Midterm – EE596
14 Feb 2002

Instructions:

• The test is open book/open notes.

• **Show all work.** Partial credit will be given for partial work; no credit will be given for no work.

• Be sure to state all assumptions made and check them when possible.

• There are four problems on eight pages, including the cover.

---

Honor Code:
This exam represents only my own work. I did not give or receive help on this exam.

__________________________  __________________________
Signature                     Date
1. (25 points)
Consider the problem of face recognition. A clever signal processing colleague has implemented feature extraction so that the vector $x$ can be represented using a common covariance. In other words, you have to work on designing a classifier where you can assume each person is characterized by $p(x|\omega_i) \sim N(\mu_i, \Sigma)$. Assume that each person is equally likely, and the classification costs $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ are $\lambda_{ii} = 0$ and $\lambda_{ij} = 1$ for $i \neq j$ and $i, j = 1, \ldots, m$.

(a) First, let's assume that only people who are allowed to use the system have access to it, so there are no imposters.

i. In this case, the Bayes minimum error classification rule can be reformulated as a minimum distance classifier:

$$\omega^* = \arg\min_i d(x, \nu_i).$$

Express the distance function $d(x, \nu_i)$ and face “template” vector $\nu_i$ in terms of the parameters of the Gaussian distributions.

Since $\Sigma_i = \Sigma \forall i$, this corresponds to a minimum Mahalanobis distance classifier with $\nu_i = \mu_i$:

$$\omega^* = \arg\min_i d(x, \mu_i) = \arg\min_i (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$

(Note: $g_i(x) = q_i^T x + a_{i0}$ is the simplest form but not the minimum distance form)

ii. The probability of correct recognition for a particular person $\omega_i$ can be expressed as

$$P_c(i) \geq 1 - \sum_{j \neq i} P_{ij}.$$ 

Give a formula for $P_{ij}$ in terms of the Gaussian parameters and the $Q$ function:

$$P_c(i) = 1 - \sum_{j \neq i} P(D_j|\omega_i) \geq 1 - \sum_{j \neq i} P_{ij}$$

where $P_{ij} = 2\text{ class } Pe$

$$= Q\left(\frac{1}{2} \left( (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) \right) \right)$$
(b) Now let’s refine the model to allow for the possibility of an imposter trying to break in, where $\omega_0$ corresponds to the imposter class. Let $\lambda_{00} = 0, \lambda_{0i} = 1, \text{and } \lambda_{i0} = 10$ for $i \neq 0$. 

i. Give an expression for $R(\alpha_i | x), i \neq 0$ and $R(\alpha_0 | x)$ using the standard definition of risk, i.e. in terms of $p(\omega_i | x)$ and the numerical values for $\lambda_{ij}$.

\[
R(\alpha_i | x) = \sum_{j=1}^{m} \lambda_{ij} p(\omega_j | x) = 10 p(\omega_0 | x) + \sum_{j \neq i}^{m} p(\omega_j | x)
\]

\[
= 10 p(\omega_0 | x) + (1 - p(\omega_i | x)) - p(\omega_0 | x)
\]

\[
= 9 p(\omega_0 | x) + 1 - p(\omega_i | x)
\]

\[
R(\alpha_0 | x) = \sum_{j=0}^{m} \lambda_{0j} p(\omega_j | x) = 1 - p(\omega_0 | x)
\]

ii. Let $g_i(x) = d(x, v_i)$ be the distance function found in part 1(a)i. Assume you are given $p(\omega_0)$ and $p(x | \omega_i) \sim N(\mu_i, \Sigma)$. Show

\[
\arg\min_i R(\alpha_i | x) = \arg\min_i g_i(x)
\]

by finding $g_0(x)$ given $p(x | \omega_0) \sim N(\mu_0, \Sigma)$. In other words, you can add imposter rejection without redesigning the main classifier from part (a) if the imposter distribution has the same form as the speaker distributions.

\[
\arg\min_i R(\alpha_i | x) = \arg\max_i g_i(x)
\]

\[
g_i(x) = \sum_{j=0}^{m} p(\omega_j | x) \left( 10 p(\omega_0 | x) - R(\alpha_i | x) \right)
\]

\[
= \arg\max_i \sum_{j=0}^{m} p(\omega_j | x), 10 p(x | \omega_0) \]

for equal priors $p(\omega_i) = \frac{1}{M}$, \(M \neq 1\)

\[
\mathcal{G}_0 = \arg\max_i \sum_{j=0}^{m} p(\omega_j | x), k p(x | \omega_0) \]

\[
k = \frac{10M p(\omega_0)}{1 - p(\omega_0)}
\]

Taking logs & canceling constants

\[
= \arg\max_i \sum_{j=0}^{m} d(x, \mu_j), d(x, \mu_0) = 2 \log k
\]

Need the additive constant in $g_0(x)$ here.

If we used $g_i(x) = a_i d(x, x_i) + a_{i0}$, the constant would be incorporated in $a_{i0}$
2. (20 points)
Consider the hypothetical ROC curves below in answering the first two parts of this question.

(a) Which of the curves above correspond to a valid ROC curve? Which corresponds to the best classifier, assuming they are designed for the same pattern recognition problem? Which could correspond to a problem where the classes are described by scalar Gaussians with the same variance?

(I) & (II) are valid, (III) is not — need $P_d \geq P_f$

(II) is the best classifier since for any $P_f$ it has highest $P_d$ (area under ROC is greatest)

(I) could correspond to same-variance Gaussians, since it is symmetric about $P_d = 1 - P_f$

(b) On all of the valid ROC curves, draw a star at the point that corresponds to the equal error rate (EER) operating point, i.e., where $P_m = P_f$. In a few words or with an equation, explain why you chose this/these points.

$$P_m = 1 - P_d = P_f \quad \Rightarrow \quad P_d = 1 - P_f$$

(c) Assume zero cost of being correct, i.e. $\lambda_{00} = \lambda_{11} = 0$. Under what conditions is the minimax operating point equal to the EER operating point? Under what conditions is the minimum error solution at the EER operating point?

$$\text{Minimax: } P_d = \frac{d_{01} - d_{00}}{d_{01} - d_{11}} - P_f \left( \frac{d_{10} - d_{00}}{d_{01} - d_{11}} \right) = \frac{1 - \frac{d_{10}}{d_{01}} P_f}{P_f}$$

for $P_d = 1 - P_f$ need $d_{10} = d_{01}$

minimum error for Gaussians with the same variance, need equal priors $p(w_0) = p(w_1)$

more generally, if $P_d = h(P_f)$, need $h'(P_f) = p(w_1)/p(w_0)$
3. (25 points)

The Gamma distribution has two parameters and two sufficient statistics, using the definitions in the Duda, Hart and Stork table on common distributions in the exponential family. Assume you know \( \theta_1 \) and want to estimate \( \theta_2 \).

(a) Find the maximum (ML) likelihood estimate of \( \theta_2 \). Do you need both statistics or is just one sufficient?

\[
g(s, \theta)^n = \frac{\theta_2^{(\theta_1+1)} s_{\theta_1} \cdot e^{-\theta_2 s}}{\Gamma(\theta_1+1)}
\]

\[
\hat{\theta}_2 = \arg\max_{\theta_2} \frac{1}{n} \sum \log g(s, \theta)^n
\]

\[
= \arg\max_{\theta_2} \frac{1}{n} \left( (\theta_1+1) \log \theta_2 + \theta_1 \log s + -\theta_2 s - \log \Gamma(\theta_1+1) \right)
\]

\[
\frac{d}{d\theta_2} \left( \right) = \frac{n(\theta_1+1)}{\theta_2} + n s = 0 \quad \Rightarrow \quad \hat{\theta}_2 = \frac{\theta_1+1}{s}
\]

only \( s_2 = \frac{1}{n} \sum x_i \) is needed.

(b) Find the MAP estimate of \( \theta_2 \) assuming that the prior distribution has an exponential form with parameter \( \alpha \). Show that the MAP estimate is asymptotically equal to the ML estimate.

\[
p(\theta_2) = \alpha e^{-\alpha \theta_2}
\]

\[
\hat{\theta}_2 = \arg\max_{\theta_2} \left( \frac{1}{n} \sum \log g(s, \theta)^n + \log p(\theta_2) \right)
\]

\[
= \arg\max_{\theta_2} \left( (\theta_1+1) \log \theta_2 - n \theta_2 s + \text{const} + \log \alpha - \alpha \theta_2 \right)
\]

\[
\frac{d}{d\theta_2} \left( \right) = \frac{n(\theta_1+1)}{\theta_2} - n s - \alpha = 0
\]

\[
\hat{\theta}_2 = \frac{n(\theta_1+1)}{n s + \alpha} = \frac{\theta_1+1}{s + \alpha/n} \quad \text{as} \quad \frac{s + \alpha/n}{\theta_1+1} \to \frac{\theta_1+1}{\theta_1+1} \quad \Rightarrow \hat{\theta}_2 \to \hat{\theta}_2^	ext{ML}
\]
(c) Could you use this prior and this sufficient statistic with another distribution assumption? If so, give an example. Justify your answer.

\[ S = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ is also sufficient for the exponential at the Gaussian with known variance.} \]

Of these, the exponential makes more sense since it has \( \theta > 0 \) which is consistent with an exponential prior.

(d) Going back to the original Gamma distribution: Could you use a Rayleigh prior for \( \theta_2 \) without changing the sufficient statistic(s)? Explain.

Yes, the prior does not impact the sufficient statistics, though you do want to choose a prior that is consistent with the constraint that \( \theta_2 > 0 \), which is the case for the Rayleigh prior.
4. (30 points)
Four different parameter estimation problems are given below, where in each case you observe i.i.d. random variables or vectors $Y_i$. For each case, state whether you need the EM algorithm to solve the problem and explain why or why not. For cases where you do not need the EM algorithm, find the ML parameter estimate. (No need to rederive the estimate if we've seen it before in lecture, HW or the text.)

(a) Estimate $\lambda$ to describe the distribution of a doubly stochastic process random variable $Y$ that is Poisson with (random) parameter $A$, where $A$ is exponential with unknown parameter $\lambda$.

$$\lambda$$ is the parameter of $p(\lambda)$ & $A$ is a hidden variable, so $\text{EM is needed}$.

$$p(Y|A) = \int p(Y|a) p(a|\lambda) \, da$$

$$= \int A^{-a_d} \exp \left( -\frac{\sum_i y_i}{a} \right) \exp(-a) \, da$$

hard to max with respect to $\lambda$

(b) Let $X_i$ be described by a geometric distribution with unknown parameter $q$, e.g. the number of trials until success in a particular game. We observe:

$$Y_i = \begin{cases} X_i & \text{for } X_i \leq K \\ K & \text{for } X_i > K \end{cases}$$

$$p(X_i = j) = q (1-q)^j$$

Since $P(Y_i \geq K) = \sum_{j=K}^{\infty} q (1-q)^j = \frac{q (1-q)^K}{1-(1-q)}$.

Note: there is another defn. of the geometric for $j=1,2,\ldots$ I used the defn. in the course handout. can be expressed in terms of $q$, you don't need EM

$$\log p(Y|q) = \sum_{i: Y_i < K} \log q + Y_i \log (1-q) + \sum_{i: Y_i = K} \log (1-q)$$

$$\cdot$$

Let $n_k = \# \text{ of } Y_i \text{ that are } = K$

$$\frac{d}{dq} \log p(Y|q) = \frac{n - n_k}{q} + \frac{-\sum_{i: Y_i < K} Y_i}{1-q} + \frac{n_k K}{(1-q)^2} = 0$$

$$\Rightarrow \hat{q} = \frac{n - n_k}{\sum_{i} Y_i + n - n_k}$$
(c) The observation vectors $Y_i$ occasionally have a missing element, and the complete vector
has a Gaussian distribution with unknown mean $\mu$ and covariance $\sigma^2 I$. The unknown
parameters are $\mu$ and $\sigma^2$.

\[ EM \text{ is not needed here because diagonal Gaussian means that the elements } Y_{ij} \text{ are independent, so } \mu_j \text{ and } \sigma^2 \text{ can be estimated by separating the vector elements.} \]

Let $N_j = \# \text{ of } Y_i \text{ where } Y_{ij} \text{ is observed}
\mu_{ij} = \frac{1}{N_j} \sum_{i \cdot Y_{ij} \text{ obs}} Y_{ij}
\hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_d \end{bmatrix}
\sigma^2 = \frac{1}{d \sum_{j=1}^d \sum_{i \cdot Y_{ij} \text{ obs}} (Y_{ij} - \mu_{ij})^2}

(d) The observation $Y_i$ is described by a mixture distribution

\[ p(y) = \sum_{j=1}^m \lambda_j p_j(y) \quad \text{where } p_j(y) \sim N(\mu_j, \sigma^2) \]

and the parameters $\lambda_j$ and $\sigma^2$ are known. The unknown parameters are $\mu_j$ for $j = 1, \ldots, m$.

\[ \text{Even though } \lambda_j \text{ is known, the mixture mode for any particular observation } \]
\[ Y_i \text{ is not known, so you still need } EM \]

End Of Exam