SPRING 2013
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, February 4, 2013

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:

- both questions in the section labeled 760 – MACHINE LEARNING
- 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 Dimensionality reduction

Suppose you are given a dataset with a million numerical features and only several thousand labeled examples, and you wish to build a classifier for a binary-valued output.

(a) Describe three algorithms you might apply for *dimensionality reduction*.

(b) For each algorithm, describe a set of conditions under which you expect this dimensionality reduction approach to be better than the other two. Conditions you might take into account include properties of the application domain, the target concept to be learned, the motivation for the learning task, and the learning algorithm(s) you will employ for the classification task. Be sure to justify your answers.

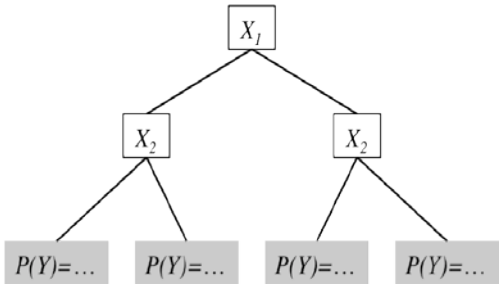
760-2 Decision trees and Bayes networks

Consider a classification task in which we are given n binary features, $X_1 \ldots X_n$, and a binary class variable, Y.

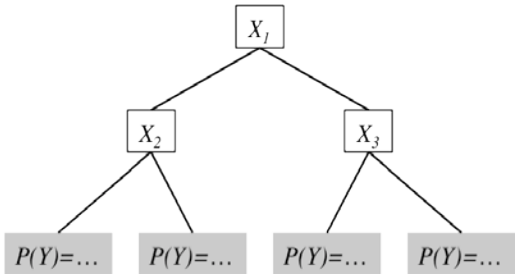
(a) A *probability estimation tree* is a decision tree in which a leaf represents the conditional probability of each class value for instances that reach the leaf. Suppose we have an ID3-like algorithm for learning trees. Describe how you would estimate class conditional probabilities for the leaves using a *maximum likelihood* approach?

(b) Describe an alternative way to estimate the class conditional probabilities at the leaves using a *maximum a posteriori* approach?

(c) Assume one has learned the following probability estimation tree. Is there a compact Bayes network that provides an equivalent representation of $P(Y \mid X_1 \ldots X_n)$? If there is, show the structure of the network. Otherwise, explain why not.



(d) Repeat Part (c), but this time use the tree below, where one of the second-level nodes has been changed from X_2 to X_3.



(e) Suppose that some of the X_i variables are irrelevant for predicting the class variable. Compare and contrast how a decision tree learner ideally should handle irrelevant variables versus how a *naïve Bayes* network ideally should handle them.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Variational representation of Bayes Rule

Let $\pi(\theta)$ be a prior distribution over a model family specified by parameter θ. Let $p(x|\theta)$ be the corresponding likelihood function. Let the data be $x_1, \ldots, x_n$. Consider the following optimization problem, where Δ is the space of all probability distributions over θ, and KL() is the Kullback-Leibler divergence:

$$\min_{q \in \Delta} KL(q \| \pi) - \int q(\theta) \log \left(\prod_{i=1}^{n} p(x_i \mid \theta) \right) d\theta.$$

(a) Prove that the minimizer is the posterior distribution $p(\theta \mid x_1, \ldots, x_n)$.

(b) By constraining the minimization problem over a subset of Δ, one in general arrives at a different solution than the posterior. This can be viewed as a way to incorporate domain knowledge into learning. Consider the example where θ is the head probability of a coin and x is the Bernoulli random variable associated with the coin flip. Discuss the crucial difference between the following two cases:

   i. We add a constraint $E_q[\theta] < 0.5$ to the above optimization problem.

   ii. We do not add any constraints, but use a prior where $\pi(\theta) = 0$ for $\theta \geq 0.5$.

(c) Propose a different constraint that incorporates the spirit of large margin classifiers. Explain its effect on learning. For this question you are not limited to the coin-flip example.

761-2 Growth functions

Let $x \in R$. Consider interval classifiers

$$f(x; a, b, s) = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

Let

$$G = \{f(\cdot; a, b, s) \mid s \in \{-1, 1\}, 0 \leq a \leq b \leq 1\}.$$

Recall that the growth function is the maximum number of ways into which any n points can be classified by the function class G.

(a) What is the VC-dimension of G?

(b) Derive the growth function of G.

766 – COMPUTER VISION QUESTIONS

766-1 3D reconstruction

(a) What is the definition of fundamental matrix F? Given a pair of corresponding points x and x' in two images of the same 3D scene, what is the relationship between F, x and x'?

(b) What is the number of degrees of freedom in F? Explain the reason. Then give a procedure that estimates F given two images. The procedure should be robust to outliers in the feature extraction procedure.

(c) What is geometric meaning of the left null space of F? Similarly, what is the geometric meaning of the right null space of F? Explain using equations.

(d) What is essential matrix E? What is its relationship to F? What is the number of degrees of freedom in E? What geometric quantities do we usually want to recover from E?

(e) Given two images, describe a procedure that computes 3D positions of some salient key points in the images by estimating F and E. You can assume that the intrinsic parameters (e.g., focal length) of the images are known but the extrinsics (e.g., pose) are not.

766-2 Energy minimization in early vision

Let $\{x_1, \ldots, x_n\}$ be a set of {0,1} variables. Consider a class of functions that can be expressed as

$$E(x_1, \ldots, x_n) = \sum_i E_i(x_i) + \sum_{(i,j)} E_{i,j}(x_i, x_j)$$

In other words, a sum of functions of one or at most two binary variables.

(a) Many problems in early vision can be expressed as the minimization of an objective function in the above form where we want to find

$$\arg\min_{\{x_1, \ldots, x_n\}} E(x_1, \ldots, x_n)$$

Briefly describe for any **one** vision problem (e.g., stereo, image restoration, segmentation), how to write your problem in the above form. Explain the role of the two energy terms and the unknown variables for your specific problem.

(b) Describe a situation where it would be appropriate to use a Potts model within the above objective function.

(c) Consider the function $E_{i,j}(x_i, x_j)$ of pairwise variables. Let the notation $E_{i,j}(0,1)$ refer to the penalty associated with the assignment where x_i takes value 0 and x_j takes value 1. Assume each pairwise term satisfies the inequality $E_{i,j}(0,0) + E_{i,j}(1,1) \leq E_{i,j}(0,1) + E_{i,j}(1,0)$. What is the practical significance of this assumption? Include in your answer how this information affects the efficiency of solving the problem.

(d) Consider a generalization of $E_{i,j}$ to $E_{i,j,k}$ which now refers to an assignment to a *triple* of pixels instead of a pair of pixels. Explain if it is possible to find global or local optimal solutions of such minimization problems in vision.

769 – ADVANCED NATURAL LANGUAGE PROCESSING QUESTIONS

769-1 Semantic parsing

Suppose we are given a set of training sentences annotated with lambda calculus *logical forms*. For example:

What states border the state that borders the most states

$\lambda x.state(x) \wedge borders(x, \text{argmax}(\lambda y.state(y), \lambda y.count(\lambda z.state(z) \wedge borders(y, z))))$

Our goal is to induce a semantic parser that can predict the correct logical forms for future sentences.

(a) Describe the Combinatory Categorial Grammar (CCG) formalism used by Zettlemoyer and Collins for combining derivations of logical form with syntactic parsing.

(b) Write down the logical form for the sentence: *What states border Texas*, and show one step of the CCG parse for this sentence.

(c) Describe a probabilistic model that defines $P(L,T \mid S;\theta)$, where L is the logical form, T is the CCG derivation, S is the sentence, and θ is a parameter vector. The model should make use of a feature vector $f(L,T,S)$. Describe the form of the model, the training algorithm, as well as any restrictions imposed on the feature vector.

(d) Describe a non-probabilistic global linear model that can be used in conjunction with your probabilistic model. The model should make use of an unrestricted global feature vector $g(L,T,S)$. Describe an algorithm for training the model.

769-2 Conditional Random Fields (CRFs)

One advantage of the CRF for part-of-speech tagging is that it allows dependent and overlapping emission features. For example, the identity of a word as well as its suffix can be used as features. The suffix feature allows the model to make reasonable predictions even for unseen words.

(a) Describe a CRF model for supervised part-of-speech tagging. Your model should make use of a feature vector $f(w,t)$, where w is a sentence and t is a corresponding sequence of tags. Describe the form of the model, the training algorithm, and any restrictions imposed on the feature vector.

(b) Explain why the CRF cannot be used in the unsupervised learning scenario.

(c) Change the CRF in a small way to turn it into an MRF (M=Markov). Theoretically, the model could now be trained in an unsupervised setting. Explain why this is impractical.

(d) Describe how the technique of Contrastive Estimation due to Smith and Eisner could be used to get around this practical limitation.

776 – ADVANCED BIOINFORMATICS QUESTIONS

776-1 Learning cellular networks from genome-wide expression datasets

Consider the task of learning cellular networks from genome-wide expression datasets.

(a) Given measured expression levels of *n* genes from *p* different conditions, describe how you will infer the hidden network of gene interactions.

(b) Given that *n* is typically in the thousands, reliably inferring such networks is difficult. However we can use some knowledge about the structure of biological networks that enable us to constrain the problem. Describe two structural constraints that you can impose on your learning problem and devise a more efficient algorithm that uses these constraints.

(c) If you were given some information about what are likely edges, how would you use these examples to help your algorithm?

a) Finally, for some of your arrays, you notice that the measurements of some of their genes are missing. How would you use your approach to infer likely values for these genes?

776-2 Whole-genome alignment

(a) Describe an approach for computing an alignment of the genomes from two different species. Assume that the species are eukaryotic, are moderately diverged from one another (at least as diverged as human and mouse), and have large (> 1 billion bases) multi-chromosomal genomes. In your description, explain how each of the techniques listed below plays a role, if any, in your approach. Note that you are not limited to these techniques, nor are you required to use any of them.

- The Needleman-Wunsch global alignment algorithm
- The Smith-Waterman local alignment algorithm
- Profile hidden Markov models
- BLAST
- Suffix trees
- A motif-finding method, such as MEME
- A gene-finding method, such as GENSCAN
- A multiple alignment method, such as CLUSTALW

(b) For each technique listed above that is **not** part of your approach, explain **why** it is either (i) not appropriate for the task, or (ii) appropriate for the task but not used in your approach because its functionality is provided by some other technique.

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