CSE 555 Spring 2010 Final Exam

Jason J. Corso
Computer Science and Engineering
SUNY at Buffalo
jcorso@buffalo.edu

Friday 30 April 2010, 8:00 - 11:00, Norton 218

Brevity is the soul of wit.
-Shakespeare

There are 6 questions each worth 20pts; ***choose 5 of these 6 questions to answer****. There is no additional credit for answering 6 questions and we will only grade the first 5 answers we see. Hence, the exam is worth 100pts. The questions are of varying difficulty levels; wisely choose which 5 you will answer. Be sure to read the whole exam before attempting any of it. 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.*

Problem 1: "Recall" Questions (20pts)

Answer each in one or two sentences.

1. (4pts) What quantity is Fisher Linear Discriminant maximizing during dimension reduction?
2. (4pts) What is the basic assumption on the distribution of the data in k-Means clustering?
3. (4pts) Describe how cross-validation is used in the training of a general classifier.
4. (4pts) What is the key idea of the Ugly Duckling Theorem?
5. (4pts) What is Bagging?

Solution:

Answers are all directly in the notes.

Problem 2: Dimension Reduction, Manifolds and Classifiers (20pts)

We spent a good amount of time throughout the semester discussing dimension reduction, specifically PCA, in both the class and in homeworks. So, in this question, you need to demonstrate that you've understood and internalized these ideas.

1. (2 pts) What quantity is PCA maximizing during dimension reduction?
2. (2 pts) What are the underlying assumptions on the manifold made by PCA?
3. (4 pts) Describe the PCA representation in terms of the Mahalanobis distance:

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \quad . \tag{1}$$

   Use a figure to help explain the concept.

4. (6 pts) PCA is not intrinsically a classifier. Yet, it is quite often used during classifier construction. Explain how PCA can be used for pattern classification. Use mathematics and give examples.

5. (6 pts) What are pros and cons for nonlinear and linear manifold/subspace learning methods, e.g. LLE and LEA, as opposed to PCA?

Solution:

The nonlinear methods are more effective to visualize the manifold structures, while the linear methods are more applicable to real-world applications. The linearization bridges the manifold representation to subspace learning, which is useful for pattern analysis. The nonlinear methods can not be directly applied to the testing data as they are fully defined on the training data. Any new data come in, the entire training process has to re-run with all the testing data added. Even the out-of-sample extension can be applied, the storage cost may be large for the large-scale case. However the linear methods may not be able to capture all the details of the latent manifold structure. It is a trade-off between sufficient modeling and computing/storage cost.

Problem 3: Bayesian Reasoning (20pts)

Formulate and solve this classical problem using the Bayes rule. There are three criminals A, B, and C waiting in three separate jail cells. One of them will be executed in the next morning when the sun rises. A is very nervous, as he has 1/3 chance to be the one. He tries to get some information from the janitor: "I know you cannot tell me whether I will be executed in the next morning, but can you tell me which of my inmates B and C will not be executed? Because one of them will not be executed anyway, by pointing out who will not be executed, you are not telling me any information." This sounds quite logical. So the Janitor tells A that C won't be executed. At a second thought, A gets much more worried. Before he asked the janitor, he thought he had 1/3 chance, but with C excluded, he seems to have 1/2 chance. A says to himself: "What did I do wrong? Why did I ask the janitor?"

1. (3pts) Formulate the problem using the Bayes rule, i.e. what are the random variables and the input data? What are the meaning of the prior and posterior probabilities in this problem?
2. (3pts) What are the probability values for the prior?
3. (3pts) What are the probability values for the likelihood?
4. (4pts) Calculate the posterior probability (you need to derive the probability values with intermediate steps, not simply showing the final values).
5. (pts) What is the probability of A being executed after he knows that C is excluded?
6. (3pts) Did the janitor tell us any information about A's fate?
7. (3pts) Explain how the Bayes rule helps you.

Problem 4: Decision Trees (20pts)

1. (2pts) 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.

2. (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 198W 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 decision tree. Show your calculations, work, and the resulting tree. Tell me, which route should I take home tonight, Thursday, May 7. I will grade the exams before leaving, so it will be quite late. There is no major sports game tonight.

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

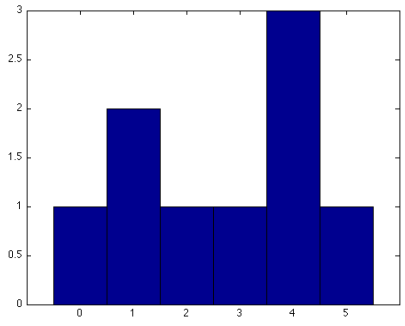
4. (10pts) Now, consider a *softening* of the decision tree. Rather than a hard classification such as that provided by the decision trees we discussed, our pattern classification scenario (with nominal data) requires that we provide a distribution over possible classes for a given data sample. Explain how you can adapt the standard CART-based methodology to handle this scenario.

Problem 5: Non-Parametric Methods (20pts)

You are given a dataset $\mathcal{D} = \{0, 0, 2, 3, 3, 3, 4, 4, 5, 5\}$. 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 $\{0, 1, 2, 3, 4, 5\}$.

   Solution:



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

   Solution:
   Straight from the notes (not defining the notation):

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

3. (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. 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.

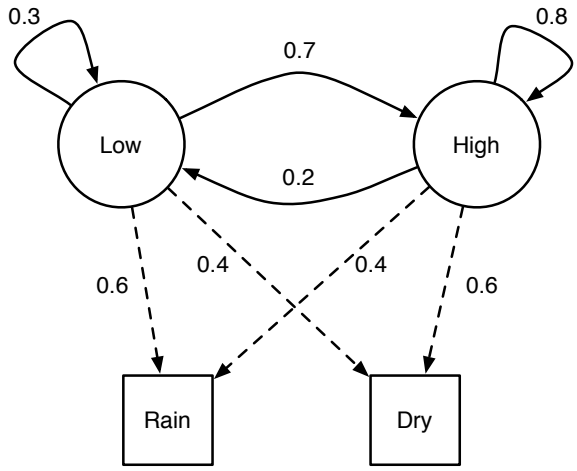
   Solution:

4. (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. If they are possible, then compute them.

   Solution:

Problem 6: Sequential Modeling (20pts)

1. (2pts) What are the three core problems in hidden Markov modeling?
2. Consider the following HMM, which we saw directly in class.



Circles denote hidden states and squares denote observables. Consider a uniform initial distribution.

(a) (2pts) What is the probability of observing the sequence: Rain, Rain, Dry, Rain? Show your trellis.

(b) (2pts) What is the most likely sequence of hidden states for this sequence (Rain, Rain, Dry, Rain)?

(c) (7pts) Now, I travel often and sometimes I don't know what the weather is on any given day. Yet, I still might want to know how well my model fits over a time-period. What is the probability of observing the sequence: Rain, *, Dry, Rain; where the * means I did not observe the weather on that day.

3. (7pts) Now, consider the situation in Problem 4 above where I am selecting which route to use on my drive home from campus in the evening. Problem 4 considered the decision for each day independently. However, certain circumstances, such as the road construction on Rt. 5, would suggest that my day-to-day choices correlate. Describe how you would incorporate a sequential model into this pattern recognition problem. What are the observables? What are the hidden states? How would you do the classification?