CSE 555 Spring 2009 Final Exam

Jason J. Corso
Computer Science and Engineering
University at Buffalo SUNY
jcorso@cse.buffalo.edu
Thursday 7 May 2009, 11:45 AM - 02:45 PM, KNOX 14

Brevity is the soul of wit.
-Shakespeare

The exam is worth 100 points total and has five questions. The questions are of varying difficulty levels and points. 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 (25pts)

Answer each in one or two sentences.

1. (5pts) Describe Bayes Decision Rule.
2. (5pts) What quantity is Fisher Linear Discriminant maximizing during dimension reduction?
3. (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?
4. (5pts) Describe how cross-validation is used in the training of a general classifier.
5. (5pts) What is the key idea of the Ugly Duckling Theorem?

Problem 2: Dimension Reduction, Manifolds and Classifiers (15pts)

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. (5 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.

Problem 3: 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 4: Unsupervised Methods and Clustering (20pts)

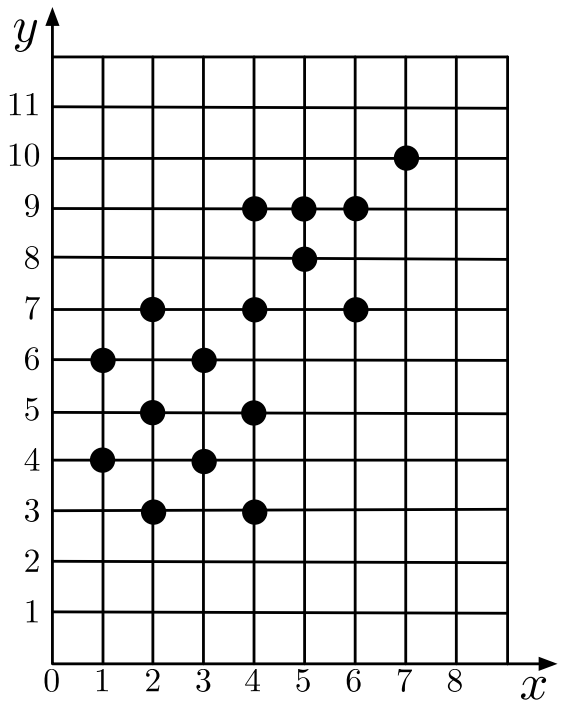
1. (3pts) What is the key difference between supervised and unsupervised methods?

2. (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:

$$\mathcal{O}(\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}|\mathcal{D}) = \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \ . \tag{3}$$

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

   Consider the following dataset in the Euclidean plane.



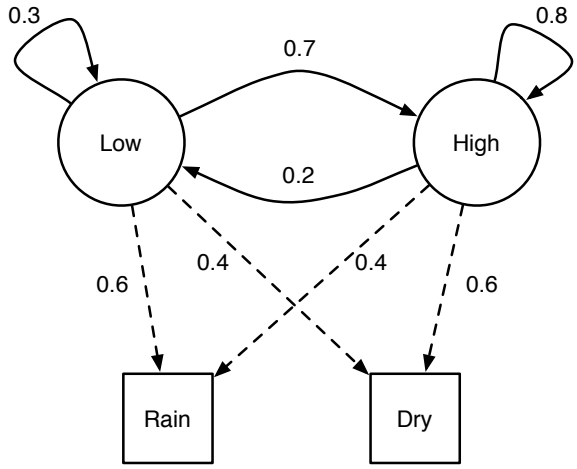
   For the case of $k = 2$ with initial cluster centers of $\boldsymbol{\mu}_1 = (4, 5)$ and $\boldsymbol{\mu}_2 = (6, 7)$, execute the k-Means algorithms for two steps. Clearly indicate the cluster centers after both steps. Show work.

3. (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, $\boldsymbol{\Sigma}$. 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 5: 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?

   (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 3 above where I am selecting which route to use on my drive home from campus in the evening. Problem 3 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?