CSE 455/555 Spring 2011 Final Exam

Jason J. Corso
Computer Science and Engineering
SUNY at Buffalo
jcorso@buffalo.edu
Date 10 May 2011

Brevity is the soul of wit.
-Shakespeare

Name:

Nickname:

Section: 455 or 555

25	25	25	25	25	100

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

Directions – Read Completely

The exam is closed book/notes. You have 180 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 are graded independently but have the same questions. Questions 1 and 5 are required. Then answer two of questions 2, 3, and 4. In total, you will answer 4 of the 5 questions. No extra credit is given for answering all 5 questions. Good luck.

Problem 1: "Recall" Questions (25pts)

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

1. (5pts) What is the basic assumption on the distribution of the data in k-Means?
2. (5pts) What is the key of of the Ugly Duckling Theorem?
3. (5pts) What is the convergence rate of AdaBoost?
4. (5pts) What is an ROC curve?
5. (5pts) What are the four components of a grammar?

Problem 2: Non-Parametric Methods (25pts)

You are given a dataset $\mathcal{D} = \{0, 1, 1, 1, 1, 3, 3, 4, 4, 4, 5\}$. Using techniques from 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. (2pts) 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. (6pts) 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 $x = \{2, 4\}$ bandwidths of 2. Compare the kernel density estimate of these values to the histogram estimate.

5. (10pts) Describe a method for estimating the bias and the variance of the kernel density estimate with the triangle kernel, if you think it is possible. If not, state why you it is not possible to compute estimates of the bias and variance. In any case, do not try to compute them for this exam.

Problem 3: Unsupervised Methods and Clustering (25pts)

1. (3pts) What is the key difference between supervised and unsupervised methods?

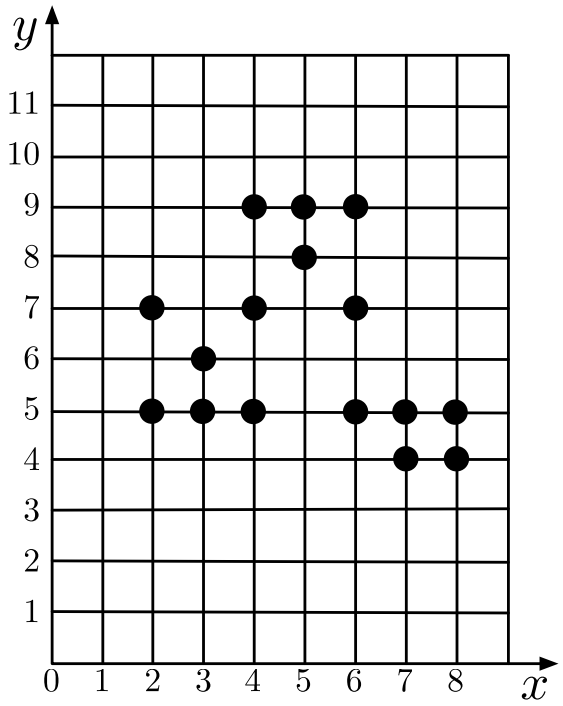
2. (5pts) PCA is an example of an unsupervised method. Now, PCA may be used for classification as well; how? However, what property of PCA makes it generally ill-suited to classification? Name a related linear dimension reduction method that is better-suited to discrimination, in general.

3. (7pts) The k-Means algorithm with Euclidean distances is a very popular and widely used method for data clustering. Assume you have a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. The task is the to find the k cluster centers $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ such that the following objective function is minimized:

$$\mathcal{O}(\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}|\mathcal{D}) = \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \ . \tag{1}$$

Note, this is but one way of writing the objective function.

Consider the following dataset in the Euclidean plane.



For the case of $k = 2$ with initial cluster centers of $\boldsymbol{\mu}_1 = (4, 5)$ and $\boldsymbol{\mu}_2 = (6, 7)$, execute the k-Means algorithms for two steps. Clearly indicate the cluster centers after both steps. Show work.

4. (10pts) This k-Means with Euclidean distances makes a key assumption about the underlying data. Namely, that the clusters are isotropic (e.g., Gaussian with covariance matrices of the form $\sigma^2\mathbf{I}$). However, in many circumstances, we may wish to generalize this underlying assumption. Consider the case in which we relax the isotropic assumption and rather permit a full covariance matrix, $\mathbf{\Sigma}$. Describe how you would do the clustering in this case (8pts)? Be sure to (i) pay attention to the distance function and (ii) clearly describe your algorithm. Finally, apply the algorithm to the data from step 2 (2pts) (for two steps).

Problem 4: Decision Trees (25pts)

1. (3pts) True or False: Decision tree are suitable only for categorical data.

2. (3pts) Consider the entropy impurity function for a node N over c categories:

$$i(N) = -\sum_{j=1}^{c} P(\omega_j) \log P(\omega_j) \ . \tag{2}$$

   As we induce the decision tree, we iteratively select a feature based on how it affects the impurity of the resulting sub-trees. Explain, with mathematics, this process.

3. (6pts) Use the ID3 method with the entropy impurity function (above) to construct a decision tree for the following scenario. Each day I commute from my house in Elmwood Village to the UB North campus. On the return trip, I have three possible routes to choose from: (i) 290W to 190S to 198E to city, call this the *westerly* route, (ii) 290E to 90W to 33W to city, call this the *easterly* route, (iii) 265W to 5W to city, call this the *urban* route. Although there is not very much traffic in Buffalo, certain routes will result in substantially longer driving times for a variety of reasons. For example, the urban route is only 8 miles whereas the other two are 12+ miles, but it has stoplights. Here are some data that I have accumulated using a variety of variables, which follow:

   Sports Event Is there any sporting event happening today? This is a binary feature.

   Time of Day At what time will I be driving home. Consider the following groupings: 1-4,4-6,6-.

   Weather What is the weather? Consider the following: Clear, Rain, Snow, Blizzard

   School in Session Is school in session? This is a binary feature.

   In the following data table, the first word of each feature is used as an indicator:

Index	Route	Sports	Time	Weather	School
1	Easterly	Yes	1-4	Clear	No
2	Urban	Yes	6-	Clear	No
3	Westerly	No	4-6	Snow	Yes
4	Urban	No	6-	Clear	Yes
5	Easterly	Yes	4-6	Rain	Yes
6	Westerly	No	4-6	Rain	Yes
7	Urban	Yes	4-6	Blizzard	Yes
8	Westerly	No	1-4	Rain	No
9	Easterly	Yes	1-4	Snow	Yes
10	Urban	No	1-4	Blizzard	No
11	Westerly	No	4-6	Clear	No
12	Easterly	Yes	4-6	Clear	No
13	Urban	Yes	6-	Rain	No
14	Westerly	No	1-4	Clear	Yes

Using this (limited) data set, construct a three-class decision tree (for the route). Show your calculations, work, and the resulting tree limited to just the first two levels (i.e., the root and the subsequent level).

4. (3pts) In the ID3 case as you explored in step 2, what is the computational complexity for constructing the tree given n data samples?

5. (10pts) Now, consider a *softening* of the decision tree. Rather than a hard classification such as that provided by the decision trees we discussed, our pattern classification scenario (with nominal data) requires that we provide a distribution over possible classes for a given data sample. Explain how you can adapt the standard CART-based methodology to handle this scenario.

Problem 5: General Pattern Classification (25pts)

This required question is of a more general nature than the others above. It is designed to evaluate your ability to use various concepts that have been covered throughout the semester in a critical thinking exercise. There are many "correct" answers to the question, but the goal is to see if you can intelligently construct a solution to a new pattern recognition problem, and explain why it is an appropriate one. Use any and all appropriate tools from the semester.

Use at max 1 sheet of paper and write legibly. Partial credit will be given.

Fresh out of school, you are hired by the prestigious consulting firm PRI (Pattern Recognition International). First day on the job, your boss gives you your first assignment: network intrusion detection. The client, who shall remain nameless to you, is a multinational firm that deals with massive amounts of users accessing data on their network every day. Each access is logged with a variety a variables, including but not limited to, username and user demographics, access IP, access device (computer, smartphone, other machine), access location and a complete record of the session. All counted, thousands of variables are recorded for each session and some are mapped back to databases of users. The firm has uncovered what may be a plot to illegally access vast portions of their network; what could amount to hundreds of millions of dollars in damages. They have provided PRI with full access to previous logs and with a handful of questionable accesses they believe have originated from the thieves. Your job is to build a classifier that can distinguish a network access as a normal pattern from one that is potentially malicious. The method needs to work fast so that the firm can quickly act on its assailant.

Describe a potential approach, in detail. What type of method(s) will you use? Why? How will it be trained? How will it be executed at run time? What are the potential limitations of the method?

Give me a clear picture of (1) the proposed method and (2) that you have learned the material in this course. (You need not answer each of the above questions in detail; rather you need to thoroughly justify your proposed method.)