Linear Discriminant Analysis

Jia Li

Department of Statistics
The Pennsylvania State University

Email: jiali@stat.psu.edu
http://www.stat.psu.edu/~jiali
Notation

- The prior probability of class $k$ is $\pi_k$, $\sum_{k=1}^{K} \pi_k = 1$.
  - $\pi_k$ is usually estimated simply by empirical frequencies of the training set
    $$\hat{\pi}_k = \frac{\# \text{ samples in class } k}{\text{Total } \# \text{ of samples}}$$

- The class-conditional density of $X$ in class $G = k$ is $f_k(x)$.

- Compute the posterior probability
  $$Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

- By MAP (the Bayes rule for 0-1 loss)
  $$\hat{G}(x) = \arg \max_k Pr(G = k \mid X = x)$$
  $$= \arg \max_k f_k(x)\pi_k$$
Class Density Estimation

- Linear and quadratic discriminant analysis: Gaussian densities.
- Mixtures of Gaussians.
- General nonparametric density estimates.
- Naive Bayes: assume each of the class densities are products of marginal densities, that is, all the variables are independent.
Linear Discriminant Analysis

- Multivariate Gaussian:

\[ f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \]

- Linear discriminant analysis (LDA): \( \Sigma_k = \Sigma, \forall k \).

- The Gaussian distributions are shifted versions of each other.
Linear Discriminant Analysis

Jia Li  http://www.stat.psu.edu/~jiali
Optimal classification

\[ \hat{G}(x) = \arg \max_k \Pr(G = k \mid X = x) \]

\[ = \arg \max_k f_k(x)\pi_k = \arg \max_k \log(f_k(x)\pi_k) \]

\[ = \arg \max_k \left[ -\log((2\pi)^{p/2}|\Sigma|^{1/2}) \right. \]

\[ - \frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k) + \log(\pi_k) \left. \right] \]

\[ = \arg \max_k \left[ -\frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k) + \log(\pi_k) \right] \]

Note

\[ -\frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x \]
To sum up

$$\hat{G}(x) = \arg\max_k \left[ x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right]$$

- Define the *linear discriminant function*

  $$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k).$$

  Then

  $$\hat{G}(x) = \arg\max_k \delta_k(x).$$

- The decision boundary between class $k$ and $l$ is:

  $$\{x : \delta_k(x) = \delta_l(x)\}.$$

Or equivalently the following holds

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) = 0.$$
Linear Discriminant Analysis

Binary classification \((k = 1, l = 2)\):

- Define \(a_0 = \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)\).
- Define \((a_1, a_2, \ldots, a_p)^T = \Sigma^{-1}(\mu_1 - \mu_2)\).
- Classify to class 1 if \(a_0 + \sum_{j=1}^p a_j x_j > 0\); to class 2 otherwise.
- An example:
  - \(\pi_1 = \pi_2 = 0.5\).
  - \(\mu_1 = (0, 0)^T, \mu_2 = (2, -2)^T\).
  - \(\Sigma = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 0.5625 \end{pmatrix}\).
  - Decision boundary:
    \[
    5.56 - 2.00x_1 + 3.56x_2 = 0.0.
    \]
Linear Discriminant Analysis
In practice, we need to estimate the Gaussian distribution.

\[ \hat{\pi}_k = \frac{N_k}{N}, \text{ where } N_k \text{ is the number of class-}k \text{ samples.} \]

\[ \hat{\mu}_k = \frac{\sum_{g_i=k} x^{(i)}}{N_k}. \]

\[ \hat{\Sigma} = \frac{\sum_{k=1}^{K} \sum_{g_i=k} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T}{(N - K)}. \]

Note that \( x^{(i)} \) denotes the \( i \)th sample vector.
Diabetes Data Set

- Two input variables computed from the principal components of the original 8 variables.
- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1.6634 \end{pmatrix}$

Classification rule:

$$\hat{G}(x) = \begin{cases} 1 & 0.7748 - 0.6771x_1 - 0.3929x_2 \geq 0 \\ 2 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & 1.1443 - x_1 - 0.5802x_2 \geq 0 \\ 2 & \text{otherwise} \end{cases}$$
The scatter plot follows. Without diabetes: stars (class 1), with diabetes: circles (class 2). Solid line: classification boundary obtained by LDA. Dash dot line: boundary obtained by linear regression of indicator matrix.
Within training data classification error rate: 28.26%.
Sensitivity: 45.90%.
Specificity: 85.60%.
Contour plot for the density (mixture of two Gaussians) of the diabetes data.
Simulated Examples

- LDA is not necessarily bad when the assumptions about the density functions are violated.
- In certain cases, LDA may yield poor results.
LDA applied to simulated data sets. Left: The true within class densities are Gaussian with identical covariance matrices across classes. Right: The true within class densities are mixtures of two Gaussians.
Linear Discriminant Analysis

Left: Decision boundaries by LDA. Right: Decision boundaries obtained by modeling each class by a mixture of two Gaussians.

Jia Li  http://www.stat.psu.edu/~jiali
Quadratic Discriminant Analysis (QDA)

- Estimate the covariance matrix $\Sigma_k$ separately for each class $k$, $k = 1, 2, ..., K$.

- Quadratic discriminant function:

  $$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k.$$ 

- Classification rule:

  $$\hat{G}(x) = \arg \max_k \delta_k(x).$$

- Decision boundaries are quadratic equations in $x$.

- QDA fits the data better than LDA, but has more parameters to estimate.
Diabetes Data Set

- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma}_1 = \begin{pmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{pmatrix}$
- $\hat{\Sigma}_2 = \begin{pmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{pmatrix}$
Within training data classification error rate: 29.04%.
Sensitivity: 45.90%.
Specificity: 84.40%.
Sensitivity is the same as that obtained by LDA, but specificity is slightly lower.
LDA on Expanded Basis

► Expand input space to include $X_1 X_2$, $X_1^2$, and $X_2^2$.
► Input is five dimensional: $X = (X_1, X_2, X_1 X_2, X_1^2, X_2^2)$.

\[
\hat{\mu}_1 = \begin{pmatrix} -0.4035 \\ -0.1935 \\ 0.0321 \\ 1.8363 \\ 1.6306 \end{pmatrix}, \quad \hat{\mu}_2 = \begin{pmatrix} 0.7528 \\ 0.3611 \\ -0.0599 \\ 2.5680 \\ 1.9124 \end{pmatrix}
\]

\[
\hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 & -0.6254 & 0.3548 & 0.5215 \\ -0.1461 & 1.6634 & 0.6073 & -0.7421 & 1.2193 \\ -0.6254 & 0.6073 & 3.5751 & -1.1118 & -0.5044 \\ 0.3548 & -0.7421 & -1.1118 & 12.3355 & -0.0957 \\ 0.5215 & 1.2193 & -0.5044 & -0.0957 & 4.4650 \end{pmatrix}
\]
Classification boundary:

\[0.651 - 0.728x_1 - 0.552x_2 - 0.006x_1x_2 - 0.071x_1^2 + 0.170x_2^2 = 0\].

If the linear function on the right hand side is non-negative, classify as 1; otherwise 2.
Classification boundaries obtained by LDA using the expanded input space $X_1$, $X_2$, $X_1X_2$, $X_1^2$, $X_2^2$. Boundaries obtained by LDA and QDA using the original input are shown for comparison.
Within training data classification error rate: 26.82%.
Sensitivity: 44.78%.
Specificity: 88.40%.
The within training data classification error rate is lower than those by LDA and QDA with the original input.
Figure 4.6: Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $x_1, x_2, x_{12}, x_1^2, x_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.