

Mixture Discriminant Analysis

Jia Li

Department of Statistics
The Pennsylvania State University

Email: jjali@stat.psu.edu
<http://www.stat.psu.edu/~jiali>

Mixture Discriminant Analysis

- ▶ A method for classification (supervised) based on mixture models.
- ▶ Extension of linear discriminant analysis
- ▶ The mixture of normals is used to obtain a density estimation for each class.

Linear Discriminant Analysis

- ▶ Suppose we have K classes.
- ▶ Let the training samples be $\{x_1, \dots, x_n\}$ with classes $\{z_1, \dots, z_n\}$, $z_i \in \{1, \dots, K\}$.
- ▶ Each class, with prior probability a_k , is assumed to follow a Gaussian distribution: $\phi(x|\mu_k, \Sigma)$.
- ▶ Model estimation:

$$a_k = \frac{\sum_{i=1}^n I(z_i = k)}{n}$$

$$\mu_k = \frac{\sum_{i=1}^n x_i I(z_i = k)}{\sum_{i=1}^n I(z_i = k)}$$

$$\Sigma = \frac{\sum_{i=1}^n (x_i - \mu_{z_i})(x_i - \mu_{z_i})^t}{n}$$

- ▶ Given a test sample $X = x$, the Bayes classification rule is:

$$\hat{z} = \arg \max_k a_k \phi(x | \mu_k, \Sigma)$$

The decision boundary is linear because Σ is shared by all the classes.

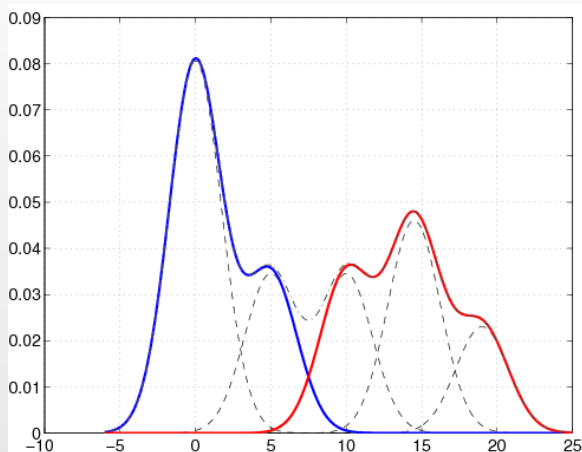
Mixture Discriminant Analysis

- ▶ A single Gaussian to model a class, as in LDA, is too restricted.
- ▶ Extend to a mixture of Gaussians. For class k , the within-class density is:

$$f_k(x) = \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$

- ▶ A common covariance matrix is still assumed.

A 2-classes example. Class 1 is a mixture of 3 normals and class 2 a mixture of 2 normals. The variances for all the normals are 3.0.



Model Estimation

- ▶ The overall model is:

$$P(X = x, Z = k) = a_k f_k(x) = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$

where a_k is the prior probability of class k .

- ▶ The ML estimation of a_k is the proportion of training samples in class k .
- ▶ EM algorithm is used to estimate π_{kr} , μ_{kr} , and Σ .
- ▶ Roughly speaking, we estimate a mixture of normals by EM for each individual class.
- ▶ Σ needs to be estimated by combining all the classes.

► EM iteration:

- E-step: for each class k , collect samples in this class and compute the posterior probabilities of all the R_k components. Suppose sample i is in class k ,

$$p_{i,r} = \frac{\pi_{kr} \phi(x_i | \mu_{kr}, \Sigma)}{\sum_{r'=1}^{R_k} \pi_{kr'} \phi(x_i | \mu_{kr'}, \Sigma)}, \quad r = 1, \dots, R_k$$

- M-step: compute the weighted MLEs for all the parameters.

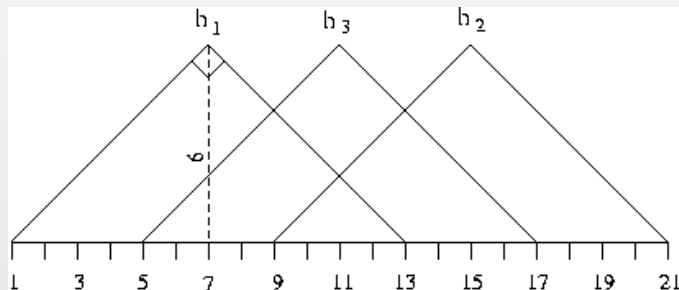
$$\pi_{kr} = \frac{\sum_{i=1}^n I(z_i = k) p_{i,r}}{\sum_{i=1}^n I(z_i = k)}$$

$$\mu_{kr} = \frac{\sum_{i=1}^n x_i I(z_i = k) p_{i,r}}{\sum_{i=1}^n I(z_i = k) p_{i,r}}$$

$$\Sigma = \frac{\sum_{i=1}^n \sum_{r=1}^{R_{z_i}} p_{i,r} (x_i - \mu_{z_i r})(x_i - \mu_{z_i r})^t}{n}$$

Waveform Example

- ▶ Three functions $h_1(\tau)$, $h_2(\tau)$, $h_3(\tau)$ are shifted versions of each other, as shown in the figure.
- ▶ Each h_j is specified by the equal-lateral right triangle function. Its values at integers $\tau = 1 \sim 21$ are measured.



- ▶ The three classes of waveforms are random convex combinations of two of these waveforms plus independent Gaussian noise. Each sample is a 21 dimensional vector containing the values of the random waveforms measured at $\tau = 1, 2, \dots, 21$.
 - ▶ To generate a sample in class 1, a random number u uniformly distributed in $[0, 1]$ and 21 random numbers $\epsilon_1, \epsilon_2, \dots, \epsilon_{21}$ normally distributed with mean zero and variance 1 are generated.

$$x_{\cdot j} = uh_1(j) + (1 - u)h_2(j) + \epsilon_j, \quad j = 1, \dots, 21.$$

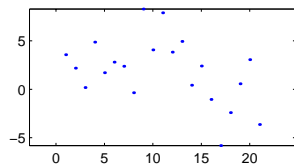
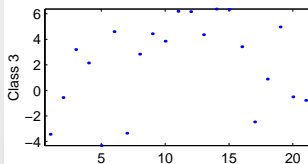
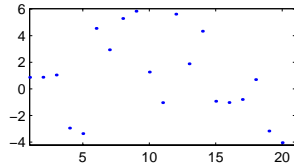
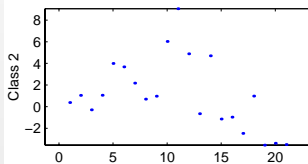
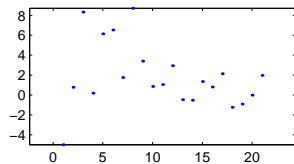
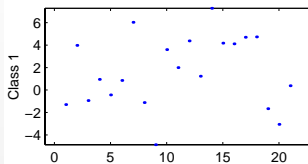
- ▶ To generate a sample in class 2, repeat the above process to generate a random number u and 21 random numbers $\epsilon_1, \dots, \epsilon_{21}$ and set

$$x_{\cdot j} = uh_1(j) + (1 - u)h_3(j) + \epsilon_j, \quad j = 1, \dots, 21.$$

- ▶ Class 3 vectors are generated by

$$x_{\cdot j} = uh_2(j) + (1 - u)h_3(j) + \epsilon_j, \quad j = 1, \dots, 21.$$

Example random waveforms



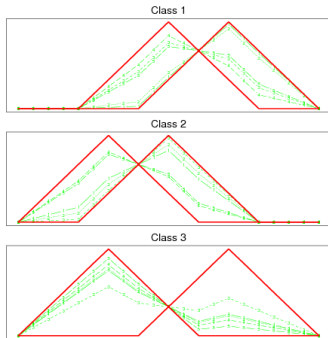


Figure 12.12: *Some examples of the waveforms generated from model (12.62) before the Gaussian noise is added.*

- ▶ A three component mixture of normals is assumed for each class.
- ▶ The Bayes risk has been estimated to be about 0.14.
- ▶ MDA outperforms LDA, QDA, and CART.
 - ▶ Training data size: 300. Test data size: 500. Ten simulations are performed.
 - ▶ Error rates for MDA (3 components per class) and other methods are compared below.

Method	Training	Test
LDA	0.121(0.006)	0.191(0.006)
QDA	0.039(0.004)	0.205(0.006)
CART	0.072(0.003)	0.289(0.004)
MDA	0.087(0.005)	0.169(0.006)

- ▶ Low dimension views are obtained from projecting on to canonical coordinates.

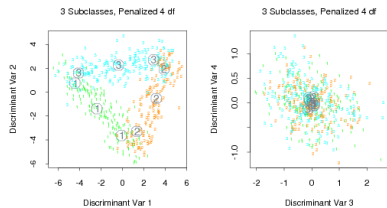


Figure 12.13: *Some two-dimensional views of the MDA model fitted to a sample of the waveform model. The points are independent test data, projected on to the leading two canonical coordinates (left panel), and the third and fourth (right panel). The subclass centers are indicated.*