Fall 2011
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN – MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, September 19, 2011

GENERAL INSTRUCTIONS:

(a) This exam has **15** 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:

Answer:

- **both** questions in the section labeled B760 or B766, corresponding to your chosen focus area, ***and***
- any **two** additional questions in the sections Bxxx, where these two questions need *not* come from the same section, ***and***
- **both** questions in the section labeled A760 or A766, again corresponding to your chosen focus area.

Hence, you are to answer a total of **exactly six** questions.

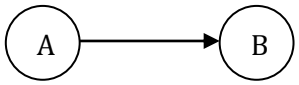
POLICY ON MISPRINTS AND AMBIGUITIES:

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.

731 Advanced Artificial Intelligence: Basic Questions

B731-1. Gibbs Sampling

Consider the following Bayesian Network:



Both A and B are binary random variables. The conditional probability distributions are defined by:

$P(A = 1) = 1/2$
$P(B = 1|A = 1) = p$
$P(B = 0|A = 0) = p$

where parameter $p \in [0,1]$.

(a) You decide to draw samples from the joint distribution $P(A, B)$ using Gibbs sampling. Write down your Gibbs sampling algorithm in pseudocode. Be sure to specify any distributions you use by defining them using **only** numbers and p (if applicable).

(b) How do you estimate $E(A + B)$, the expected value of $A + B$, from your samples?

(c) Discuss **one** major problem of your Gibbs sampler when p is very close to 1.

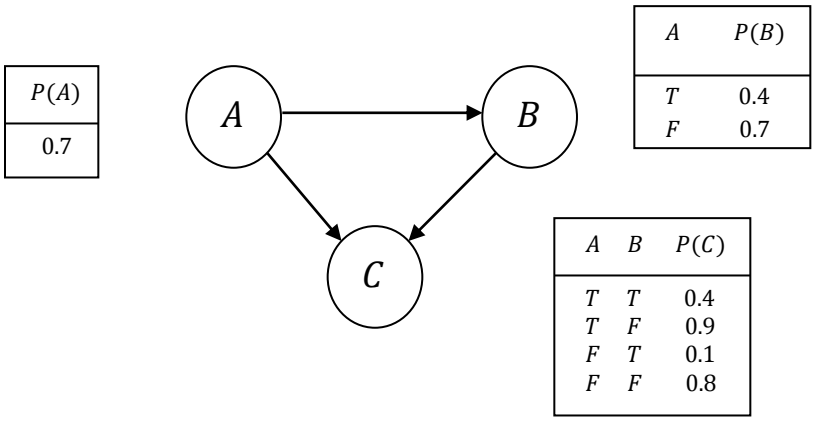
B731-2. Inference in Graphical Models

Suppose you wish to do inference in a graphical model with loops (e.g., a Markov network with cycles or a Bayesian network with cycles in the underlying undirected graph).

(a) Describe **one** situation where you would expect inference by a clique tree (junction tree) to be preferable to other methods.

(b) Describe **one** situation where you would expect belief propagation to be preferable to other inference methods.

(c) Use variable elimination to answer the query $P(A|C = F)$ from the Bayesian network below.



$P(A)$
0.7

A	$P(B)$
T	0.4
F	0.7

A	B	$P(C)$
T	T	0.4
T	F	0.9
F	T	0.1
F	F	0.8

760 Machine Learning: Basic Questions

B760-1. Kernels

Kernels play an important role in machine learning.

(a) Explain what a *kernel* is and discuss **one** key property kernels provide in machine learning.

(b) For **each** of the following approaches to machine-learning: (i) describe how a learning-system designer can use kernels **and** (ii) explain when doing so might be a good idea:

1. Support-vector machines

2. Nearest-neighbor methods

3. Decision-tree induction

4. Reinforcement learning

B760-2. Decision Trees

Decision trees are a successful representation of learned models. For each of the following problem variations, discuss how you would extend Quinlan's standard decision-tree approach:

(a) Feature values are not informative by themselves, but what may matter is how the value of feature *i* compares to the value of feature *j* (for all cases where $i \neq j$).

(b) Rather than returning the most likely class label given an example's features, the induced decision tree should return the conditional probability for each possible class.

(c) Examples are NOT described using *fixed-length feature vectors*. Instead, each example is described by a set of facts, where the number of facts varies from one example to another.

769 Advanced Natural Language Processing: Basic Questions

B769-1. Language Models

Consider a vocabulary with three word types {**<s>**, **A**, **B**} where **<s>** is a special BEGIN-STRING word. All strings we consider start with **<s>**, and that is the only place that **<s>** appears. Also, we only consider strings of length 4, for example, **<s> A A A** or **<s> B A B**.

(a) Define a bigram language model on this vocabulary. List the free parameters needed to fully determine this language model. Hint: P(**<s>** | **A**) would not be a free parameter because we know it is 0, as **<s>** cannot appear after A.

(b) Under your bigram language model, what is the probability of the string **<s> B A B**?

(c) Let's say we don't get to observe the strings. Instead, we see their bag-of-word representations. For example, **<s> B A B** is represented as (**<s>**:1, **A**:1, **B**:2). Under your bigram language model, what is the probability of the bag-of-words (**<s>**:1, **A**:1, **B**:2)? Hint: more than one string maps to this bag-of-words.

(d) Give another possible bag-of-words, and express its probability under your bigram language model.

B769-2. HMM for Part-of-Speech Tagging

Suppose we have trained a bigram hidden Markov model for part-of-speech tagging, and we now wish to predict the most likely part-of-speech tag sequence for a test sentence. In other words, we are given transition parameters θ, emission parameters φ, an input sentence consisting of words $w_1,\ldots,w_n$, and we wish to predict the most likely tag sequence:

$$\text{argmax}_{t_1 \ldots t_n} P(t_1 \ldots t_n \mid w_1 \ldots w_n;\ \theta, \varphi) \qquad (1)$$

For example, if the input test sentence were: "*I love fish*," our model would ideally predict the tag sequence: ***PRONOUN VERB NOUN***.

(a) Draw the directed graphical model structure of the HMM for the example sentence.

(b) Suppose our vocabulary consists of V words and T part-of-speech tags. How many emission parameter values do we have, and how many probability distributions do these values form?

(c) How many transition parameter values do we have, and how many probability distributions do these values form?

(d) Rewrite Equation (1) for the example sentence explicitly in terms of the parameters. (Hint: the parameters should be indexed by the words of the sentence and the three tag variables t_1, t_2, t_3.)

776 Advanced Bioinformatics: Basic Questions

B776-1. Pair HMMs

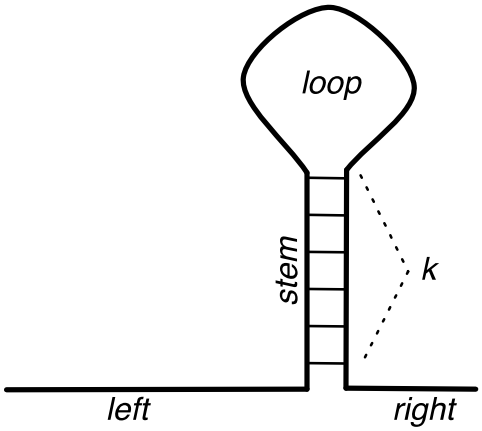
Assume you are given pairs of orthologous proteins from organisms *A* and *B* (i.e. each pair consists of a protein from *A* and a protein from *B* that evolved from a common ancestral protein). Your task is to partition the given set of orthologous pairs into subsets that seem to have diverged in similar ways between the two species. For example, one subset might be characterized by unusual amino-acid substitution frequencies and relatively long insertions/deletions.

Assume that you are given about 1,000 orthologous protein pairs, and you want to partition this set into 10 subsets. Describe how you would solve this problem using pair hidden Markov models.

B776-2. Find the RNA stem loop

Suppose we are interested in a certain class of RNA sequences. Each sequence in this class has the following features:

- It contains a single stem loop preceded by a sequence *left* and followed by a sequence *right.*
- The stem consists of *k* pairs of base-paired nucleotides.
- The lengths of *loop, left* and *right* are all geometrically distributed with a mean length of 10.
- It only contains the nucleotides C and G and has an expected composition of 50% C and 50% G.



(a) Specify a SCFG for modeling this class of RNA sequences with $k = 2$. Give both productions and their probabilities.

(b) Specify an HMM for modeling this class of RNA sequences with $k = 2$. Give a state transition diagram with transition and emission probabilities.

(c) Suppose you are given a set of RNA sequences from this class and you wish to predict the most likely position of the stem loop in each sequence. Each sequence is of length n. Under what conditions (in terms of the relative magnitudes of k and n) would you prefer to use a SCFG over an HMM for this task? **Briefly** justify your answer.

766 Computer Vision: Basic Questions

B766-1. RANSAC

Consider the problem of finding the relation between two images when there may be a large change in viewpoint. Given a set of local feature points detected in each image, assume that the relation between the two images can be modeled as a *2D affine transformation*.

(a) At least how many matching pairs of points, n, are necessary to solve for the 2D affine transformation that relates two input images?

(b) Describe the main steps of the RANSAC algorithm for estimating a good transformation. In your answer use n as described in (a), and assume t is a threshold value for determining when a candidate pair of matching points fits a given 2D transformation model, and d is the number of matching pairs needed to decide that a given model fits well.

(c) If the fraction of feature point matches that are correct is q, what is the probability after h iterations of choosing n candidate matching pairs, that all h model estimates are bad in that they are contaminated by at least one mismatching pair? How can the parameter h be set?

(d) If q is small, i.e., there are a large number of possible mismatches, RANSAC may fail. Why? How could this be improved?

B766-2. Hough Transform

The Hough Transform is a technique for detecting low-level visual features such as straight lines.

(a) Given a set of 2D points where most of the points are on a straight line, explain how to use the Hough Transform to find the line's equation.

(b) Continuing from (a), if most of the points are instead distributed on a *set* of straight lines (not just a single line), how can the Hough Transform be used to find the equations for all these lines?

(c) Consider *two* sets of 2D points. Assume both sets have the same number of points and each point in one set has a unique corresponding point in the other set. Further assume that the two sets are related by an unknown similarity transformation. In the presence of significant point matching errors (outliers), explain how to use the Hough Transform to find the similarity transformation between the two sets.

(d) Describe one potential limitation of using the Hough Transform to detect high dimensional curves or surfaces.

760 Machine Learning: Advanced Questions

A760-1. Naïve Bayes and Logistic Regression

Naïve Bayes (NB) and logistic regression (LR) can be used for very similar purposes. However, a fundamental difference between them is that naïve Bayes is a generative classifier and logistic regression is a discriminative classifier. We say that NB is generative, because it directly models parameters for the joint distribution $P(x, y)$ in terms of $P(x \mid y)$ and $P(y)$ using maximum likelihood. In contrast, we say LR is a discriminative model, as it directly models parameters for $P(y \mid x)$ using maximum conditional likelihood.

Assume both models are trained on the same finite data set with no priors.

(a) Provide high-level pseudocode for using naïve Bayes to generate a sample of N data points.

(b) How would your choice of model vary with the amount of training data available? **Briefly** explain **<u>two</u>** tradeoffs involved in making this decision.

(c) Does logistic regression require the assumption that the input features x_i are conditionally independent of each other given Y? **Briefly** explain.

(d) Assume the features are ***not*** independent in a given data set. Would you expect logistic regression to outperform naïve Bayes? **Briefly** explain.

A760-2. Supervised Learning and Relational Data

Most supervised machine learning algorithms assume a feature-vector representation of data, but much real-world data exists in relational databases consisting of multiple tables. One general approach to this problem is to convert the data from multiple tables into a single table, sometimes called "propositionalization," and then run the standard supervised learning algorithms.

(a) One way to propositionalize the data is to perform a database JOIN operation on the multiple tables. Describe **one** problem with this approach.

(b) Describe a way to use inductive logic programming (ILP) in order to propositionalize the data.

(c) Describe **one** way to run support vector machines (SVMs) and k-nearest neighbor (kNN) on the relational data *without* first propositionalizing it.

766 Computer Vision: Advanced Questions

A766-1. Mean-Shift

The Mean-Shift algorithm is a technique for segmenting images and videos.

(a) Describe the main steps of the Mean-Shift algorithm.

(b) Compare Mean-Shift clustering and Gaussian Mixture Models (GMM). Specifically, describe one situation where one may work better than the other and compare their computational complexity.

(c) Compare Mean-Shift and Normalized Cut in terms of computational complexity. Specifically, if you have multi-core computers and can spawn a large number of threads, discuss how much parallel computation can help to speed up Mean-Shift and Normalized Cut.

(d) Discuss one main difference between using Mean-Shift on a video volume (x, y, t) and using Mean-Shift on volumetric medical data (x, y, z) such as CT scans. Suggest one solution to address this difference when applying Mean-Shift to the volumetric data sets.

A766-2. Pyramid Match Kernel

The Pyramid Match Kernel (PMK) by Grauman and Darrell is extensively used for discriminative learning in image categorization problems. This question deals with various properties of this algorithm.

(a) Assume that each image is represented by unordered sets of features or parts. Consider a setting where we are given a group of images such that (1) their representative sets may *not* have the same size and (2) we have no prior knowledge of feature correspondences across images. **Briefly** describe the PMK procedure and point out if (and how) PMK can still operate in the above situation.

(b) Does PMK afford the ability to perform classification with unsegmented images (with small variations in background and/or occlusion)? If not, why? If yes, discuss why it provides robustness with respect to clutter.

(c) What is the likelihood that a very strong match (between features) is counted more than once (e.g., at different resolutions) when computing the PMK score? **Briefly** explain why.

(d) The PMK method and the histogram intersection function have a number of special cases that relate to well-known distance measures and/or matching algorithms. Discuss any **one**.