MLT - Assignment 3

Ayush Sekhari (12185)

March 11, 2015

1 Part I and II

Part I and II were hand solved and are attached with the report. Please scroll down to view them.

2 Part III

2.1 Part III-a

2.1.1 Generating Data

I used inbuilt runif, rnorm functions to generate the data. 600 points for each class were used. 400 of them were used to train and 100, 100 for validation and test respectively. The metadata of the data is :

1. Training Data : 1200
2. Validation Data : 300
3. Test Data : 300
4. Priors = [1/3,1/3,1/3]

2.1.2 Calculation of the Initial Model

Each class is to be represented as a mixture of three gaussians. I used **K-means clustering** on training data for each class with k = 3. So each class data is clustered into 3 groups. These give 3 cluster means and co-variance matrices for each class. These lead to 9 mean vectors and 9 covariance matrices. I also estimate initial priors for each cluster as the cluster group strength.

2.1.3 Estimation of Co-Variance Matrix

I took the pooled co-variance matrix as the weighted average of the cluster co-variance matrices. The weights used were (priors - 1) in order to ensure unbiased estimator of the co-variance. This is kept constant for the problem.

Approach	CV error	Test-Set error
k-means {training error = 12.333}	16.666	15.666
Stochastic Variation	16.333	15.333
Varying all of the 21 components simultaneously with c = {-0.5,0,0.5}	14	17.333
Varying all of the 21 components simultaneously with c = {-0.1,0,0.1}	15.666	16.333

Table 1: Observed Best errors while searching the space around means using various approaches

2.1.4 Generic Approach:

I have an initial model with 9 mean vectors and a SIGMA matrix. I vary the mean vectors locally using various approaches, recompute the priors each time and then use the mixture model to estimate the error on the validation set. The one with the best is used and the errors are reported on the test set. Performance and search results using various search approaches are listed in the table below.

2.2 Part III-b – LDA and QDA

1. LDA gives the error of 0.136 on the test set
2. QDA gives the error of 0.153 on the test set
3. **LDA performs the best**

Que 1:

$$\Sigma_j = \sigma_j^2(1-\rho_j)\,I + \sigma_j^2\rho_j\,11^T$$

Using $(A+B)^{-1} = A^{-1} - \frac{1}{1+g}A^{-1}BA^{-1}$ where $g = \text{trace}(BA^{-1})$

(as a special case of Woodbury Matrix Inversion Lemma).

[math.stackexchange.com/questions/17776/inverse-of-the-sum-of-matrices].

Here, $A = \sigma_j^2(1-\rho_j)\,I$, $B = \sigma_j^2\rho_j\,11^T$

$$A^{-1} = \sigma_j^{-2}(1-\rho_j)^{-1}\,I, \quad BA^{-1} = \rho_j(1-\rho_j)^{-1}\,11^T \Rightarrow g = \rho_j(1-\rho_j)^{-1}d$$

$$\therefore \Sigma_j^{-1} = (A+B)^{-1} = \sigma_j^{-2}(1-\rho_j)^{-1}\,I - \frac{1 \times \sigma_j^{-2}(1-\rho_j)^{-2}\rho_j\,11^T}{1+\rho_j(1-\rho_j)^{-1}d}$$

$$= C_{1j}\,I - \frac{\rho_j\,11^T}{\sigma_j^2(1-\rho_j)(1+(d-1)\rho_j)} = C_{1j}\,I - C_{2j}\,11^T = \Sigma_j^{-1}$$

Finding the Baeyes Discriminant Function:

Let $\underset{\sim}{x} \in$ Discriminant fn.

$\therefore p_1\,p(x|1) = p_2\,p(x|2)$ [Where p_1, p_2 are priors & p is the class p.d.f.]

$$p_1 \times \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \times \exp\left(\frac{-(\underset{\sim}{x}-\underset{\sim}{u})^T\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{u})}{2}\right) = p_2 \times \frac{1 \times \exp\left(\frac{-(\underset{\sim}{x}-\underset{\sim}{u})^T\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{u})}{2}\right)}{(2\pi)^{d/2}|\Sigma_2|^{1/2}}$$

Taking log on both sides,

$$-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{u})^T\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{u}) = -\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{u})^T\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{u}) + \log c \quad \left(\text{where } c = \frac{p_2}{p_1}\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)^{1/2}\right)$$

$$\therefore -\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{u})^T\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)(\underset{\sim}{x}-\underset{\sim}{u}) - \log c = 0.$$

$$\therefore \frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T(-\Sigma_1^{-1}+\Sigma_2^{-1})(\underset{\sim}{x}-\underset{\sim}{\mu}) \pm \log c = 0$$

(using $\Sigma_i = c_{1i} I - c_{2j} 11^T$)

$$(\underset{\sim}{x}-\underset{\sim}{\mu})^T\left(\frac{1}{2}(c_{21}-c_{22})11^T - \frac{1}{2}(c_{11}-c_{12})I\right)(\underset{\sim}{x}-\underset{\sim}{\mu}) = \log c.$$

$$\Rightarrow -\frac{1}{2}(c_{11}-c_{12})(\underset{\sim}{x}-\underset{\sim}{\mu})^T(\underset{\sim}{x}-\underset{\sim}{\mu}) + \frac{1}{2}(c_{21}-c_{22})(\underset{\sim}{x}-\underset{\sim}{\mu})^T 11^T(\underset{\sim}{x}-\underset{\sim}{\mu}) = \log c$$

$$\Rightarrow -\frac{1}{2}(c_{11}-c_{12})b_1 + \frac{1}{2}(c_{21}-c_{22})\left((1^T(\underset{\sim}{x}-\underset{\sim}{\mu}))^T(1^T(\underset{\sim}{x}-\underset{\sim}{\mu}))\right) = \log c$$

$$\Rightarrow -\frac{1}{2}(c_{11}-c_{12})b_1 + \frac{1}{2}(c_{21}-c_{22})b_2 = \log c$$

$\therefore$ discriminant is given by $-\frac{1}{2}(c_{11}-c_{12})b_1 + \frac{1}{2}(c_{21}-c_{22})b_2$ within a constant (i.e. $\log c$).

Question-2:

$$p(x) = \sum_{j=1}^{g} \pi_j\, p(\underset{\sim}{x}|\mu_j, m_j) \quad \text{where } p(x|\mu_j, m_j) = \frac{m_j}{(m_j-1)!\,\mu_j}\left(\frac{m_j x}{\mu_j}\right)^{m_j-1} \exp\left(-m_j x/\mu_j\right)$$

Let us assume that we know the class to which it belongs.

$\therefore$ Let $y' = (\underset{\sim}{x}, z')$ where z' is the observed latent indicator variable

bit vector with all entries at zero except the Kth entry, which is 1 if x is represented by the cluster K.

$$\therefore p(y) = \prod_{j=1}^{g}\left[p(x|\theta_j)\,\pi_j\right]^{z_j} \quad \left(\text{Since we are sure that it belongs to class given by } z'.\right)$$

$$p(\mathcal{L}') = p(y_1, y_2, y_3, \ldots, y_{|\mathcal{L}|}) = \prod_{i=1}^{|\mathcal{L}|} p(y_i) \quad [\text{using iid assumption}].$$

$$= \prod_{i=1}^{n}\left(\prod_{g=1}^{g}\left[p(x_i|\theta_j)\,\pi_j\right]^{z_{ji}}\right)$$

L = Likelihood = $\prod_{i=1}^{n} \left(\prod_{j=1}^{g} \left[p(x_i|\theta_j)\, \pi_j \right]^{z_{ji}} \right)$

l = log likelihood = $\log L = \sum_{i=1}^{n} \sum_{j=1}^{g} \left(z_{ji} \log p(x_i|\theta_j) + z_{ji} \log \pi_j \right)$

I would want to ~~minimize~~ maximize. l (log likelihood) and thus find the parameters.

z_{ji} = a latent variable, I would remove it and replace it by its expected value

E-Step :

$$w_{ji}^{(m+1)} = E\left(z_{ji} \mid x_i, \Phi^{(m)} \right) = 1 \times P\left(z_{ji} = 1 \mid x_i, \phi^{(m)} \right) + 0 \times P\left(z_{ji} = 0 \mid x_i, \phi^{(m)} \right)$$

$$= P\left(z_{ji} = 1 \mid x_i, \phi^{(m)} \right)$$

$$\boxed{w_{ji}^{(m+1)} = \frac{\pi_j^{(m)}\, p(x_i|\theta_j^{(m)})}{\sum_{k=1}^{g} \pi_k^{(m)}\, p(x_i|\theta_k^{(m)})}}$$

M-Step :

$\hat{l}$ = an estimator of l = $\sum_{i=1}^{n} \sum_{j=1}^{g} \left(w_{ji} \log p(x_i|\theta_j) + z_{ji} \log \pi_j \right)$

I want to ~~minimize~~ maximize the $\hat{l}$ under the Constraint $\sum_{j=1}^{g} \pi_j = 1$

$\therefore \hat{l}' = \hat{l} - \lambda \left(\sum_{j=1}^{g} \pi_j - 1 \right)$ (using the idea of lagrange multipliers)

$\therefore$ Now, we want to maximize $\hat{l}'$ (without any Constraints).

$\therefore \frac{\partial \hat{l}'}{\partial \pi_j} = 0 \quad \forall\, j \in 1..g \quad , \frac{\partial \hat{l}'}{\partial \mu} = 0 \, , \frac{\partial \hat{l}'}{\partial m} = 0 \, .$

$$\frac{\partial \hat{\ell}}{\partial \pi_j} = \sum_{i=1}^{n} \frac{w_{ji}}{\pi_j} - \lambda = 0 \Rightarrow \hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ji}}{\lambda}$$

$$\Rightarrow \sum_{j=1}^{g} \hat{\pi}_j = 1 \Rightarrow \frac{\sum_{i=1}^{n}\sum_{j=1}^{g} w_{ji}}{\lambda} = 1 \Rightarrow \frac{\sum_{i=1}^{n} 1}{\lambda} = 1 \Rightarrow \lambda = n$$

$$\therefore \hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ji}}{n}$$

$$\hat{\ell}' = \sum_{i=1}^{n}\sum_{j=1}^{g} w_{ji} \log \pi_j - \lambda\left(\sum_{j=1}^{g} \pi_j - 1\right) +$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{g} w_{ji} \left(\log m_j - \log(m_j - 1)! - \log y_j + (m_j - 1)\log m_j x_i - (m_j - 1)\log y_j - \frac{m_j x_i}{y_j} \right)$$

$$\frac{\partial \hat{\ell}'}{\partial y_j} = + \sum_{i=1}^{n} \sum_{j=1}^{g} \left(\frac{w_{ji}}{y} \left(-\frac{1}{y}(m-1) + \frac{m x_i}{y^2} \right) \right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{g} \left(w_{ji} \left(\frac{x_i}{y^2} - \frac{1}{y} \right) \right) = 0$$

$$\Rightarrow y_j = \frac{\sum_{i=1}^{n} \sum_{j=1}^{g} w_{ji} x_i}{\sum_{i=1}^{n} \sum_{j=1}^{g} w_{ji}} \Rightarrow \frac{1}{n} \quad \boxed{y = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{g} w_{ji} x_i}$$

$$\hat{u}_j = \sum_{i=1}^{n} w_{ji} x_i \Big/ \sum_{i=1}^{n} w_{ji}$$

$$\frac{\partial \hat{l}'}{\partial m_j} = 0 \quad \Rightarrow \sum_{i=1}^{n} w_{ji} \left[\frac{1}{m_j} - \frac{\partial}{\partial m_j} \log (m_j - 1)! + \log m_j x_i + \frac{m_j - 1}{m_j} - \log \mu_j - \frac{x_i}{\mu_j} \right] = 0$$

$$\Rightarrow \sum_{i=1}^{n} w_{ji} \left[1 + \log m_j + \log x_i - \log \mu_j - \frac{x_i}{\mu_j} - \frac{\partial}{\partial m_j} \log (m_j - 1)! \right] = 0$$

$$\Rightarrow \sum_{i=1}^{n} w_{ji} \left[\log m_j - \frac{\partial}{\partial m_j} \log (m_j - 1)! + \log x_i - \log \mu_j - \frac{x_i}{\mu_j} + 1 \right] = 0$$

Let $\log m_j - \frac{\partial}{\partial m_j} \log (m_j - 1)! = \Psi(m_j)$,

claim: We can estimate m_j from $\Psi(m_j)$ to arbitary accuracy using numerical methods.

$\therefore$ also, $\frac{\partial}{\partial m_j} \log (m_j - 1)!$ = digamma function

{ en.wikipedia.org/wiki/Digamma_function }

$$\Rightarrow \sum_{i=1}^{n} w_{ji} \left[\Psi(m_j) + \log \frac{x_i}{\mu_j} e^{-x_i/\mu_j + 1} \right] = 0 .$$

$$\Psi(m_j) = \frac{-\sum_{i=1}^{n} w_{ji} \log \frac{x_i}{\mu_j} e^{1 - x_i/\mu_j}}{\sum w_{ji}}$$

$\Downarrow$

m_j

∴ The Iteration steps are :

E - Step :

$$\hat{\omega}_{ji}^{(m+1)} = \frac{\pi_j^{(m)} \, p(x_i | \theta_j^{(m)})}{\sum_{k=1}^{g} \pi_k^{(m)} \, p(x_i | \theta_k^{(m)})}$$

M - Step :

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

$$\hat{\mu}_j^{(m+1)} = \sum_{i=1}^{n} \omega_{ji}^{(m+1)} x_i \Big/ \sum_{i=1}^{n} \omega_{ji}^{(m+1)}$$

$$\hat{v}(m_j)^{m+1} = -\frac{\sum_{i=1}^{n} \hat{\omega}_{ji}^{(m+1)} \log \frac{x_i}{\hat{\mu}_j^{(m+1)}} e^{1 - x_i / \hat{\mu}_j^{(m+1)}}}{\sum_{i=1}^{n} \hat{\omega}_{ji}^{(m+1)}}$$