Assignment 3

Saurav Kumar, 12641

March 11, 2015

Problem 1

Suppose the d-sized vector $\mathbf{x}$ in a population with two classes is normmally distributed as: $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma_j), j = 1, 2$ where $\Sigma_j = \sigma_j^2[(1-\rho_j)I + \rho_j \mathbf{1}\mathbf{1}^T]$, where $\mathbf{1}$ is a vector of all 1s and I is the identity matrix. Show that the Bayes discriminant function is given by upto a constant by:

$-\frac{1}{2}(c_{11} - c_{12})b_1 + \frac{1}{2}(c_{21} - c_{22})b_2$, where

$b_1 = (\mathbf{x} - \mu)^T(\mathbf{x} - \mu)$
$b_2 = (\mathbf{1}^T(\mathbf{x} - \mu))^2$
$c_{1j} = [\sigma_j^2(1-\rho_j)]^{-1}$
$c_{2j} = \rho_j[\sigma_j^2(1-\rho_j)(1+(d-1)\rho_j)]^{-1}$

Solution:

Probability Distribution,

$$p(\mathbf{x}|\omega_j) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma_j^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} \tag{0.1}$$

Bayes Discriminant Function,

$$g(\mathbf{x}) = \ln\left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\right) \tag{0.2}$$
$$= -\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma_1^{-1}(\mathbf{x}-\mu) + \frac{1}{2}(\mathbf{x}-\mu)^T \Sigma_2^{-1}(\mathbf{x}-\mu) + constants \tag{0.3}$$

By Ken Miller's identity[1], if A and $A+B$ are invertible, and B has rank 1, then let g=trace(BA^{-1}). Then g≠-1 and

$$(A+B)^{-1} = A^{-1} - \frac{1}{(1+g)}A^{-1}BA^{-1} \tag{0.4}$$

To calculate $g(\mathbf{x})$, we need to calculate inverse of Σ. Given:

$$\Sigma_j = \sigma_j^2[(1-\rho_j)I + \rho_j \mathbf{1}\mathbf{1}^T] \tag{0.5}$$

$$\implies \Sigma_j^{-1} = \frac{1}{\sigma_j^2}[(1-\rho_j)I + \rho_j \mathbf{1}\mathbf{1}^T]^{-1} \tag{0.6}$$

To use Ken Miller's identity, let $A = (1-\rho_j)I$ and $B = \rho_j \mathbf{1}\mathbf{1}^T$. Assuming Σ_j to be invertible, $(A+B)$ is invertible. Since A is constant times identity matrix, A is also invertible. Since all the rows of B are identical, $Rank(B) = 1$.

$$B = \rho_j \mathbf{1}\mathbf{1}^T \tag{0.7}$$

$$A^{-1} = \frac{1}{1-\rho_j}I \tag{0.8}$$

$$BA^{-1} = \frac{\rho_j}{1-\rho_j}\mathbf{1}\mathbf{1}^T \tag{0.9}$$

$$g = trace(BA^{-1}) \tag{0.10}$$

$$= \frac{d.\rho_j}{1-\rho_j} \tag{0.11}$$

Using (0.4),

$$(A+B)^{-1} = \frac{1}{1-\rho_j}I - \frac{1}{1+(\frac{d.\rho_j}{1-\rho_j})}.\frac{1}{1-\rho_j}I.\frac{\rho_j}{1-\rho_j}\mathbf{1}\mathbf{1}^T \tag{0.12}$$

$$\Sigma_j^{-1} = \frac{1}{\sigma_j^2(1-\rho_j)}I - \frac{\rho_j}{\sigma_j^2(1-\rho_j)(1+d\rho_j-\rho_j)}\mathbf{1}\mathbf{1}^T \tag{0.13}$$

Neglecting *constants*, (0.3) becomes

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu)^T(\frac{1}{\sigma_1^2(1-\rho_1)}I - \frac{\rho_1}{\sigma_1^2(1-\rho_1)(1+d\rho_1-\rho_1)}\mathbf{1}\mathbf{1}^T)(\mathbf{x}-\mu) \tag{0.14}$$

$$+\frac{1}{2}(\mathbf{x}-\mu)^T(\frac{1}{\sigma_2^2(1-\rho_2)}I - \frac{\rho_2}{\sigma_2^2(1-\rho_2)(1+d\rho_2-\rho_2)}\mathbf{1}\mathbf{1}^T)(\mathbf{x}-\mu) \tag{0.15}$$

$$\tag{0.16}$$

$$= -\frac{1}{2}(\frac{1}{\sigma_1^2(1-\rho_1)} - \frac{1}{\sigma_2^2(1-\rho_2)})(\mathbf{x}-\mu)^T(\mathbf{x}-\mu) \tag{0.17}$$

$$+\frac{1}{2}(\mathbf{x}-\mu)^T(\frac{\rho_1}{\sigma_1^2(1-\rho_1)(1+d\rho_1-\rho_1)}\mathbf{1}\mathbf{1}^T - \frac{\rho_2}{\sigma_2^2(1-\rho_2)(1+d\rho_2-\rho_2)}\mathbf{1}\mathbf{1}^T)(\mathbf{x}-\mu) \tag{0.18}$$

[1] `http://math.stackexchange.com/questions/17776/`

Substituting b_1, b_2, c_{1j}, c_{2j},

$$g(\mathbf{x}) = -\frac{1}{2}(c_{11} - c_{12})b_1 + \frac{1}{2}(c_{21} - c_{22})[(\mathbf{x} - \mu)^T(\mathbf{1}\mathbf{1}^T)(\mathbf{x} - \mu)] \tag{0.19}$$

Let the column vector $(\mathbf{x} - \mu)$ be $\mathbf{z}$. We have to calculate $\mathbf{z}^T(\mathbf{1}\mathbf{1}^T)\mathbf{z}$.

$$\mathbf{z}^T(\mathbf{1}\mathbf{1}^T) = \Sigma_{i=1}^d z_i \mathbf{1}^T \tag{0.20}$$

$$\begin{aligned} \mathbf{z}^T(\mathbf{1}\mathbf{1}^T)\mathbf{z} &= \Sigma_{i=1}^d z_i \mathbf{1}^T \mathbf{z} && (0.21) \\ &= (\Sigma_{i=1}^d z_i)(\Sigma_{i=1}^d z_i) && (0.22) \\ &= (\mathbf{1}^T \mathbf{z})^2 && (0.23) \\ &= (\mathbf{1}^T(\mathbf{x} - \mu))^2 && (0.24) \\ &= b_2 && (0.25) \end{aligned}$$

Hence, the discriminant function $g(\mathbf{x})$ becomes,

$$g(\mathbf{x}) = -\frac{1}{2}(c_{11} - c_{12})b_1 + \frac{1}{2}(c_{21} - c_{22})b_2 \tag{0.26}$$

Problem 2

Let $p(x)$ be modelled as mixture $p(x) = \Sigma_{j=1}^{g} \pi_j p(x|\mu_j, m_j)$, where $p(x|\mu, m)$ is given by a gamma distribution with mean μ and order parameter m as follows:

$$p(x|\mu, m) = \frac{m}{(m-1)!\mu} (\frac{mx}{\mu})^{m-1} exp(-\frac{mx}{\mu}) \tag{0.27}$$

Derive the EM parameter update equations for π_j, μ_j and m_j.

Solution:

Given a set $\mathcal{L}$ of n independent observations, $\{\mathbf{x}\} = \{\mathbf{x_1}, ..., \mathbf{x_n}\}$, the Likelihood function for the mixture distribution $p(x) = \Sigma_{j=1}^{g} \pi_j p(x|\mu_j, m_j)$ is given by:

$$L(\psi) = \prod_{i=1}^{n} \left(\sum_{j=1}^{g} \pi_j p(x|\mu_j, m_j) \right)$$

where ψ denotes the set of parameters $\{\pi_1, .., \pi_g; (\mu_1, m_1), .., (\mu_g, m_g)\}$.
Our aim is to maximize $L(\psi)$.
Let us augment the component labels z_i to x_i to form y_i, such that $y_i^T = (x_i^T, z_i)$.
So,

$$p(\mathbf{y}|\psi) = p(y_1, .., y_n|\psi)$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{g} [\pi_j p(x_i|\mu_j, m_j)]^{z_{ji}}$$

where z_{ji} are the indictor variables, $z_{ji} = 1$ if the pattern $\mathbf{x_i}$ is in group j, zero otherwise.
Hence,

$$log(p(y_1, .., y_n|\psi)) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ji} log(\pi_j p(x_i|\mu_j, m_j))$$
$$= \sum_{i=1}^{n} z_i^T \mathbf{l} + \sum_{i=1}^{n} z_i^T \mathbf{u}_i(\mu, \mathbf{m})$$

where the vector $\mathbf{l}$ has jth component $\log(\pi_j)$; $\mathbf{u}_i(\mu, \mathbf{m})$ has jth component $\log(p(\mathbf{x}_i|\mu, \mathbf{m}))$; and z_i has components z_{ji}, $j = 1, ..., g$.

E-Step: Q is formed as

$$\begin{aligned} Q(\psi, \psi^{(m)}) &= E[log(p(y_1, .., y_n|\psi))|\{\mathbf{x}\}, \psi^{(m)}] \\ &= \sum_{i=1}^{n} w_i^T \mathbf{l} + \sum_{i=1}^{n} w_i^T \mathbf{u}_i(\mu, \mathbf{m}) \end{aligned}$$

where $w_i = E[z_i|\mathbf{x}_i, \psi^{(m)}]$ with jth component the probability that $\mathbf{x}_i$ belongs to class j given the current paramter estimates, $\psi^{(m)}$.

$$w_{ij} = \frac{\pi_j^{(m)} p\left(\mathbf{x}_i|\mu_j^{(m)}, \mathbf{m}_j^{(m)}\right)}{\sum_k \pi_k^{(m)} p\left(\mathbf{x}_i|\mu_k^{(m)}, \mathbf{m}_k^{(m)}\right)}$$

M-Step:

This consists of maximizing Q with respect to ψ. Consider the parameters $\pi_j, \mu_j, \mathbf{m}_j$ in turn.

- Estimation of π:

  Maximum Q can be obtained by differentiating $Q - \lambda(\Sigma_{j=1}^{g} \pi_j - 1)$ with respect to π_j, where λ is a Lagrange multiplier.

$$\frac{\partial}{\partial \pi_j}\left(Q - \lambda \sum_{j=1}^{g} \pi_j - 1\right) = 0$$

$$\implies \frac{\partial}{\partial \pi_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln(p(x_i|\mu_j, m_j)) + \sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln\pi_j\right) - \lambda \frac{\partial}{\partial \pi_j}\left(\sum_{j=1}^{g} \pi_j - 1\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}.\frac{1}{\pi_j} - \lambda = 0$$

  The constraint $\Sigma_{j=1}^{g} \pi_j = 1$ gives $\lambda = \Sigma_{j=1}^{g} \Sigma_{i=1}^{n} w_{ij} = n$ and we have the estimate of π_j as

$$\hat{\pi}_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$$

- Estimation of μ:

$$\frac{\partial}{\partial \mu_j}\left(Q - \lambda \sum_{j=1}^{g} \pi_j - 1\right) = 0$$

$$\frac{\partial}{\partial \mu_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln(p(x_i|\mu_j, m_j)) + \sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln\pi_j\right) - \lambda \frac{\partial}{\partial \mu_j}\left(\sum_{j=1}^{g} \pi_j - 1\right) = 0$$

Ignoring terms independent of μ_j as their derivative would vanish, we get:

$$\frac{\partial}{\partial \mu_j}\left(\sum_{i=1}^{n} w_{ij} ln(p(x_i|\mu_j, m_j))\right) = 0$$

$$\frac{\partial}{\partial \mu_j}\left(\sum_{i=1}^{n} w_{ij} ln\left(\frac{m_j}{(m_j-1)!\mu_j}(\frac{m_j x_i}{\mu_j})^{m_j-1} exp(-\frac{m_j x_i}{\mu_j})\right)\right) = 0$$

Again, ignoring terms independent of μ_j as their derivative would vanish, we get:

$$\frac{\partial}{\partial \mu_j}\left(\sum_{i=1}^{n} w_{ij} ln\left(\frac{1}{(\mu_j)^{m_j}}.exp(-\frac{m_j x_i}{\mu_j})\right)\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(-\frac{m_j}{\mu_j} + \frac{m_j x_i}{\mu_j^2}\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(x_i - \mu_j\right) = 0$$

$$\implies \hat{\mu}_j^{(m+1)} = \frac{\sum\limits_{i=1}^{n} w_{ij} x_i}{\sum\limits_{i=1}^{n} w_{ij}} = \frac{1}{n\hat{\pi}_j}\sum_{i=1}^{n} w_{ij} x_i \qquad ...(1)$$

- Estimation of m

$$\frac{\partial}{\partial m_j}\left(Q - \lambda \sum_{j=1}^{g} \pi_j - 1\right) = 0$$

$$\frac{\partial}{\partial m_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln(p(x_i|\mu_j, m_j)) + \sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij} ln\pi_j\right) - \lambda\frac{\partial}{\partial m_j}\left(\sum_{j=1}^{g} \pi_j - 1\right) = 0$$

Ignoring terms independent of m_j as their derivative would vanish, we get:

$$\frac{\partial}{\partial m_j}\left(\sum_{i=1}^{n} w_{ij} ln(p(x_i|\mu_j, m_j))\right) = 0$$

$$\implies \frac{\partial}{\partial m_j}\left(\sum_{i=1}^{n} w_{ij} ln\left(\frac{m_j}{(m_j-1)!\mu_j}(\frac{m_j x_i}{\mu_j})^{m_j-1} exp(-\frac{m_j x_i}{\mu_j})\right)\right) = 0$$

$$\implies \frac{\partial}{\partial m_j}\left(\sum_{i=1}^{n} w_{ij}\left(ln(m_j) - ln((m_j-1)!) - ln(\mu_j) + (m_j-1)ln(\frac{m_j x_i}{\mu_j}) - \frac{m_j x_i}{\mu_j}\right)\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(\frac{1}{m_j} - \frac{\partial}{\partial m_j} ln((m_j-1)!) + (m_j-1)(\frac{1}{m_j}) + ln(\frac{m_j x_i}{\mu_j}) - \frac{x_i}{\mu_j}\right) = 0$$

Using Stirling's Approximation,

$$ln(n!) = nln(n) - n + O(ln(n))$$

$$\begin{aligned}\frac{\partial}{\partial n} ln((n-1)!) &\approx \frac{n-1}{n-1} + ln(n-1) - 1\\ &= ln(n-1)\end{aligned}$$

Substituting this in our equation, we get:

$$\sum_{i=1}^{n} w_{ij}\left(\frac{1}{m_j} - ln(m_j-1) + (m_j-1)(\frac{1}{m_j}) + ln(\frac{m_j x_i}{\mu_j}) - \frac{x_i}{\mu_j}\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(-ln(m_j-1) + (m_j-1+1)(\frac{1}{m_j}) + ln(\frac{m_j x_i}{\mu_j}) - \frac{x_i}{\mu_j}\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(-ln(m_j-1) + 1 + ln(\frac{m_j x_i}{\mu_j}) - \frac{x_i}{\mu_j}\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij}\left(1 + ln(\frac{m_j x_i}{\mu_j(m_j-1)}) - \frac{x_i}{\mu_j}\right) = 0$$

To simplify, let us substitute $(\frac{m_j}{\mu_j(m_j-1)})$ as m'_j:

$$\implies \sum_{i=1}^{n} w_{ij}\left(1 + ln(m'_j x_i) - \frac{x_i}{\mu_j}\right) = 0$$

$$\implies \sum_{i=1}^{n} w_{ij} ln(m'_j x_i) = \frac{1}{\mu_j}\sum_{i=1}^{n} w_{ij}x_i - \sum_{i=1}^{n} w_{ij}$$

Substitute $\sum\limits_{i=1}^{n} w_{ij}$ as W_j.
Using (1), substituting estimate of μ_j, we get:

$$\implies \sum_{i=1}^{n} w_{ij} ln(m'_j x_i) = n.\hat{\pi}_j - W_j$$

$$\implies \sum_{i=1}^{n} w_{ij} ln(m'_j) + \sum_{i=1}^{n} w_{ij} ln(x_i) = n.\hat{\pi}_j - W_j$$

$$\implies ln(m'_j) = \frac{n.\hat{\pi}_j - W_j - \sum\limits_{i=1}^{n} w_{ij} ln(x_i)}{\sum\limits_{i=1}^{n} w_{ij}}$$

$$\implies \frac{m_j}{\mu_j(m_j-1)} = exp\left(\frac{n.\hat{\pi}_j - W_j - \sum\limits_{i=1}^{n} w_{ij} ln(x_i)}{\sum\limits_{i=1}^{n} w_{ij}}\right)$$

$$\implies \hat{m_j}^{(m+1)} = \frac{\hat{\mu_j}^{(m+1)}.X}{\hat{\mu_j}^{(m+1)}.X - 1},\ where\ X = exp\left(\frac{n.\hat{\pi}_j^{(m+1)} - \sum\limits_{i=1}^{n} w_{ij} - \sum\limits_{i=1}^{n} w_{ij} ln(x_i)}{\sum\limits_{i=1}^{n} w_{ij}}\right)$$

Problem 3

a. Classification Error Rate on Test Set: 13.6%

b. Classification Error Rate on Test Set (using LDA): 13.8%

c. Classification Error Rate on Test Set (using QDA): 14.0%

Corresponding code files are attached.