Spring 2016
COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, Feb 1, 2016

GENERAL INSTRUCTIONS

1. This exam has 8 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 ROC and PR Curves

(a) Define ROC curves.

(b) Define Precision-Recall (PR) curves.

(c) Describe a set of conditions under which you would prefer to use ROC curves over PR curves.

(d) Name a machine learning application area where PR is regularly used rather than ROC, and discuss why you think PR curves are used in this area. If you can't think of such an area, discuss an area where you think PR *should* be used and why.

760-2 Support vector machines, perceptrons, and *k*-nearest neighbor

(a) Show the primal linear SVM formulation with slack variables.

(b) Show the (not necessarily linear) dual SVM formulation with slack variables. Do not assume any particular kernel; rather, use the notation $K_{i,j}$ for the value of the kernel function between examples i and j.

(c) Discuss one major similarity and one major dissimilarity between linear SVMs and perceptrons.

(d) Discuss one major similarity and one major dissimilarity between SVMs and k-nearest neighbor.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Markov Networks

Suppose you are given the task of developing a system to predict whether there is more than 10cm of snow at each of a set of markers along a mountain trail. At each marker, the system is able to observe a set of discrete variables summarizing the daily temperature, solar intensity, and wind speed. The system is also able to observe whether the daily snow depth at the first marker is more than 10cm or not, but it must infer the binarized snow depth at all other markers.

To train the system, you are given a data set that has the daily observations for all variables, including the snow depth measurements, for each marker over the past winter. Keep in mind that, after training, the system will not be able to observe snow depth measurements at most of the markers.

(a) Describe a Markov network to implement this system.

(b) Describe how the Markov network would be used to estimate the binarized snow depths given the observed variables for each day during the current year.

(c) In addition to the given training data, suppose the system can do active learning. Specifically, assume that each day it can select one marker at which it is able to obtain the measured snow depth. Describe how your system would select the marker to be queried in this way.

761-2 Learning an Interval Classifier

Let the input space be $X = [0, 1]$. Consider a hypothesis space $H = \{h_{a,b} : 0 \le a < b \le 1\}$ where each $h_{a,b}$ is an interval classifier:

$$h_{a,b}(x) = \begin{cases} 1, & x \in [a, b] \\ -1, & \text{otherwise} \end{cases}$$

Let the target concept be $h_{c,d} \in H$. The test distribution $p(x, y)$ is $p(x) = \text{uniform}(X)$ and $p(y = 1 \mid x) = (h_{c,d}(x) + 1)/2$.

(a) What is the error rate of a classifier $h_{a,b}$? Derive it.

(b) Consider a training set with $k + 1$ items on an evenly spaced grid $x_1 = 0, x_2 = 1/k, \ldots, x_{k+1} = 1$, and correct (with respect to the target concept) labels $y_1, \ldots, y_{k+1}$. The version space consists of all classifiers in H that agree with the training set. What value of k can guarantee that the error rate of the worst classifier in the version space is no more than some fixed $\epsilon > 0$ (call this the label complexity)? Derive it.

(c) This question asks you to design a deterministic learning algorithm with the following structure. Initially, the oracle gives your algorithm an arbitrary positive item $x_0 \in [c, d], y_0 = 1$. (Call this a "warm start".) Your algorithm then proceeds in iterations. In iteration $i = 1, 2, \ldots$ it can choose a query $x_i \in X$ based on previous queries $x_0, \ldots, x_{i-1}$ and labels $y_0, \ldots, y_{i-1}$. An oracle answers the correct label y_i. Your algorithm updates the version space accordingly. We ask that your algorithm be deterministic in that it should not use randomness (it should not draw samples). For this question, we ask you to go beyond the grid idea in part (b). Define how your algorithm chooses x_i, and derive its label complexity.

(d) Now assume there is no warm start initially. How would you modify your algorithm? Derive its label complexity.

769 – ADVANCED NATURAL LANGUAGE PROCESSING QUESTIONS

769-1 Probabilistic Sequence Models for Named Entity Recognition

The *named entity recognition* (NER) task involves recognizing instances of named entities (e.g., people, places, organizations) in text strings. Consider an NER task in which we want to recognize names of places (i.e., geographic locations) in English sentences. Suppose we are given a corpus of sentences in which each token has been assigned one of three labels: begin_location, internal_location, other. The begin_location label indicates the first token in a place name, the internal_location label indicates a token that is part of a place name but is not the first token, and the other label is assigned to tokens that are not part of place names.

(a) Describe how you would approach this problem using a hidden Markov model (HMM).

(b) Describe how you would use this trained HMM to recognize named entities in a test set.

(c) Describe one discriminative sequence modeling approach you could use instead of an HMM. Be sure to describe the aspects of the model that would differ from your HMM approach.

(d) Discuss one advantage of the discriminative approach over the HMM approach.

769-2 Solving Arithmetic Word Problems

Design a system that reads an arithmetic word problem (in plain text), and outputs a numerical answer (e.g., 13 for the balloon question below). Here are some example arithmetic word problems:

- Bob gave 9 balloons to his friends. He now has 4 balloons. How many balloons did he have to start with?
- There are 42 walnut trees currently in the park. Park workers will cut down 13 walnut trees that were damaged. How many walnut trees will there be in the park when the workers are finished?
- Liz had 9 black kittens. She gave some of her kittens to Joan. Liz has 5 kittens left. How many kittens did Joan get?

Your answers should include enough technical details to be a plausible road map. You can use standard NLP tools such as POS taggers, named entity recognizers, coreference resolution, parsers, etc.

(a) You will inevitably restrict the space of problems that your system can handle. Clearly state the restrictions and assumptions you make.

(b) How do you internally represent an arithmetic word problem?

(c) How do you convert the input plain text to your internal representation?

(d) How do you compute the numerical answer?

(e) How do you evaluate your system?

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