CS 479/679 Pattern Recognition
Sample Final Exam

1. **[20 pts]** True/False Questions – To get credit, you **must** give brief reasons.

T F The decision boundary of a two-class classification problem where the data of each class is modeled by a multivariate Gaussian distribution is always linear.

F – when the covariance matrices are not equal (case III), then the decision boundary is quadratic.

T F The Maximum Likelihood estimation solution is always of the assumed parametric form.

T – ML estimates the optimum parameters $\hat{\theta}$ given p(x/θ) (i.e., views the parameters θ as quantities whose values are fixed but unknown).

T F Genetic Algorithms is an example of a heuristic search algorithm for feature selection.

F – GAs is an example of randomized search.

T F Linear Discriminant Analysis (LDA) finds a space of lower dimensionality by choosing the directions where the data varies most.

F – LDA chooses the most discriminative directions

T F The convergence of the EM algorithm is highly dependent on the choice of the learning rate.

F – there is no learning rate in the EM algorithm

2. **[15 pts]** What numerical computational issues arise in practice when using PCA or LDA on data of high dimensionality (e.g., images)? How do we deal with them?

PCA: the covariance matrix AA^T could be very large (i.e., N^2 x N^2), for example, when dealing with images (i.e., N x N). To deal with this issue, we consider the matrix A^TA which is M x M (where M is the number of training images). First, we find the eigenvectors v of A^TA. Then, we find the eigenvectors u of AA^T using u=Av.

LDA: The solution of LDA can be obtained by solving the generalized eigenvalue problem

$$S_b u_k = \lambda_k S_w u_k$$

This problem could be converted to a conventional eigenvalue problem using

$$S_w^{-1} S_b u_k = \lambda_k u_k$$

In practice, S_w is singular due to the high dimensionality of the data. To alleviate this problem, PCA could be applied first:

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \text{--->} PCA \text{--->} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix}$$

LDA is applied next to find the most discriminative directions:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix} \text{--->} LDA \text{--->} \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_{C-1} \end{bmatrix}$$

3. **[15 pts]** (a) What is the EM algorithm?

EM is an iterative ML estimation method. It starts with an initial estimate for θ and refines the current estimate iteratively to increase the likelihood of the observed data:

$$p(D/\theta)$$

(b) What kind of problems is EM best for?

EM works best for problems where the data is incomplete OR can be thought as being incomplete.

(c) Explain the main idea behind applying EM algorithm for estimating the parameters of a Mixture of Gaussians (MoGs).

Assuming that we model the data in a class using a MoGs, the key question is which data was generated by which sub-model; this is the "missing" information for this problem. To use EM, we introduce "hidden" variables z_i for each data x_i

$$y_i=(x_i, z_i)=(x_i, z_{i1}, z_{i2}, \ldots, z_{iK})$$

where z_i is a class indicator vector (hidden variable):

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ was generated by } j-th \text{ component} \\ 0 & \text{otherwise} \end{cases}$$

EM estimates $E[z_i]$ in the expectation step. The parameters of the Gaussians are then updated in the maximization step.

4. **[15 pts]** How would one perform Bayesian classification assuming that classification errors are not all equally important?

Employ a more general error function (i.e., expected "risk") by associating a "cost" (based on a "loss" function) with different errors.

A **loss function** $\lambda(\alpha_i / \omega_j)$ indicates the cost associated with taking action α_i when the correct classification category is ω_j

The **conditional risk** (or expected loss) with taking action α_i is defined as:

$$R(a_i / \mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i / \omega_j) P(\omega_j / \mathbf{x})$$

The Bayes rule minimizes R by:
(i) Computing $R(\alpha_i / x)$ for every α_i given an x
(ii) Choosing the action α_i with the **minimum** $R(\alpha_i / x)$

5. **[15 pts]** (a) What is the criterion being optimized by SVMs?

SVMs perform "structural risk" minimization by (1) minimizing the training error and (2) maximizing the margin of separation.

(b) What is the meaning of the support vectors and why are they important?

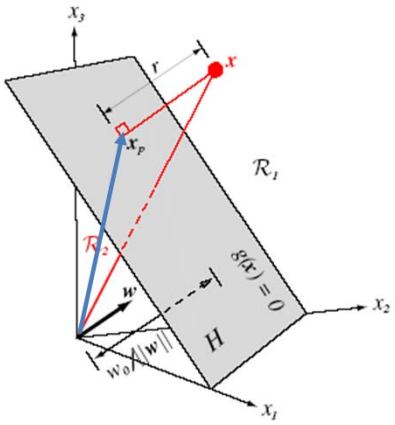
The support vectors define the margin of separation and correspond to the nearest training samples from the decision boundary (i.e., most difficult samples to classify). They are important because the SVM solution (i.e., decision boundary) depends **only** on them.

(c) How do SVMs handle the case of non-linearly separable data?

By mapping the data to a space of very high dimensionality (i.e., using the "kernel" trick) and applying a linear classifier in that space.

6. **[10 pts]** What is the geometric interpretation of the value g(**x**) computed by a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ given some feature vector **x**? Prove it.

g(x) provides an algebraic measure of the **distance** of x from the hyperplane.



$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 = \mathbf{w}^t(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}) + w_0 =$$

$$= \mathbf{w}^t\mathbf{x}_p + r\frac{\mathbf{w}^t\mathbf{w}}{\|\mathbf{w}\|} + w_0 = r\|\mathbf{w}\|$$

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

7. **[10 pts]** Derive the "Perceptron" rule and explain how it works.

The perceptron rule minimizes the following error:

$$J_p(\boldsymbol{\alpha}) = \sum_{\mathbf{y} \in Y(\boldsymbol{\alpha})} (-\boldsymbol{\alpha}^t \mathbf{y})$$

where Y(α) is the set of samples misclassified by α.

Compute the gradient of $J_p(\alpha)$

$$\nabla J_p = \sum_{\mathbf{y} \in Y(\boldsymbol{\alpha})} (-\mathbf{y})$$

Apply gradient descent:

$$\mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k) \nabla J(\mathbf{a}(k))$$

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \eta(k) \sum_{\mathbf{y} \in Y(\boldsymbol{\alpha})} \mathbf{y}$$