consider logistic regression with binary features $f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}) \in \{-1, 1\}$ and binary labels $y \in \{-1, 1\}$:

$$p(y \mid \mathbf{x}) = \frac{1}{1 + e^{-y(\sum_{i=1}^{d} w_i f_i(\mathbf{x}) + w_0)}}. \tag{1}$$

You may assume there is a large unlabeled data set available to you.

1. The standard way to train the model is to collect a labeled training set $(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_n, y_n)$. Write down the optimization problem for finding the maximum likelihood estimate of $\mathbf{w}$. (Hint: use log likelihood)

2. In addition to the labeled training set, suppose a domain expert also provides *feature labels* for some of the features. A feature label for feature f_j is a binary variable $z_j \in \{-1, 1\}$. It intuitively means that feature f_j is indicative of class z_j. Note we have intentionally left the definition vague for you to have your own interpretation. As an example, let $\mathbf{x} = (x_1, x_2) \in [0, 1]^2$ and let the true class labels be in the following figure:



   Let $f_1(\mathbf{x}) = bool(x_1 \geq b)$ be the Boolean function which takes value 1 if $x_1 \geq b$, and value -1 otherwise. The domain expert labels f_1 as $z_1 = 1$ to indicate that f_1 is a positive feature. Intuitively, when f_1 "fires" ($f_1(\mathbf{x}) = 1$ in the rectangle $[b, 0], [1, 1]$) the label is always positive. Similarly, let $f_2(\mathbf{x}) = bool(x_1 \geq a)$. The domain expert also labels f_2 as $z_2 = 1$ to indicate that f_2 is (mostly) a positive feature.

   John thinks he knows how to incorporate feature labels into logistic regression training. His idea is straightforward: add constraints to the weights in the optimization problem. If feature f_j has label z_j, John's constraint is

$$z_j w_j \geq 0. \tag{2}$$

   What do you think of John's approach? Be sure to justify your answer. You may use the figure to help make your case.

3. Propose another approach to incorporate feature labels. Clearly state your assumptions. Explain your approach in sufficient detail.

761-2 From Word Embedding to Document Distances

Consider the problem of clustering documents by their semantic similarity. Consider these two documents:

(doc1) Obama addresses the media in Illinois

(doc2) The President greets the press in Chicago

We can represent each document as a bag-of-words (BOW) vector as follows. We first define a vocabulary with V distinct words. The BOW vector has V dimensions. The ith dimension takes the integer value of the number of times the ith vocabulary word occurs in the document.

1. We can define a distance between two documents as the Euclidean distance between their BOW vectors. With respect to the ultimate goal of document clustering by semantic similarity, what is one major disadvantage of this distance? Use doc1 and doc2 as an example in your answer.

2. Word embedding maps the ith vocabulary word w_i to a m-dimensional real-valued vector $x_i \in \mathbb{R}^m$. Recent advances in word embedding such as word2vec map semantically similar words to nearby points in $\mathbb{R}^m$. For instance, if w_i =Obama and w_j =President then the Euclidean distance $\|x_i - x_j\|$ is small. If each document has length one, we can simply use word embedding Euclidean distance as the distance between documents. But what if each document has length two? Define a distance using word embedding Euclidean distances as building blocks and explain your idea. Your distance should make the four documents "Obama media", "President press", "media Obama", "press President" all close to each other (assuming Obama and President are close, and media and press are close).

3. Now define a document distance for the situation when the documents have arbitrary (not necessarily the same) lengths. Your distance should seek the overall "best word match" for the two documents, again using the word embedding Euclidean distances as building blocks. Note you may need to scramble the word order to achieve the best match, and you need to handle different document lengths. (Hint: you may normalize each BOW vector so its elements sum to one.) We ask you to precisely define this document distance by formulating it as an optimization problem. Be sure to include the variables, the objective function, and any constraints if appropriate. Be sure to explain your design with sufficient details.

766 – ADVANCED COMPUTER VISION QUESTIONS

766-1 Lucas-Kanade Optical Flow

The Lucas-Kanade optical flow algorithm is among the most widely used methods for image alignment. This question deals with some of the formulation and optimization aspects of this algorithm.

1. Briefly describe the objective function that the Lucas-Kanade algorithm seeks to optimize. Provide some intuition behind the objective, describe the variables being optimized and how they correspond to a solution to the image alignment problem.

2. Describe any one difference between the (a) Newton and (b) Gauss-Newton approaches when used within the Lucas-Kanade algorithm.

3. A key computational issue in an efficient implementation of the Lucas-Kanade algorithm is efficiently computing the Hessian. Briefly discuss this issue and identify at least one heuristic used in practice to reduce the computational burden.

4. Most Newton-type approaches in Lucas-Kanade implementations are used when we are close to a local minimum. Describe any one strategy you could use to start the estimation process when it is far away from the local optimal solution.

5. The classical formulations of optical flow seem to work best when the lighting variations across the two image frames are fairly small. Describe a variation of the standard procedure that you could use (or whether you would use a completely different algorithm) when there are significant changes in lighting.

766-2 Pyramid Match Kernels

The pyramid match kernel is an important technique used for image categorization problems in computer vision. This question deals with various technical details of this idea.

1. Briefly describe how the Pyramid Match kernel algorithm computes similarities between unordered sets of features (for each image) to finally obtain a kernel matrix that can be used for regression and classification tasks.

2. Consider an alternative to the pyramid match kernel constructed in the following way: Given a set of real-valued feature vectors derived from an image, construct a single flat histogram based on a number of pre-defined, quantized bins. This will give a *vocabulary of words* representation and we simply count the frequency of occurrences of individual features over these quantized bins. Identify an advantage or limitation of the pyramid match kernel formulation over this alternative approach.

3. A key property of Pyramid Match kernels is that it satisfies Mercer's condition. Briefly describe why this property is relevant in image categorization experiments. Will this property be essential if we were using a k-nearest neighbors classifier?

4. Assume that our interest is not in image categorization, rather we want to use Pyramid Match kernels simply to identify similar features (or objects) across a set of images. Describe a reasonable strategy for achieving partial match correspondences across images using Pyramid Match kernels.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Genome Analysis without DNA sequence

Suppose we are interested in studying the genome of species X, but instead of knowing the DNA sequence of the genome, we have multiple measurements of biochemical activity (e.g., transcriptional activity, levels of transcription factor binding, levels of histone modification) at each position of the genome. Specifically, we have m different real-valued biochemical measurements across the n positions of the genome, and thus the data may be represented by an $m \times n$ matrix, with measurements indexing rows and genome positions indexing columns.

1. Suppose we believe that each position of the genome belongs to one of k functional classes and that positions belonging to the same class have similar biochemical activity profiles. Describe an unsupervised approach for classifying each position as one of k functional classes that does not take positional information into account.

2. After classifying the genomic positions using your approach from (1), you wish to determine whether there are statistically significant dependencies between the functional classes of nearby positions. Describe an approach for detecting such dependencies, should they exist.

3. Assuming you detect dependencies in (2), describe an unsupervised approach for classifying the genomic positions that takes positional information, and thus the detected dependencies, into account.

4. Suppose we obtain the same types of biochemical activity measurements for the genome of species Y. Describe an approach for aligning the genomes of X and Y using the biochemical activity measurements instead of DNA sequence. You may assume that the genomes are collinear and composed of a single chromosome.

776-2 Bayesian Networks for Gene Expression Networks

Recall the Bayesian network representation of gene networks. Suppose that you had gene expression levels of N genes measured in m different experimental conditions. That is, each gene has m measurements.

1. Let X_i denote the random variable for the expression level of the i^{th} gene and let $Pa(X_i)$ denote the parents of X_i in a Bayesian network. Give two ways to model the Conditional Probability Distributions (CPD) $P(X_i|Pa(X_i))$, and describe two distinguishing properties for each. State the assumptions you need to make to use these forms of CPDs for gene expression data.

2. Briefly describe CPDs in the context of Module networks and how they are estimated.

3. Suppose that the module membership of a gene is also influenced by the promoter sequence of that gene. How would you change the Module network algorithm to incorporate this property?

4. Suppose you were told that the m different conditions come from k different classes. How would you extend Module networks to integrate the class of the experimental condition?

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