CSE 455/555 Spring 2012 Mid-Term Exam

Jason J. Corso, jcorso@buffalo.edu

Computer Science and Engineering, SUNY at Buffalo

Date 18 Mar 2012

**Brevity is the soul of wit.
-Shakespeare**

Name:

Nickname:

25	25	25	25	100

Section: 455 or 555

Nickname is a unique identifier only you know. I will try to place grades on my door using this nickname.

Directions – Read Completely

The exam is closed book/notes. You have 75 minutes to complete the exam. Use the provided white paper, write your name on the top of each sheet and number them. Write legibly. Turn in both the question sheet and your answer sheet.

455 and 555: Answer all of the questions. Your exam is out of 100 points.

Problem 1: "Recall" Questions (25pts)

Answer each in one or two sentences **max**.

1. (5pts) What is the Bayes Risk and what does it tell us about making a decision?

2. (5pts) Suppose we have built a classifier on multiple features. What do we do if one of the features is not measurable for a particular case?

3. (5pts) What is the fundamental difference between Maximum Likelihood parameter estimation and Bayesian parameter estimation?

4. (5pts) What is the role of the windowing function φ in the Parzen window density estimator?

5. (5pts) In decision tree learning, when is a node *pure*?

Problem 2: Decision Trees (25pts)

Suppose there is a student that decides whether or not to go in to campus on any given day based on the weather, wakeup time, and whether there is a seminar talk he is interested in attending. There are data collected from 13 days.

Day	Wakeup	HaveTalk	Weather	GoToSchool
D1	Normal	No	Sunny	No
D2	Normal	No	Rain	No
D3	Early	No	Sunny	Yes
D4	Late	No	Sunny	Yes
D5	Late	Yes	Sunny	Yes
D6	Late	Yes	Rain	No
D7	Early	Yes	Rain	Yes
D8	Normal	No	Sunny	No
D9	Normal	Yes	Sunny	Yes
D10	Late	Yes	Sunny	Yes
D11	Normal	Yes	Rain	Yes
D12	Early	No	Rain	Yes
D13	Early	Yes	Sunny	Yes

1. (10pts) Build a decision tree based on these observations, using ID3 algorithm with entropy impurity. Show your work and the resulting tree.

2. (5pts) Classify the following sample using your algorithm: Wakeup=Late, HaveTalk=No,Weather=Rain.

3. (10pts) Consider the case, where some data is missing on a given day. For example, the student lost his/her clock and does not know if he/she woke up early, normal or late. Given an existing tree that has been trained on the full set of variables, what can you do to classify a sample with missing data? For example: Wakeup=???,HaveTalk=Yes,Weather=Sunny. Is this possible? If not explain, why not. If so, explain how and give the answer for that example above using your tree from problem 2.1.

Problem 3: Parametric and Non-Parametric Methods (25pts)

You are given a dataset $\mathcal{D} = \{0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5\}$. Using techniques from parametric and non-parametric density estimation, answer the following questions:

1. (2pts) Draw a histogram of $\mathcal{D}$ with a bin-width of 1 and bins centered at $\{0, 1, 2, 3, 4, 5\}$.

2. (3pts) Write the formula for the kernel density estimate given an arbitrary kernel K.

3. (5pts) In terms of their respective algorithms and their asymptotic performance, compare the Parzen window method and the k-NN method of non-parametric density estimation.

4. (5pts) Select a triangle kernel as your window function:

$$K(u) = (1 - |u|)\delta(|u| \leq 1).$$

where u is a function of the distance of sample x_i to the value in question x divided by the bandwidth: $u = \frac{x - x_i}{h}$. Compute the kernel density estimates for the following values of $x = \{0, 1, 2, 3, 4, 5\}$ bandwidths of 2.

5. (5pts) Now, what if you make an assumption that, rather, the density is a parametric density: it is a Gaussian. Compute the maximum likelihood estimate of the Gaussian's parameters.

6. (5pts) Compare the histogram, the triangle-kernel density estimate, and the maximum-likelihood estimated Gaussian. Which best captures the data? What does each miss? Why would you choose one of another if you were forced to?

Problem 4: General Pattern Classification (25pts)

This question is adapted from Prof. Jain's Homework 1 at MSU this spring. `http://www.cse.msu.edu/~cse802/`

Do not answer in more than one page.

In the first lecture (and Chapter 1 of DHS book), we discussed an example of a pattern classification system to separate sea bass from salmon. Along the same lines, consider the following classification scenario:

A fruit company packages four types of fruits, viz. apples (different kinds, say green, pink, red go to the same box), oranges, grapefruits and grapes. The fruits come to the boxing area on a conveyer belt, and they have to be sorted into their corresponding boxes, based one of these four classes.

For the above problem, briefly explain the following:

1. (5pts) Why would a company be interested in automating these classification problems?

2. (5pts) What kind of sensors and preprocessing can be used?

3. (5pts) What features will be useful for separating the classes involved in these two problems?

4. (10pts) What are the challenges involved in representing the fruits, and how are the above features useful in overcoming some of them?