CSE 455/555 Spring 2012 Final Exam
Jason J. Corso, jcorso@buffalo.edu
Computer Science and Engineering, SUNY at Buffalo
Date 3 May 2012, 11:45 - 14:45
Location Knox 04

Brevity is the soul of wit.
-Shakespeare

Name:

Nickname:

Section: 455 or 555

5	20	20	20	20	20	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 180 minutes to complete the exam. Use the question sheet and 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. There are 5 extra credit points for the 0th question.

Problems 2-5 each have an "easy" and a "hard" part. Be sure to answer those questions you know best, first.

Problem 0: Extra Credit Question (5pts)

Answer in one sentence **in the space provided**.

RA Grace Murray Hopper was a pioneer in computer science. Name one significant contribution she made to the field.

Problem 1: "Recall" Questions (20pts)

Answer each in one or two sentences **in the space provided**.

1. (4pts) What is the key idea of the No Free Lunch Theorem?

2. (4pts) What does LLE focus on preserving during learning?

3. (4pts) What is the basic assumption on the distribution of the data in k-Means clustering?

4. (4pts) What is the convergence rate for the AdaBoost training method?

5. (4pts) What is an ROC curve?

Problem 2: Dimension Reduction (20pts)

This question is about dimensionality, PCA, and it's related methods.

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:

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

   Use a figure to help explain the concept.
4. (4 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. (8pts) Selecting the dimensionality in the reduced PCA basis is a challenge; a variety of methods have been proposed, including those based on the distribution of the eigenvalues, those incorporating model complexity terms, and heuristics. Here, we suggest selecting the dimensionality based on the bias and variance of the ultimate PCA classifier that you designed in the previous questions. Describe a method for estimating the bias and the variance of the classifier you proposed in the previous question. Then describe an algorithm for selection the dimensionality of the reduced PCA basis using the bias and variance ideas.

Problem 3: Bayesian Reasoning (20pts)

Monty Hall Formulate and solve this classifical problem using the Bayes rule. Imagine you are on a gameshow and you're given the choice of three doors: behind one door is a car and behind the other two doors are goats. You have the opportunity to select a door (say No. 1). Then the host, who knows exactly what is behind each door and will not reveal the car, will open a different door (i.e., one that has a goat). The host then asks you if you want to switch your selection to the last remaining door.

1. (2pts) 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 the posterior probabilities in this problem (one sentence each).

2. (2pts) What are the probability values for the prior?

3. (2pts) What are the probability values for the likelihood?

4. (2pts) Derive the posterior probability (include intermediate steps).

5. (2pts) Is it in the contestant's advantage to switch his/her selection? Why?

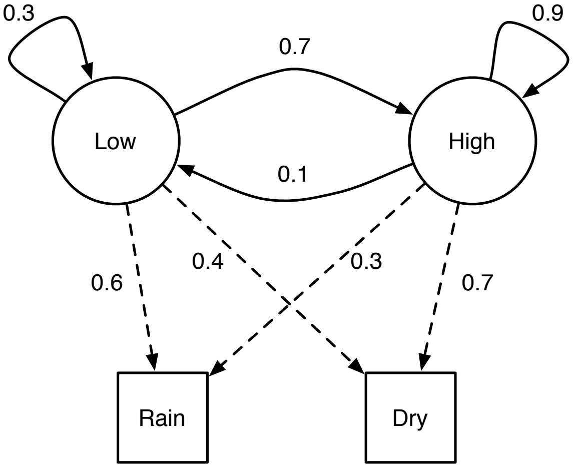
6. (10pts) Now, consider the following twist. The host is having trouble remembering what is behind each of the doors. So, we cannot guarantee that he will not accidentally open the door for the car. We only know that he will not open the door the contestant has opened. Indeed, if he accidentally opens the door for the car, the contestant wins. How does this change the situation? Is it now in the contestant's advantage to switch? Rederive your probabilities to justify your answer.

Problem 4: Sequential Modeling (20pts)

1. (2pts) What is the key difference between Markov models and hidden Markov models?

2. (2pts) What are the three core problems in hidden Markov modeling?

3. Consider the following HMM, which we saw directly in class.



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

   (a) (4pts) What is the probability of observing the sequence: Rain, Rain, Dry? Show your trellis.

   (b) (4pts) What is the most likely sequence of hidden states for this sequence (Rain, Rain, Dry)?

   (c) (8pts) 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; where the * means I did not observe the weather on that day.

Problem 5: Discriminant Functions (20pts)

We covered linear discriminants in significant depth this semester. These questions require you to display you've understood the basic concepts. The last part of the question challenges you to relate MSE training of linear discriminants directly to the Fisher Linear Discriminant.

1. Consider the general linear discriminant function

$$g(x) = \sum_{i=1}^{\hat{d}} a_i \phi_i(\mathbf{x}) \tag{2}$$

   with augmented weight vector $\mathbf{a}$. Let $\mathbf{y} = [\phi_1(\mathbf{x}) \dots \phi_{\hat{d}}(\mathbf{x})]^{\mathsf{T}}$.

(a) (1pts) What are the role of the ϕ functions?

(b) (1pts) Write the equation for the plane in $\mathbf{y}$-space that separates it into two decision regions.

(c) (2pts) A weight vector $\mathbf{a}$ is said to be a solution vector if $\mathbf{a}^\mathsf{T}\mathbf{y}_j > 0 \quad \forall j \in 1, \ldots, n$, assuming the y samples are normalized based on their class label as discussed in class. In general, is this solution vector unique? Why?

2. Consider the relaxation criterion function, let b be a margin,

$$J_r(\mathbf{a}) = \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{Y}} \frac{\left(\mathbf{a}^\mathsf{T}\mathbf{y} - b\right)^2}{\|\mathbf{y}\|^2} \tag{3}$$

(a) (2pts) Compare this relaxation criterion function to the perceptron criterion function.

(b) (2pts) For a single-sample relaxation learning procedure, what is the update rule?

(c) (2pts) What is the geometrical interpretation of the update rule? Draw a figure to help explain.

3. (10pts) Relating the Minimum-Squared Error procedure for training generalized linear discriminants to the Fisher linear discriminant. Recall the MSE procedure for estimating a discriminant. In matrix Y let each row be a data sample, and let $\mathbf{b}$ be a vector of margin values for each point. The MSE procedure turns the inequalities into equalities:

$$Y\mathbf{a} = \mathbf{b} \tag{4}$$

Ultimately, the solution we seek minimizes the sum-of-squared error criterion over all of the samples:

$$J_s(\mathbf{a}) = \sum_{i=1}^{n} (\mathbf{a}^\mathsf{T}\mathbf{y}_i - b_i)^2. \tag{5}$$

Taking the derivative and equating it to 0 gives us the necessary conditions:

$$Y^\mathsf{T}Y\mathbf{a} = Y^\mathsf{T}b. \tag{6}$$

The pseudoinverse is how we solve it.

$$\mathbf{a} = (Y^\mathsf{T}Y)^{-1}Y^\mathsf{T}\mathbf{b}. \tag{7}$$

Consider the following specific augmentation and selection of the margin vector. Assume (1) the augmentation is to simply add a constant 1 to the top of each sample vector, (2) we normalize the class ω_2 samples by multiplying by -1, and (3) the first n_1 samples are labeled ω_1 and the second n_2 samples are labeled ω_2. X_1 is the matrix of class 1 samples x with each row a sample, and X_2 is the matrix for class 2 samples. $\mathbf{1}_i$ is a column vector of n_i ones.

$$Y = \begin{bmatrix} \mathbf{1}_1 & X_1 \\ -\mathbf{1}_2 & -X_2 \end{bmatrix} \tag{8}$$

Let $\mathbf{b}$ be set as the following.

$$\mathbf{b} = \begin{bmatrix} \frac{n}{n_1}\mathbf{1}_1 \\ \frac{n}{n_2}\mathbf{1}_2 \end{bmatrix} \tag{9}$$

Show that for this choice of the margin vector, the MSE solution is equivalent, up to a scale factor, to the Fisher linear discriminant.