CSE 455/555 Spring 2011 Mid-Term Exam

Jason J. Corso
Computer Science and Engineering
SUNY at Buffalo
jcorso@buffalo.edu
Date 11 Mar 2011

Brevity is the soul of wit.
-Shakespeare

Directions – Read Completely

The exam is closed book/notes. You have 50 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: Your exam is out of 67 points. You must answer everything except for 3.3 and 3.4, which are considered extra credit, if you answer them

555: Answer all of the questions. Your exam is out of 85 points.

Problem 1: "Recall" Questions (25pts)

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

1. (5pts) In Bayes Decision Theory, what does the posterior probability capture?
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 quantity is PCA maximizing during dimension reduction?
4. (5pts) What does LLE focus on preserving during learning?
5. (5pts) What is the fundamental difference between Maximum Likelihood parameter estimation and Bayesian parameter estimation?

Problem 2: Discriminant Functions (35pts)

1. Consider the general linear discriminant function

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

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

   (a) (3pts) What are the role of the ϕ functions?
   (b) (3pts) Write the equation for the plane in $\mathbf{y}$-space that separates it into two decision regions.
   (c) (3pts) A weight vector $\mathbf{a}$ is said to be a solution vector if $\mathbf{a}^\mathsf{T}\mathbf{y}_j > 0 \quad \forall j \in 1, \dots, n$, assuming the y samples are normalized based on their class label as discussed in class. In general, is this solution vector unique? Why?
   (d) (3pts) If we were to add a margin b to the discriminant, what role would this margin plan, i.e., $\mathbf{a}^\mathsf{T}\mathbf{y}_j > b \quad \forall j \in 1, \dots, n$.

2. Computing a linear discriminant in the plane. You are given a 4-sample data set of points in the 2D Cartesian plane. The samples, given by $\left(\begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}, \omega_i \right)$, are

$$\left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, +1 \right), \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, +1 \right), \left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, -1 \right) \right\} . \tag{2}$$

(a) (1pt) Plot a class-normalized version of these points on the plane.

(b) (4pts) Using the Batch-Perceptron algorithm compute the discriminant vector a_B. Assume a fixed increment $\eta = 1$ and an initial vector of $\begin{bmatrix}0 & 1\end{bmatrix}^{\mathrm{T}}$. Show all work. Do not normalize the discriminant vector (this will result in messy calculation). Plot a_B.

(c) (4pts) Using the Single-Sample Perceptron algorithm, compute the discriminant vector a_S. Assume a fixed increment $\eta = 1$ and an initial vector of $\begin{bmatrix}0 & 1\end{bmatrix}^{\mathrm{T}}$. Show all work. Indicate which points were selected as "correction" points. Again, do not normalize the vector. Plot a_S.

(d) (4pts) Compare the batch and single sample perceptron algorithms in light of this example.

3. (10pts) Now consider the following dataset,

$$\left\{ \left(\begin{bmatrix}1\\0\end{bmatrix}, +1 \right), \left(\begin{bmatrix}1\\1\end{bmatrix}, -1 \right), \left(\begin{bmatrix}-1\\0\end{bmatrix}, -1 \right), \left(\begin{bmatrix}-1\\1\end{bmatrix}, +1 \right) \right\} . \tag{3}$$

Can the perceptron or relaxation algorithms we discussed converge to a correct solution vector? If so, sketch it. If not, why not?

Problem 3: Expected Loss / Conditional Risk (25pts)

This problem is in the area of Bayesian Decision Theory and, specifically, expected loss / conditional risk.

Recall, the **expected loss** or conditional risk is by definition

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \tag{4}$$

where we have c classes, $\lambda(\alpha_i|\omega_j)$ is the (given) loss associated with classifying a sample as ω_i given it is really ω_j, and the posterior probabilities $P(\omega_j|\mathbf{x})$ are all known.

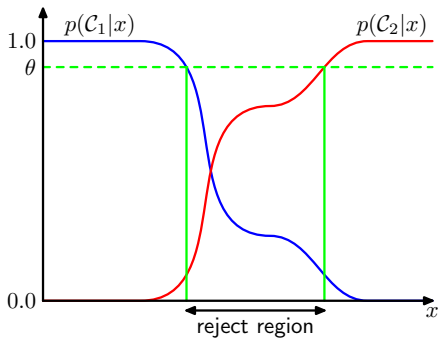
1. (3pts) Bayesian decision theory instructs us to take what action in the context of Equation (4)?

2. (4pts) In what case is the expected loss exactly the minimum probability of error?

Below is required for 555 and extra-credit for 455

Consider the real-world scenario where we have the option of avoiding any decision if we are too unsure about it. For example, if we are designing a computer-aided diagnosis system, we would want to defer the diagnosis to a physician if the X-Ray is of poor quality and we are not very sure.

It has been observed that the highest classification error result from regions of the input space where the largest of the posterior probabilities $P(\omega_j|\mathbf{x})$ is significantly less than unity. So, we can define a threshold θ and decide to avoid making a decision when the maximum posterior is less than θ. In the Bishop PRML book, this has been called *the reject option* and is illustrated on the right.



Assume the loss function as above for cases in which we do not reject and assume a fixed loss of λ_r for cases in which we do reject.

3. (8pts) Write down the decision criterion that will give us the minimum expected loss.

4. Assuming the reject criterion is to reject if the maximum posterior, say $P(\omega|\mathbf{x})$, is less than θ, the fraction of samples that are rejected is clearly controlled by the value of θ.

   (a) (1pt) What is the value of θ to ensure no samples are rejected?

   (b) (1pt) What is the value of θ to ensure all samples are rejected?

   (c) (8pts) What is the relationship between the value of θ and rejection loss λ_r.