FALL 2012
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 24, 2012

GENERAL INSTRUCTIONS:

(a) This exam has 12 numbered pages.

(b) Answer each question in a separate book.

(c) 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.*

(d) Return all answer books in the folder provided. Additional answer books are available if needed.

SPECIFIC INSTRUCTIONS:

You should answer:

- <u>both</u> questions in the section labeled 760 – MACHINE LEARNING

- <u>two</u> additional questions in another selected section, 7xx, where both questions *must* come from the same section

Hence, you are to answer a total of <u>four</u> 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 Support vector machines

(a) What is the primal form of the constrained optimization problem solved by support vector machines?

(b) What is the dual form of the constrained optimization problem solved by support vector machines?

(c) Given the small dataset below that involves one feature (A) and the Category (-1 for negative and +1 for positive), what will be the output of Platt's Sequential Minimal Optimization (SMO) algorithm on this data? Assume that α_1, α_2, and b are initialized to 0, and show their final values. Use a linear kernel (dot product). Assume C = 10.

	A	*Category*
Ex1	0	+1
Ex2	1	-1

(d) Why can an SVM algorithm learn the exclusive-NOR function (XNOR, or 2-bit even parity) of two input binary features if it uses a quadratic kernel but not if it uses a linear kernel? (You may use any representation for binary values that you like in answering this question.)

760-2 Active learning

Assume you have previously used 1000 labeled training examples to learn a model for a binary classification task using each of the following learning approaches (each learning approach is run independently of the others).

- decision trees
- Bayesian networks
- support-vector machines
- an ensemble of 25 backpropagation-trained neural networks

Of the 1000 training examples, 600 are positive and 400 negative. Based on some cross-validation experiments, you believe each of your learned models will have a future accuracy of about 85%.

(a) You are now given 500 unlabeled examples drawn from the same distribution, and you can have a domain expert label 100 of these. Assume that you will iterate between acquiring one additional training example and then learning a new model. For **each** of the above learning approaches, describe and justify a way to choose the next example to be labeled. Each proposed example-selection strategy should exploit a key aspect of the representation the learning approach uses for the models it learns. Do not use a given strategy more than once (i.e., your answer should involve four reasonably different example-selection strategies).

(b) Now assume that instead of having the domain expert label one example at a time, you must select batches consisting of 20 examples to be labeled in each iteration. Describe one other consideration the active learning system should take into account in addition to the criteria you listed in (a).

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Bayesian optimization

Let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be a function that might not be concave and might not have derivatives. In fact, the only available operation is point evaluation $f(\boldsymbol{x})$ at any $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$. The task is to attempt to fine the *global* maximizer

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbb{R}^D} f(\mathbf{x})$$

with a small number of evaluations. We ask you to use Gaussian process (GP) as the main tool in designing your algorithm to search for the maximizer. Some properties of GP are reviewed below to help you with your design. Specifically, you should answer the following questions with a combination of intuition (in English) and key math formulas.

(a) How can one use GP to represent the uncertainty in the target function *f*, given *n* point evaluations $f(\boldsymbol{x}_1), \ldots, \mathrm{f}(\boldsymbol{x}_n)$?

(b) Given those *n* point evaluations, present one approach to select the next evaluation point $\boldsymbol{x}_{n+1}$. Your approach should consider both exploration and exploitation. Be sure to explain your approach in the context of GP.

(c) Briefly contrast your approach with active learning: Discuss one major similarity and one major dissimilarity.

A review of GP (you might not need all of these properties): A GP is specified by its mean function:

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$$

and covariance function:

$$k(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}\big[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))\big].$$

We write the Gaussian process as

$$f \sim GP(m,k).$$

For any $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, the random vector $f(X)$ follows a Gaussian distribution

$$f(X) \sim \mathcal{N}(m(X),\ K(X,X)),$$

where

$$m(X) = (m(\boldsymbol{x}_1),\ \ldots\ ,m(\boldsymbol{x}_n))^{\mathsf{T}}$$

and $K(X, X)$ is an $n \times n$ kernel matrix with $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Let $\boldsymbol{y}$ and $\boldsymbol{z}$ be jointly Gaussian random vectors

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}, \begin{bmatrix} A & C \\ C^{\mathsf{T}} & B \end{bmatrix}\right),$$

then the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}(\mu_y, A)$$

and the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{z}$ is

$$\boldsymbol{y} \mid \boldsymbol{z} \sim \mathcal{N}(\mu_y + CB^{-1}(\boldsymbol{z} - \mu_z),\ A - CB^{-1}C^{\mathsf{T}})$$

761-2 Bilingual latent topic modeling

Consider two corpora in English and French (or any two languages), respectively. Each English document has a corresponding French document and vice versa. The pair, however, are not necessarily word-by-word translations of each other – i.e., they are not so-called parallel corpora. Instead, they are only loosely comparable. For example, the pair may be news articles describing the same event but written independently.

Latent Dirichlet Allocation (LDA) is a popular model to extract latent topics from a single corpus. A review of LDA is provided below. Your task is to extend LDA so that it can model the paired corpora described above.
Specifically,

(a) Describe the generative process in pseudo code, similar to the LDA review below.
(b) Draw the corresponding graphical model.
(c) Describe one major advantage of your model over LDA.
(d) Describe one major limitation of your model.

Your answer should be more advanced than the following approach: translating the French documents into English and then running standard LDA on the union of the original English documents and the translated documents.

A review of LDA:

1. Sample K multinomial distributions (each of size *V*) $\phi_{1:K}$ from a Dirichlet distribution $Dir(\beta)$; these are the *K* topics. Note β is a parameter vector of length *V*, the vocabulary size.

2. For each document

   (a) Sample a topic multinomial (of size *K*) θ from a Dirichlet distribution $Dir(\alpha)$
   (b) For each word position
      i. Sample topic index $z \sim \theta$
      ii. Sample a word from the topic $w \sim \phi_z$

---

766 – COMPUTER VISION QUESTIONS

766-1 Efficient belief propagation

Belief Propagation is a technique for MRF inference.

(a) Given an image, we model it using a grid graph, in which each pixel is a node and each node is connected to its four nearest neighbors (top, bottom, left, right). Assume the image size is $N \times N$ and each pixel has K states. Please give the main steps of MAX-PRODUCT (or MAX-SUM) and SUM-PRODUCT (or SUM-SUM) versions of Belief Propagation.

(b) Give the computational complexity of the two versions per iteration over the whole image.

(c) The Distance Transform has been proposed to accelerate one version of Belief Propagation. Please state what types of MRF can exploit the Distance Transform and how the Distance Transform works in Belief Propagation.

(d) Describe another competitive method to find a MAP solution of an MRF. Give one important advantage and one important disadvantage of BP compared with this method.

766-2 Active appearance models

Active Appearance Model (AAM) is an extensively used parametric model of visual appearance.

(a) AAMs take texture/appearance and shape into account for building a model. Briefly give the main model building steps. State what your inputs are and what sub-modules or algorithms you will use to obtain the final output.

(b) Justify your particular ordering of steps in (a). Explain whether the ordering you used is important and if the scheme will still work if you randomly shuffle the order.

(c) In the original AAM paper and many later works, the numerical update steps assume that there is a linear relationship between the error image and additive changes to the shape and appearance parameters.

1) Explain why (or why not) this assumption is justified.
2) State one advantage and one limitation of this assumption.

(d) Briefly describe how you would make use of AAM for a project where you want to design a program that recognizes different types of facial expressions. Feel free to use AAMs in conjunction with any other techniques you know of.

769 – ADVANCED NATURAL LANGUAGE PROCESSING QUESTIONS

769-1 Bayesian part-of-speech tagging

Our goal is to tag a corpus of Swedish sentences with parts-of-speech. We are given a Swedish dictionary that lists the possible parts-of-speech for each word, but no annotated text.

(a) Describe a Bayesian HMM for performing unsupervised inference for this task. Your description should include:

- A graphical model representation,
- A brief description of the priors and what effect they will have,
- A brief description of the learning method.

(b) What are the possible advantages of Bayesian inference over training a standard HMM with EM in this scenario?

(c) Now suppose that a small number of gold-standard tagged sentences are made available to you, but you still want to use the much larger number of untagged sentences at your disposal. Discuss a way that you can incorporate the tagged examples into the Bayesian learning method.

769-2 Multilingual dependency parsing

We have a parallel corpus of Faroese and English sentences. Our goal is to parse the Faroese sentences with dependency parses, but we don't have a Faroese parser. Fortunately, we have an English dependency parser to use.

(a) Describe a method for projecting dependency relationships from English word pairs to Faroese. Your method should include a brief description of a word alignment model.

(b) Now suppose that a few of the Faroese sentences have been parsed for us by a linguist. We would like to use this small amount of supervision to refine our projection method. We also realize that the English parser and the word alignment model can both be reconfigured to output *n*-best lists with probabilities. Describe a discriminative reranking approach that uses these probabilities as features.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Modeling uncertainty in sequences

We are often faced with biological sequences for which we have some uncertainty about the characters at subsets of positions and perhaps even the lengths of certain segments of the sequences. For example, current DNA sequencing machines occasionally output an incorrect base, insert extra bases, or delete bases with respect to the true sequences they are reading. Fortunately, sequences are often annotated with uncertainty information, such as the positions at which an error may have occurred during sequencing. When sequence uncertainty is high, it is valuable to model this uncertainty during sequence analyses, such as alignment.

(a) Describe how you can use an HMM to represent uncertainty in a specific DNA sequence. As part of your description, provide an example sequence that has uncertain components and the state transition diagram for an HMM that can represent this uncertain sequence. Your example must include uncertainty for both characters at certain positions as well as the lengths of certain segments within the sequence. You may assume that you are given the probabilities of alternative bases, insertions, or deletions at uncertain positions.

(b) Suppose there are dependencies between the uncertainties at two or more positions within a sequence. Briefly describe how the HMM approach you provide in (a) can take into account such dependencies.

(c) Using the ideas behind profile HMMs, describe an approach to the task of aligning two sequences, **one** of which is uncertain.

776-2 RNA phylogeny

Describe an approach for the intertwined tasks of trying to (i) recover the evolutionary relationships among a given set of orthologous RNA sequences (ii) and infer the secondary structure of each of the sequences. Assume that you don't know anything about the sequences aside from the fact that they are orthologous.

Specifically, you should describe:

- How your approach determines the secondary structure of each sequence. Be specific in describing the approach you will use.
- How your approach will infer a phylogenetic tree relating the sequences.
- How the two tasks will inform each other.

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