Fall 2018
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 17, 2018

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 Mathematical Background for Machine Learning

1. Let $\sigma(x) = \frac{e^x}{1+e^x}$. Compute the derivative $\frac{d}{dx}\sigma(x)$.

2. Compute the mean and the variance of the numbers $\{2, 3, 5, 6, 9\}$.

3. Prove that the smallest Euclidean distance from the origin to some point $\mathbf{x}$ in the hyperplane $\mathbf{w}^T\mathbf{x}+b = 0$ is $\frac{|b|}{||\mathbf{w}||_2}$. Show the steps of your proof.

4. Let X, Y be two random variables, which take values in a finite discrete set V. Let $H(Y)$ denote the entropy of Y, and let $H(Y|X)$ denote the conditional entropy of Y conditioned on X. Prove that if X and Y are independent, then $H(Y|X) = H(Y)$.

760-2 Supervised Learning, Generative Learning, and Methodology

Suppose that the electronic health record (EHR) company Epic gives you data on drug prescriptions and disease diagnoses for patients. Data consist of tuples of the form $\langle$*patient*, *date*, *event*$\rangle$ where *event* may be any drug or disease diagnosis. They want you to build a model that associates drugs and diseases with one another. You may ignore the date component if you wish in your answers.

1. What will constitute your variables, or features for this task?

2. What will constitute a training example for this task?

3. What type of model (what representation) will you learn?

4. What learning algorithm will you use? Naming an algorithm is not sufficient; please provide the algorithm in sufficient detail that a knowledgeable computer scientist could implement it.

5. How will you evaluate your approach? Please describe both your methodology and your precise metric(s) for the quality of the learning algorithm's result.

6. Name and discuss one advantage of your approach (representation and/or algorithm) over other possible approaches.

7. Name and discuss one shortcoming or disadvantage of your approach (representation and/or algorithm).

8. Describe one alternative learning algorithm you could use to at least partially alleviate the above shortcoming (possibly at the cost of other shortcomings instead). Your alternative learning algorithm could use the same representation or a different one.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Classification

Consider a multiclass classification problem with K classes. Let the feature space be the unit hypercube $[0,1]^d$, for some $d \geq 1$. Assume that you are given n iid training examples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in [0,1]^d$ and $y_i \in \{1, 2, \ldots, K\}$.

1. Describe how you would construct a histogram classifier with $M = m^d$ equal-sized bins based on the n training examples. Explain how it predicts the label for a new example $\boldsymbol{x} \in [0,1]^d$.
2. Let $\widehat{y}$ denote the predicted label for the new example $\boldsymbol{x}$. $\widehat{y}$ is a random variable depending on the training data. What is the conditional probability distribution of $\widehat{y}$ given $\boldsymbol{x}$?
3. Let $\mathcal{H}$ denote the set of all M-bin histogram classification rules. For any $h \in \mathcal{H}$, let $\widehat{R}(h)$ denote the empirical error on the training data. Give a mathematical expression for $\widehat{R}(h)$.
4. Derive the mean, variance, and probability distribution of $\widehat{R}(h)$. The value of the mean corresponds to what key quantity?
5. Explain how the histogram classifier you designed in (1) can be viewed as the result of an empirical risk minimization process.
6. Recall that uniform deviation bounds are central to the analysis of empirical risk minimization. Derive a uniform deviation bound of the form

$$\mathbb{P}(\max_{h \in \mathcal{H}} |R(h) - \widehat{R}(h)| > \epsilon) \ \leq \ \delta$$

   by specifying ϵ in terms of δ, M, n, d, and K.

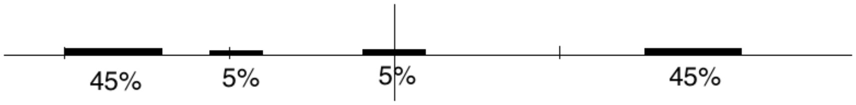
761-2 Something Wrong with Uncertainty-Based Active Learning

Consider a binary classification task with 0-1 loss.

1. Suppose a hypothesis class $\mathcal{H}$ has a finite VC dimension. Let
$$h^* \in \mathcal{H}$$
be the minimizer of the risk under distribution D:
$$h^* \in \arg\min_{h \in \mathcal{H}} L_D(h).$$
Here $L_D(h) = \mathbb{E}_{(x,y)\sim D} 1_{[h(x)\neq y]}$. For an iid training set S drawn from D, let $\hat{h}$ be the empirical risk minimizer. What does PAC learning say about $L_D(\hat{h})$ compared to $L_D(h^*)$ as the size of S approaches infinite?

2. Active learning does not receive iid training data. One popular active learning "principle" among practitioners is uncertainty based. In words, uncertainty based active learning queries the point where the learner has the highest uncertainty. Let $\mathcal{H}$ be 1D threshold classifiers. Let us define the highest uncertainty point to be the midpoint between the inner-most pair of positive query and negative query within historical queries (you may think of the learner as an SVM, then the most uncertain point is where the decision boundary is). Write a pseudo code to define an uncertainty-based active learning algorithm that fits the above specification.

3. We give you a specific D in the following picture:



The percentages are the marginal probability of x in each thick line segment. Your job is to define the conditional distribution $P_D(y \mid x)$ in such a way that the risk of your uncertainty-based active learning algorithm often fails to converge to $L_D(h^*)$ even with an infinite amount of queries. Describe your $P_D(y \mid x)$, and explain why uncertainty-based active learning can fail.

Hints:

(a) make the conditional distribution $P_D(y = 1 \mid x)$ "crisp", i.e. it is either 0 or 1 for any x.

(b) D does not need to be linearly separable.

766 – COMPUTER VISION QUESTIONS

766-1 Optical Flow

1. Consider two images captured by a camera in quick succession, at time instants t and $t + \partial t$. Let the image intensity at a pixel (x, y) in the first image be $I(x, y, t)$. Due to scene motion, suppose the scene point imaged at (x, y) in the first image gets imaged at pixel $(x + u, y + v)$ in the second image. Here, (u, v) is called the optical flow vector and it describes the motion vector at pixel location (x, y) in the image's space. Assuming that the intensity of the scene point remains constant over time, we get the following relationship between image brightness values:

   $I(x, y, t) = I(x + u, y + v, t + \partial t)$.

   Suppose that the amount of motion (u, v) is very small. Use a first-order Taylor approximation of the above equation to derive a relationship between image gradients I_x (x-gradient), I_y (y-gradient), I_t (time gradient), and the optical flow vector (u, v). This is the well-known optical flow constraint equation from the paper "Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods".

2. Is it possible to recover the two unknowns (u, v) uniquely for each pixel (x, y), simply from the above equation? Why or why not?

3. Next, assume that the optical flow vector (u, v) is constant within a local neighborhood in the image, i.e., (u, v) is the same for all the pixels in a local window. Describe a method for computing the optical flow using this assumption. Explain when this method will work, and when it will not work (in terms of the scene texture within the image window).

4. 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 Eigen-faces and Face Recognition

Eigen-faces is a widely used algorithm in computer vision both for face recognition as well as a variety of other applications.

1. What are the basic steps of the Eigen-faces algorithm? Describe the form of the input data, what is internally optimized by the algorithm and how this procedure can be used to discriminate between face images and non-face images. Define the notations you use clearly.
2. Suppose that the Eigen-faces algorithm is being used for person identification for access control in a secure facility. If the dataset you are provided for training includes face images of the participants with significant lighting and viewpoint variations, will Eigen-faces be invariant to these issues, by construction? Explain why or why not in enough detail to support your answer.
3. Consider a situation where you are provided a small set of face images where a part of the face is occluded. Assume that you also have a large set of non-occluded face image data available. Describe how you can perform *inpainting* (filling missing parts of faces) on the occluded faces dataset using Eigen-faces.
4. Consider a setting where the input data for training your algorithm included short (2–3 second) “selfie” videos of individuals with varying background and some camera shake, instead of individual face images. Describe the steps you will take so that Eigen-faces or another similar algorithm can be used with these data.
5. Describe how you will set up and train an artificial neural network that provides similar or identical functionality as Eigen-faces.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Cancer genome sequencing

DNA alterations are often a cause of the cellular changes that lead to cancer. DNA mutations, insertions, and deletions can alter protein function. Large-scale DNA abnormalities such as increases or decreases in copy number and translocations also disrupt normal cellular processes and promote tumor growth. Mapping the differences between an individual's normal (germline) DNA and the DNA of their cancer cells can be a starting point for understanding the causes of cancer in that individual or the common cancer-promoting alterations across a population of cancer patients.

Here you will explore the computational challenges of mapping these DNA alterations in human cancer. For all four subproblems below you may assume you already have the complete and correct normal DNA sequence (genome) for the individual. For simplicity, further assume that the cancer DNA comes from a single homogeneous population of cancer cells from one individual such that all of your cancer sequencing reads are generated from the same underlying sequence.

1. Suppose you are using traditional high-throughput sequencing to sequence the cancer genome and aligning these reads to the normal genome. Your sequencer produces single-end reads that are 100 base pairs (bp) in length. On average, each position in the cancer genome is covered by 50 reads. Describe why these reads will or will not be sufficient to detect whether there is a single nucleotide mutation in a cancer gene of interest. State any additional assumptions that you make.

2. Now consider a large-scale copy number amplification in which a contiguous region of the genome is duplicated many times. The duplicate copies appear consecutively such that the last nucleotides in one copy are immediately followed by the first nucleotides of the next copy. For example, the amplification may be a continuous 100,000 bp region of a chromosome copied 5 times. Given the same 100 bp single-end reads, describe why they will or will not be sufficient to detect how many copies of the genomic region are added in the copy number amplification. State any additional assumptions that you make.

3. Suppose you are instead using a long-read sequencing technology. The reads are long enough that you are able to sequence entire chromosomes from the cancer genome. The cancer genome is hypermutated, containing an extreme number of single nucleotide mutations, insertions, and deletions. In addition, the genome contains a small number of amplifications, deletions, and translocations. Describe an algorithm to align the normal and cancer genomes with special emphasis on the data structures you will use.

4. Given the same long-read sequencing technology, now consider a cancer genome with a much lower mutation rate. There are single nucleotide mutations, insertions, and deletions on each chromosome, but much fewer than in part 3. As in part 3, there are amplifications, deletions, and translocations as well. Describe an algorithm to align the normal and cancer genomes with special emphasis on the data structures you will use. Your approach should be different from part 3 in order to take advantage of the lower mutation rate.

776-2 Cell differentiation trajectory inference

A single-cell gene expression assay is applied to a population of cells that are undergoing a differentiation process. We model this differentiation process by considering each cell to be in one of a number of discrete *cell states* at a given time, and with cells transitioning over time from one state to another. The structure of the possible cell state trajectories within this differentiation process is known to be a tree, with the root node representing the progenitor cell state (i.e., cells that have not begun differentiating), the leaf nodes representing fully differentiated cell states, and the internal nodes representing intermediate cell states. Each internal node in the tree has at most two children. An individual cell from the population may come from any node (cell state) along this tree structure.

You are given a data set consisting of gene expression profiles from 1,000 cells sampled from this population. Each gene expression profile is a binary vector of length 20,000 with the ith entry of the vector indicating whether ith gene is expressed ("on") or not expressed ("off") within the cell. The assay used to produce these gene expression profiles is not perfectly accurate, with each entry of the gene expression profile being incorrect with probability 0.01 (and with errors being independent of each other). You should assume that the number of sampled cells is much greater than the number of possible cell states along the tree structure.

1. With the goal of inferring the tree structure for this differentiation process, a researcher decides to naïvely apply a phylogenetic tree reconstruction algorithm to these data. For a particular phylogenetic tree reconstruction algorithm of your choosing, describe how it could be applied to these data.
2. Explain why a straightforward application of a phylogenetic tree reconstruction algorithm will *not* succeed in inferring the tree structure of the differentiation process.
3. Describe an alternative approach for inferring the tree structure that overcomes the limitations you described in part 2.
4. For the same differentiation process, suppose that your dataset now comes from a population of cells that are all fully differentiated (i.e., they map to states corresponding to leaves of the tree). Describe an approach for inferring the gene expression profile of the progenitor cells.

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