Fall 2017
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 18, 1-5pm, 2017. 1240 CS

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.

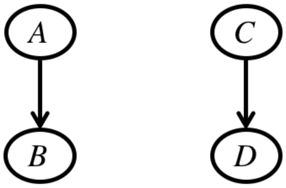
POLICY ON MISPRINTS AND AMBIGUITIES

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 Bayes Net Learning

Consider the following Bayes net structure for the four Boolean variables $A...D$.



1. Justify the statement that this Bayes net provides a *compact* representation of the joint probability distribution of the variables $A...D$.

2. Given the training set below, show the estimated parameters for the network when using Laplace estimates (i.e. add-1 smoothing).

A	B	C	D
t	f	t	f
t	t	t	t
f	t	t	f
f	t	f	t

3. Suppose you are given the additional training instance shown below, which has a missing value for variable C. Show how you would use one step of the EM algorithm to update the network parameters. **Note:** you can show your answers for this part using products/sums of fractions. You do not need to simplify these expressions.

A	B	C	D
t	t	?	t

4. Would you characterize the approach you used as *generative* or *discriminative*? Justify your answer.

760-2 Deep Neural Networks

Until a decade ago, conventional wisdom was that including more than one hidden layer in an artificial neural network (ANN) was not effective. In the last decade this story has changed dramatically.

1. Name *three* key techniques, aside from more data and powerful hardward, developed within the last decade that have led to effective learning of deep neural networks.
2. For each of the three techniques you listed, provide a paragraph description detailed enough that a neural net practitioner from a decade ago could implement it after reading your description.
3. For each of the three techniques, explain why it can improve learning in deep neural networks.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Auto-fill

Consider automatically filling tables as a machine learning problem. As our first example, consider the following table where the user just typed in 26133 in the second column:

26133 Oldenburg Germany	26133
50939 Koln Germany	
Bonn 53131 Germany	
10724 Berlin Germany	
. . .	

We want the computer to auto-fill in the second column with zip code like this:

26133 Oldenburg Germany	26133
50939 Koln Germany	50939
Bonn 53131 Germany	53131
10724 Berlin Germany	10724
. . .	. . .

Note this is non-trivial since another valid auto-fill rule is "the first word" which will fill in "Bonn" on the 3rd row. For this question, a word is an alpha-numeric string seperated by white space or punctuation.

Here is example 2, where the user typed OG in the second column:

26133 Oldenburg Germany	OG
50939 Koln Germany	
Bonn 53131 Germany	
10724 Berlin Germany	
. . .	

We want the computer to auto-fill the first and last letter of the city. But there are again many valid auto-fills:

- The constant string "OG"
- First letter of the 2nd word + "G"
- First and last letters of the 2nd word, capitalized
- First and last letters of the 1st non-number word, capitalized
- First letters of the 2nd and 3rd word
- . . .

Here is example 3, where the user typed the first 4 entries in the 2nd column, and the computer is expected to auto-fill the remaining rows in the table as standardized phone numbers:

(425)-706-7709	425-706-7709
510.220.5586	510-220-5586
1 425 235 7654	425-235-7654
425 745-8139	425-745-8139
. . .	

1. Pose the auto-fill problem as a machine learning problem. Specifically, tell us how you define a hypothesis (you can make reasonable assumptions), the hypothesis space, the learning algorithm, the training examples, the underlying distribution, and anything else you think is important.
2. Give a generalization error bound. Explain how you define the capacity of the learner.
3. What would you do to make auto-fill successful while requiring only minimal user input? Be sure to explain your answer in machine learning terms.

761-2 Gaussian Mixture Model

Imagine the following biological application. Suppose that we measure the difference in expression levels of n genes in healthy and diseased cells. Most genes will have no real difference other than random measurement noise, but a small fraction will have a nonzero difference. (It is not essential for you to be familiar with biology to answer the mathematical questions below.)

Specifically, let $x_i, i = 1 \ldots n$ denote the measured difference between the expression of gene i in diseased and healthy cells. Consider the following probabilistic model of these data.

$$x_i \overset{iid}{\sim} (1-p)\mathcal{N}(0,1) \;+\; p\mathcal{N}(\theta,1) \; .$$

This is a Gaussian mixture model. Here p represents the fraction of cells that have a different expression level in the case of disease, and θ represents the unknown difference in the expression level in such cases. We are interested in estimating p and θ.

1. Derive a formula for the likelihood of p and θ given the data $x_1, \ldots, x_n$.
2. What equations must be solved to obtain maximum likelihood estimates of p and θ? Can they be solved in closed-form? If not, suggest a numerical procedure to solve them.
3. Derive formulas in terms of p and θ for the first and second moments of the mixture distribution.
4. These formulas suggest an alternative to maximum likelihood estimation. Use these formulas to propose a different approach to estimating p and θ from the data $x_1, \ldots, x_n$. In this case you should be able to obtain closed form formulas.

766 – COMPUTER VISION QUESTIONS

766-1 Image Features and Filtering

A standard approach for low-level image processing and finding features (e.g., corners, edges, blobs, SIFT, etc.) in an image is to compute the convolution of the image with a filter.

Consider a 2D image $I(x, y)$ of size $N \times N$ pixels (N rows, N columns), and a 2D filter $f(x, y)$ of size $M \times M$ pixels (M rows, M columns). For example, $f(x, y)$ could be a 2D Gaussian filter, a wavelet filter, a Laplacian filter, etc. Suppose we wish to compute the convolution of I and f, i.e. $C(u, v) = (I \star f)(u, v)$, where $\star$ is the convolution operator. One approach for computing the value of $C(u, v)$ is to compute the dot-product of the filter f (after flipping horizontally and vertically) with the sub-image of I (of size $M \times M$) centered around the pixel (u, v). The same process is repeated for all locations $(u, v), 1 \leq u \leq N, 1 \leq v \leq N$.

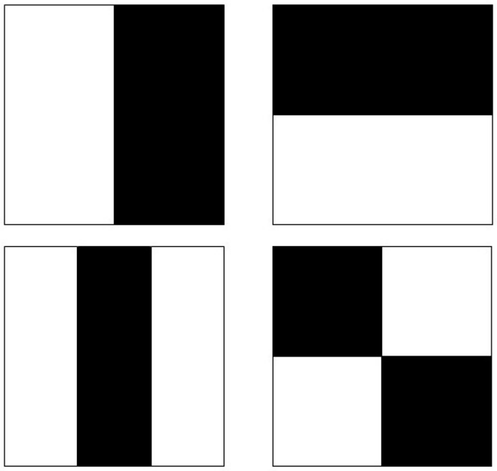
1. How many addition and multiplication operations are needed to compute the convolution of an image of size $N \times N$ with a filter of size $M \times M$, using the above procedure? (You do not need to give the exact number; it is sufficient to provide the number of additions and multiplications in order notation, in terms of M and N)

2. Now, suppose the filter is a 2D Gaussian of size $M \times M$. Describe a simple algorithm (backed by appropriate mathematical statements) to compute the convolution faster using $O(M\,N^2)$ additions and $O(M\,N^2)$ multiplications. (Hint: Can a 2D Gaussian be expressed as a convolution of two simpler filters?)

3. Next, suppose the filter is a 2D Haar filter, as used in the paper 'Robust Real-Time Face Detection, Viola and Jones'. Haar filters consist of a small number of axis-aligned binary rectangles. Some example Haar filters are shown in the figure below (black denotes 0, white denotes 1).

   Show that the convolution of an $N \times N$ image with a Haar filter can be computed with only $O(N^2)$ additions (independent of the size of the filter!).

   Does your method require any pre-computations? If so, how many computations (additions and multiplications) are required for your pre-computation routine?



Example Haar Filters

766-2 Epipolar Geometry and Fundamental Matrix

Let P and P' be two cameras with non-coincident centers. Let $\mathbf{x}_i$ and $\mathbf{x}'_i$ be the i^{th} corresponding pixels (or points) such that

$$\mathbf{x'}_i^T F \mathbf{x}_i = 0.$$

Here F is called the fundamental matrix and $\mathbf{v}^T$ gives the transpose of the vector.

1. Describe the rank, degrees of freedom, and at least one more property of F.

2. Consider the case of pure translation between the two cameras. This means that there is no rotation and no change in internal paramters. Describe two properties of F and/or the corresponding pixels in this special case.

3. Briefly describe if F can be thought of as a type of correlation. Why or why not?

4. Assume you are provided n distinct pixel correspondences (one to one) between the two views to estimate F. Assume n is sufficiently large. But $\frac{n}{4}$ of these correspondences are incorrect. Describe a reasonable procedure to estimate F with such information and briefly state any guarantees you can offer regarding your estimate $\hat{F}$ w.r.t. the optimal F^*.

5. Assume you are provided n distinct pixel correspondences (one to one) between the two views to estimate F. Unfortunately the coordinates $\mathbf{x'}_i$ and $\mathbf{x}_i$ contain measurement noise. That is, you only observe $\mathbf{x'}_i + \epsilon'_i$ and $\mathbf{x}_i + \epsilon_i$ where the ϵ's are *iid* random variables drawn from a coordinate-independent distribution. Choose any parameterization for the noise model and describe a procedure you may use to estimate F.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Motif finding on phylogenies

A DNA sequence motif is used to represent the pattern of DNA binding sites of a transcription factor protein. Such motifs are often identified by searching for over-represented DNA subsequences in the promoters of co-expressed genes. True binding sites tend to be conserved across species, that is, binding sites are likely to occur in evolutionarily conserved regions of the genome. In this question, we will aim to use evolutionary conservation to find a motif of a pre-specified length l given the following inputs:

- Genome sequence of m different species.
- A subset of genes in one of the species, called the reference species, that likely have the binding site described by the motif in their promoter sequence.
- The one-to-one orthology of genes across species.
- Genomic coordinates of genes in each species from which promoter sequences can be extracted.

1. Describe a probabilistic model for representing a sequence with one or more motif instances in one species. Describe how you will estimate the parameters of this model. Your model need not account for evolutionary conservation.
2. State the computational sub-problems you need to solve to identify a motif in the reference species while using evolutionary conservation of sequences across species.
3. Describe an approach that uses sequence conservation to find a motif in the promoters of the input subset of genes in the reference species. Briefly describe the solutions to each sub-problem stated in 2 along with any assumptions you make about your approach.

776-2 Predicting signaling pathways

Signaling pathways describe how proteins propagate information within cells in response to extracellular triggers. Computationally, they can be represented as a graph $H = (V', E')$ with protein vertices V' and protein interaction edges E' of the form (v_i, v_j). Given a large undirected protein-protein interaction network $G = (V, E)$, the signaling pathway prediction problem is to extract a subnetwork $H = (V', E')$ from G subject to some optimization criteria such that $V' \subset V$ and $E' \subset E$.

In this question, your goal is to predict a signaling pathway subnetwork given the following inputs:

- An undirected protein-protein interaction network $G = (V, E)$.
- Weights w_{ij} for all edges (v_i, v_j) in G that represent the interaction confidence. All weights are in the range $(0, 1]$, with 1 representing the most confident interactions.
- Capacities c_{ij} for all edges (v_i, v_j) in G that represent the interaction capacity, a limit on how much the interaction can contribute to the signaling response. All capacities are in the range $(0, 1]$.
- A set of receptor proteins $R \subset V$ that recognize stimuli for the pathway of interest.
- A set of transcription factor proteins $F \subset V$ that regulate important genes in the condition of interest after the pathway is activated.

1. Formally define an optimization problem to predict a signaling pathway from the above inputs. The pathway should connect the receptors to the transcription factors through 0 or more intermediate vertices. The formulation must account for edge weights and capacities. Include an objective function and any necessary constraints. Explain the biological motivation for your objective and your assumptions.
2. Can the optimization problem in part 1 be solved efficiently? Justify your answer.
3. In addition to the data above, you obtain a set of differentially phosphorylated proteins $P \subset V$ that are likely to participate in the signaling pathway. Describe how to extend your solution to part 1 to prefer subnetworks that include proteins in P.
4. What is one limitation of the optimization problem you defined in part 1? Describe a different network optimization problem that, as before, links the proteins in R and F in the weighted network G but overcomes this limitation. Your new approach does not need to use edge capacities. You do not need to provide an algorithm to solve the new optimization problem.

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