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:

Section: 455 or 555

25	25	25	25	100

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?

Problem 2

1. Build a decision tree:

The entropy equation is:

$$E = -\sum_j P(w_j) \log P(w_j)$$

According to the data, we have the total labels 9+, 4-. So the entropy for the whole data is:

$$E = -\frac{9}{13} \log \frac{9}{13} - \frac{4}{13} \log \frac{4}{13} = 0.8905$$

If we classify the data with Wakeup:

Early: 4+

$$E = -\frac{4}{4} \log \frac{4}{4} = 0$$

Normal: 2+, 3-

$$E = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} = 0.9710$$

Late: 3+, 1-

$$E = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8113$$

$$Gain(Wakeup) = 0.8905 - \frac{4}{13} \cdot 0 - \frac{5}{13} \cdot 0.9710 - \frac{4}{13} \cdot 0.8113 = 0.2674$$

If we classify the data with HaveTalk:

Yes: 6+, 1-

$$E = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7} = 0.5917$$

No: 3+, 3-

$$E = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$$

$$Gain(HaveTalk) = 0.8905 - \frac{7}{13}\cdot 0.5917 - \frac{6}{13}\cdot 1 = 0.1104$$

If we classify the data with Weather:
Sunny: 6+, 2-

$$E = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.8113$$

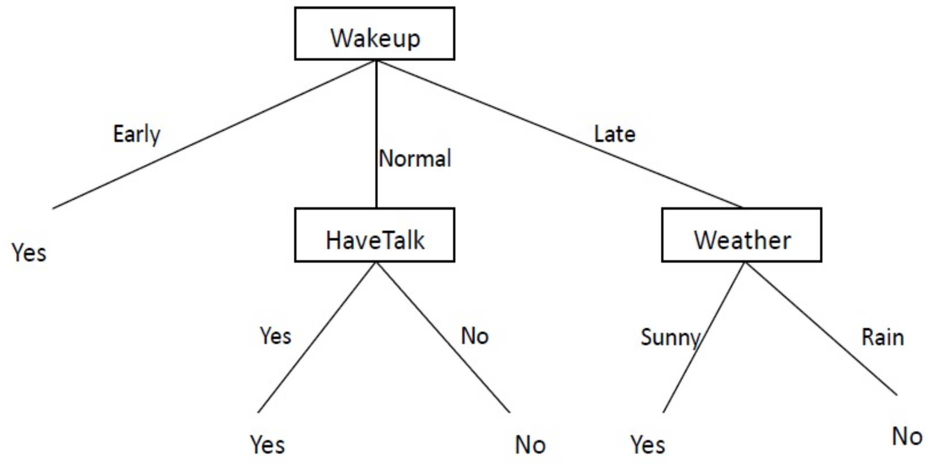
Rain: 3+, 2-

$$E = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.9710$$

$$Gain(Weather) = 0.8905 - \frac{8}{13}\cdot 0.8113 - \frac{5}{13}\cdot 0.9710 = 0.0178$$

Since $Gain(Wakeup)$ is the largest one, we choose Wakeup in this step. The following steps are skiped since they are quite similar to this one.

The final decision tree is:



2. According to the tree learned, the sample should be classified to NO.

3. Yes, the sample can be classified with missing data: ***marginalize***.

If Wakeup = Early, definitely the student will go to school.

If Wakeup = Normal, since HaveTalk = Yes, the student will go to school.

If Wakeup = Late, since Weather = Sunny, the student will go to school.

So for this sample, no matter what the wake up time is, the student will go to school.

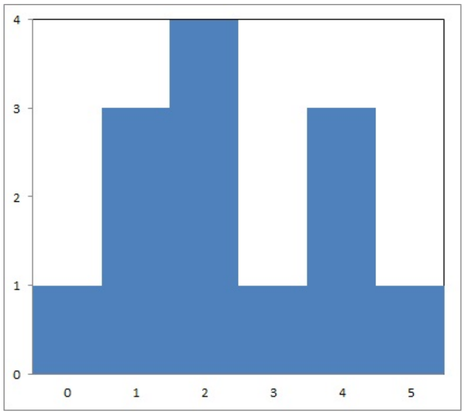


Figure 1: Histogram

1. The histogram is shown in Figure 1.

2. The kernel density function is defined as

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$$

3. Since bandwidth is 2, $V_n = h^d = h^1 = 2$, the kernel function will be

$$K(x-x_i) = (1-|\frac{x-x_i}{h}|)\delta(|x-x_i| \leq 1)$$

   Thus the estimated density for a given x will be

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{x-x_i}{h_n}) = \frac{1}{13}\sum_{i=1}^{n}\frac{1}{2}(1-|\frac{x-x_i}{h}|)\delta(|x-x_i| \leq 1)$$

   with $x_0, x_1, ..., x_n = D = 0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5.$
   Thus, we can get

$$p_n(0) = \frac{5}{52} \approx 0.096; p_n(1) = \frac{11}{52} \approx 0.21; p_n(2) = \frac{12}{52} \approx 0.23; p_n(3) = \frac{9}{52} \approx 0.17; p_n(4) = \frac{8}{52} \approx 0.15; p_n(5) =$$

4. Parzen window specifies the size of the windows as some function of n such as $V_n = 1/\sqrt{n}$, while k-nearest-neighbor specifies the number of samples k_n as some function of n such as $k_n = \sqrt{n}$. Both of them converge to $p(x)$ as $n \to +\infty$. For parzen window method, the choice of V_n has an important effect on the estimated $p_n(x)$: if V_n is too small, the estimation will depends mostly on closer points and will have too much variability based

on a limited number of training samples (over-training); if V_n is too large, the estimation will be an average over a large range of nearby samples, and will loss some details of $p(x)$. By specifying the number of samples, kNN methods circumvent this problem by making the window size a function of the actual training data. If the density is higher around a particular x, the corresponding V_n will be smaller. This means if we have more samples around x, we will use smaller window size to capture more details around x. If the density is lower around x, the corresponding V_n will be larger, which means less samples can only give us an estimation of a larger scale and cannot recover many details.

5. If assume the density is a Gaussian, the maximum likelihood estimate of the Gaussian parameters μ, σ^2 is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 2.38$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \approx 2.08$$

   The unbiased result is

$$\hat{\sigma}^2_{unbias} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \approx 2.26$$

6. Histogram captures the density in a discrete way and can have large errors around boundaries of bins. The triangle-kernel better captured the density with a smoother representation. Although Gaussian estimation is also smooth, it cannot capture the data samples very well. That's because it assumed Gaussian distribution of the samples, while the given sample data doesn't fit into a Gaussian distribution. The triangular-kernel best captured the data and I would choose this one to estimate the distribution of this particular data set.