Spring 2015

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

Artificial Intelligence

Monday, February 2, 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.

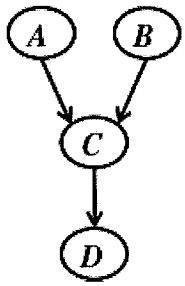
## POLICY ON MISPRINTS AND AMBIGUITIES

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 Bayes Net Learning

Suppose you are given the following Bayes net structure for five Boolean variables $A \ldots D$.




(a) In what way does this Bayes net provide a *compact* representation of the joint probability distribution for the variables?

(b) Estimate the parameters of the Bayes net using the following data set and Laplace estimates. Show your work.

| $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|
| t | f | t | f |
| t | t | t | t |
| f | t | t | f |
| f | t | f | t |

(c) How would you modify the parameter-estimation procedure used above to take into account a prior belief that $A$ is twice as likely to be *true* than *false*?

(d) 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 you calculated in part (b). **Note:** You can show your answer for this part using products and sums of fractions. You do not need to simplify these expressions.

| $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|
| t | t | ? | t |

## 760-2 Overfitting, Nearest Neighbor, and Decision Trees

(a) Define *overfitting.* Explain why it is important.

(b) Describe one common way to avoid overfitting in *decision trees.*

(c) Describe one common way to avoid overfitting in *nearest neighbor* classification.

(d) Identify and explain two advantages of decision trees over nearest neighbor classification (they need not relate to overfitting).

(e) Identify and explain two advantages of nearest neighbor over decision trees (they need not relate to overfitting).

# 761 – ADVANCED MACHINE LEARNING QUESTIONS

## 761-1 The *iid* Assumption

*iid* is the acronym for independent and identically distributed.

(a) How is the *iid* assumption used in the PAC analysis of binary classification?

(b) Tom carefully drives a golf cart on a paved road in a park every day. He instrumented his cart so that every second a camera takes an image $\mathbf{x}_t$ of what is ahead of the cart. Simultaneously Tom's steering angle $y_t$ is measured. Tom learns a regression function $f : X \mapsto Y$ from last year's $(\mathbf{x}_t, y_t)$ data for $t = 1 \ldots T$, where $T$ is the total number of frames he collected last year. Tom plans to test $f$ on this year's data of him driving the cart. Under what assumptions can we treat last year's and this year's data as *iid*?

(c) Tom plans to use $f$ to make his golf cart an autonomous vehicle. He will take an image $\mathbf{x}$ every second, and will use a motor to turn the steering wheel to angle $f(\mathbf{x})$. Ignore all other controls (e.g., do not worry about braking). There is one serious violation of *iid*-ness in this plan, even when all the assumptions you made in part (b) are true. Explain what the problem is.

(d) Suggest one major way for Tom to improve the steering of his autonomous vehicle.

## 761-2 Self-Paced Learning

Consider learning a probabilistic model with hidden variables on a classification task. The training data consists of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i$ is a feature vector and $y_i$ the label. The model contains a vector of hidden variables $\mathbf{h}_i$ which is not observed in the training data.

(a) Define a generic maximum-likelihood based learning algorithm. Your likelihood should include $x, y, h$. You can include regularization on the parameters. Explain why, with hidden variables, such a likelihood function is nonconvex in general.

(b) A major problem with a nonconvex likelihood is the learner being trapped in local minima. To hopefully arrive at a better local optimum, one idea is for the learner to perform "self-paced learning." The idea is inspired by how students learn. Instead of training on the full data set, a student can start from a subset of the training data that she is most comfortable with. She learns from this subset, becomes better, and selects another subset that she is now comfortable with. This procedure iterates until she uses all the training data. Formulate this idea of self-paced learning with respect to the learning algorithm you proposed in part (a). Be sure to give enough mathematical details on what "comfortable" means, and how the subsets are selected.

(c) Is your self-paced learning algorithm the same as active learning? Explain why or why not.

(d) Discuss whether or not your self-paced learning algorithm is guaranteed to solve the local minima issue.

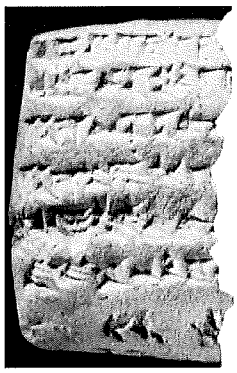
# 769 – ADVANCED NATURAL LANGUAGE PROCESSING QUESTIONS

## 769-1 Archeological Language Models

You are a computational linguist assisting a team of archeologists on the hunt for ancient human languages. On a recent field expedition the archeologists have discovered a human language called ALGOL, etched on 5,000 year old clay tablets (no relation to the ancient programming language of the same name). The team has identified which symbols on the tablets represent words. However, the team is unsure whether sentences are formed from left-to-right, right-to-left, or top-to-bottom and they ask for your help.

(a) Formulate a Bayesian model to compute the probabilities of these three possible writing directions. Your model should incorporate a **trigram language model**, as well as information about the (known) writing directions of other ancient languages from the same region. Be explicit about the components of your model and how you estimate their parameters.

(b) The team has now discovered a single broken clay tablet which is missing its right-hand side like so:



They ask for your help in computing a probability distribution over the width of the missing side. For this question, make the following assumptions:

(i) the writing direction of ALGOL is from left-to-write,

(ii) ALGOL tablets always have one sentence per line,

(iii) all words in ALGOL have equal width, and

(iv) the width of the tablet is equal to the width of the **longest sentence** in it.

Using a trigram language model, formulate the probability distribution over the width of the missing right-hand-side.

## 769-2 Archeological Parsing

You are a computational linguist assisting a team of archeologists on the hunt for ancient human languages. On a recent field expedition the archeologists have discovered a human language called ALGOL, etched on 5,000 year old clay tablets (no relation to the ancient programming language of the same name). After much effort, the team has succeeded in analyzing the syntax of the language and has produced a treebank consisting of some hand-parsed ALGOL sentences. It is your job to build a model to parse the remaining sentences.

(a) Provide an estimation procedure for a "vanilla" PCFG model using the ALGOL treebank.

(b) Describe a technique to improve the PCFG model by automatically "decorating" the non-terminal symbols with contextual information. Explain why your technique should do a better job than the vanilla PCFG.

(c) Now that you have trained your model, you must use it to parse the remaining ALGOL sentences. Unfortunately, you must deploy your model on an ancient computer that is too slow to run the $O(n^3)$ time CKY algorithm (where $n$ is the length of the sentence). However, you notice that all the grammar rules of ALGOL conform to one of the following two schemes:

$$A \rightarrow w_1 ... w_k$$
$$A \rightarrow w_1 ... w_k B$$

where $A$ and $B$ are arbitrary non-terminals, and $w_1 ... w_k$ is a sequence of arbitrary words of arbitrary length $k$. In other words, ALGOL's grammar is **right-branching**.

Formulate a parsing algorithm that will exploit this fact in order to parse ALGOL sentences in time $O(n^2)$. Provide pseudo-code for your algorithm. (**Hint:** It may be easiest to start with the $O(n)$ Viterbi algorithm and modify it to account for "variable length emissions").

# 776 – ADVANCED BIOINFORMATICS QUESTIONS

## 776-1 Multiple Motif Finding

Suppose you perform a ChIP-Seq experiment for an uncharacterized transcription factor X. After analysis of the ChIP-Seq data, you identify $N$ short genomic intervals, each of which is likely to contain a binding site for X. Your task is to identify the location of the binding site for X within each of the intervals using a motif-finding approach. Unfortunately, it is believed that X binds to multiple distinct motifs, none of which are known.

(a) Describe a probabilistic model for this task. Assume that there are $k$ distinct motifs, each of length $w$, and that each interval contains exactly one instance of one of the motifs.

(b) Describe an approach that uses your model to identify the binding sites for X within the $N$ intervals.

(c) Now suppose that the ChIP-Seq experiment was noisy and that some unknown fraction of the $N$ intervals does not contain a binding site for X. Briefly explain how you would modify your model in (a) and approach in (b) to accommodate this aspect of the data.

(d) Now suppose that $k$ is unknown. Your colleague suggests that you could learn $k$ by trying many different values of $k$, estimating the maximum likelihood parameters of your model for each value of $k$, and picking the $k$ that produces the maximum likelihood. Ignoring computational complexity issues, will your colleague's strategy work? Briefly explain your reasoning.

## 776-2 Gene Expression Clustering

Suppose you are given (1) $s$ $m$-by-$T$ gene expression matrices, one matrix for each of $s$ species, with each matrix measuring the expression of $m$ genes in $T$ different microarray experiments, and (2) a phylogenetic tree of the $s$ species; you may assume that the genes in each species are in one-to-one correspondence with the genes in the other species and this correspondence is given as input.

(a) Describe a clustering approach that identifies $k$ clusters in each species while exploiting the evolutionary relationships between the genes. You can assume that genes that share evolutionary histories often have similar function.

(b) How will you assess the quality of the results obtained from your approach compared to a baseline approach that does not use the evolutionary relationship between genes?

(c) Suppose that your collaborator measured expression in three different conditions, measuring $T_1$, $T_2$ and $T_3$ different time points in each condition producing a dataset of $T_1 + T_2 + T_3$ different measurements in each of the $s$ species. How can you extend your approach in part (a) to handle this larger dataset given that the clusters in one condition may not necessarily be the same as the clusters in a different condition? Simply concatenating the expression data into a single vector is not acceptable.

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