Regularized Discriminant Analysis and Reduced-Rank LDA

Jia Li

Department of Statistics
The Pennsylvania State University

Email: jiali@stat.psu.edu
http://www.stat.psu.edu/∼jiali

Regularized Discriminant Analysis

- A compromise between LDA and QDA.
- Shrink the separate covariances of QDA toward a common covariance as in LDA.
- Regularized covariance matrices:

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma} \ .$$

- The quadratic discriminant function $\delta_k(x)$ is defined using the shrunken covariance matrices $\hat{\Sigma}_k(\alpha)$.
- The parameter α controls the complexity of the model.

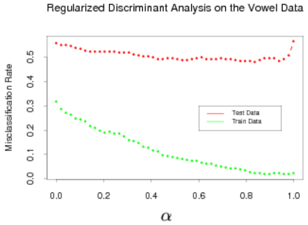


Figure 4.7: *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of* $\alpha \in [0, 1]$. *The optimum for the test data occurs around* $\alpha = 0.9$, *close to quadratic discriminant analysis.*

Computations for LDA

- Discriminant function:

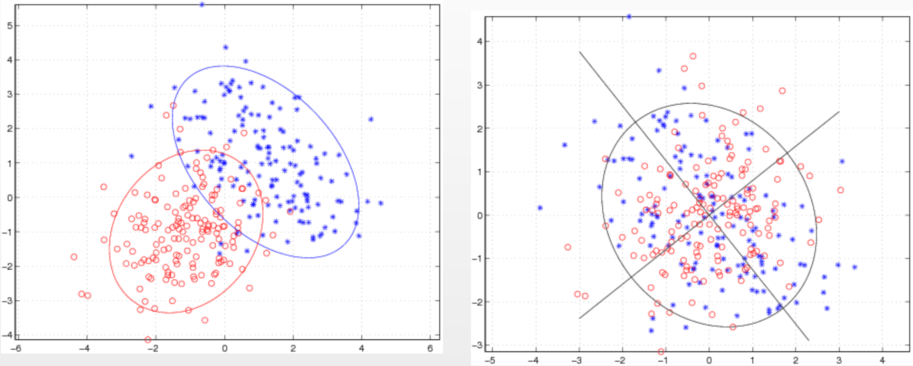
$$\delta_k(x) = -\frac{1}{2}\log|\hat{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^T\hat{\Sigma}_k^{-1}(x-\mu_k) + \log \pi_k \ .$$

- Eigen-decomposition of $\hat{\Sigma}_k$: $\hat{\Sigma}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k^T$. $\mathbf{D}_k$ is diagonal with elements d_{kl}, $l = 1, 2, ..., p$. $\mathbf{U}_k$ is $p \times p$ orthonormal.
-

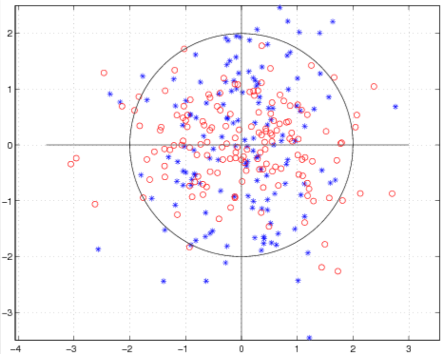
$$\begin{aligned} & (x-\hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) \\ = \ & [\mathbf{U}_k^T(x-\mu_k)]^T\mathbf{D}_k^{-1}[\mathbf{U}_k^T(x-\mu_k)] \\ = \ & [\mathbf{D}_k^{-\frac{1}{2}}\mathbf{U}_k^T(x-\mu_k)]^T[\mathbf{D}_k^{-\frac{1}{2}}\mathbf{U}_k^T(x-\mu_k)] \end{aligned}$$

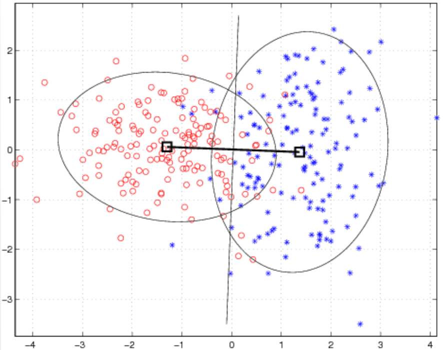
- $\log|\hat{\Sigma}_k| = \sum_l \log d_{kl}$.

- LDA, $\hat{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$:
  - Sphere the data $\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T X \rightarrow X^*$ and $\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \mu_k \rightarrow \mu_k^*$.
  - For the transformed data and class centroids, classify x^* to the closest class centroid in the transformed space, modulo the effect of the class prior probabilities π_k.



The geometric illustration of LDA. Left: Original data in the two classes. The ellipsis represent the two estimated covariance matrices. Right: The class mean removed data and the estimated common covariance matrix.





The geometric illustration of LDA. Left: The sphered mean removed data. Right: The sphered data in the two classes, the sphered means, and the decision boundary.

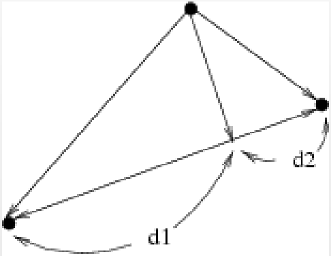
Reduced-Rank LDA

Binary classification

- Decision boundary is given by the following linear equation:

$$\log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + x^T \Sigma^{-1}(\mu_1 - \mu_2) = 0 \,.$$

- Only the projection of X on the direction $\Sigma^{-1}(\mu_1 - \mu_2)$ matters.
- If the data are sphered, only the projection of X^* on $\mu_1^* - \mu_2^*$ is needed.



- Suppose data are sphered.
- The subspace spanned by the K centroids is of rank $K-1$, denoted by H_{K-1}.
- Data can be viewed in H_{K-1} without losing any information.
- When $K > 3$, we might want to find a subspace $H_L \subseteq H_{K-1}$ optimal for LDA in some sense.

Optimization Criterion

- Fisher's optimization criterion: the projected centroids were spread out as much as possible comparing with variance.
- Find the linear combination $Z = a^T X$ such that the between-class variance is maximized relative to the within-class variance, where $a = (a_1, a_2, ..., a_p)^T$.
- Assume the within-class covariance matrix of X is $\mathbf{W}$, i.e., the common covariance matrix of the classes.

- The between-class covariance matrix is $\mathbf{B}$. Suppose μ_k is a column vector denoting the mean vector of class k.

$$\begin{aligned}\mu &= \sum_{k=1}^{K} \pi_k \mu_k \\ \mathbf{B} &= \sum_{k=1}^{K} \pi_k (\mu_k - \mu)(\mu_k - \mu)^T\end{aligned}$$

Note π_k is the percentage of class k samples in the entire data set.

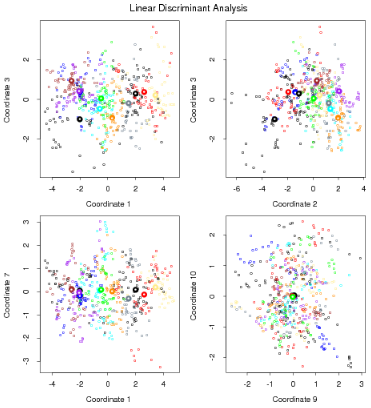


Figure 4.8: *Four projections onto pairs of canonical variates. Notice that as the rank of the canonical variates increases, the centroids become less spread out. In the lower right panel they appear to be superimposed, and the classes most confused.*

- For the linear combination Z, the between-class variance is $a^T \mathbf{B} a$ and the within-class variance is $a^T \mathbf{W} a$.
- Fisher's optimization becomes

$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a} .$$

- Eigen-decomposition of $\mathbf{W} = \mathbf{V}_W \mathbf{D}_W \mathbf{V}_W^T$.
- $\mathbf{W} = (\mathbf{W}^{\frac{1}{2}})^T \mathbf{W}^{\frac{1}{2}}$, where $\mathbf{W}^{\frac{1}{2}} = \mathbf{D}_W^{\frac{1}{2}} \mathbf{V}_W^T$.
- Define $b = \mathbf{W}^{\frac{1}{2}} a$, then $a = \mathbf{W}^{-\frac{1}{2}} b$. The optimization becomes

$$\max_b \frac{b^T (\mathbf{W}^{-\frac{1}{2}})^T \mathbf{B} \mathbf{W}^{-\frac{1}{2}} b}{b^T b}$$

- Define $\mathbf{B}^* = (\mathbf{W}^{-\frac{1}{2}})^T \mathbf{B} \mathbf{W}^{-\frac{1}{2}}$.

- Eigen-decomposition of $\mathbf{B}^* = \mathbf{V}^* \mathbf{D}_B \mathbf{V}^{*T}$.
  $\mathbf{V}^* = (v_1^*, v_2^*, ..., v_p^*)$.
- The maximization is achieved by $b = v_1^*$, the first eigen vector of $\mathbf{B}^*$.
- Similarly, one can find the next direction $b_2 = v_2^*$ that is orthogonal to $b_1 = v_1^*$ and maximizes $b_2^T \mathbf{B}^* b_2 / b_2^T b_2$.
- Since $a = \mathbf{W}^{-\frac{1}{2}} b$, convert to the original problem,

$$a_l = \mathbf{W}^{-\frac{1}{2}} v_l^* .$$

- The a_l (also denoted as v_l in the textbook) are referred to as *discriminant coordinates* or *canonical variates*.

- Summarization on obtaining discriminant coordinates:
  - Find the centroids for all the classes.
  - Find between-class covariance matrix $\mathbf{B}$ using the centroid vectors.
  - Find within-class covariance matrix $\mathbf{W}$, i.e., $\hat{\Sigma}$.
  - By eigen-decomposition

$$\mathbf{W} = (\mathbf{W}^{\frac{1}{2}})^T \mathbf{W}^{\frac{1}{2}} = (\mathbf{D}_W^{\frac{1}{2}} \mathbf{V}_W^T)^T \mathbf{D}_W^{\frac{1}{2}} \mathbf{V}_W^T \ .$$

  - Compute

$$\mathbf{B}^* = (\mathbf{W}^{-\frac{1}{2}})^T \mathbf{B} \mathbf{W}^{-\frac{1}{2}} = \mathbf{D}_W^{-\frac{1}{2}} \mathbf{V}_W^T \mathbf{B} \mathbf{V}_W \mathbf{D}_W^{-\frac{1}{2}} \ .$$

  - Eigen-decomposition of $\mathbf{B}^*$:

$$\mathbf{B}^* = \mathbf{V}^* \mathbf{D}_B \mathbf{V}^{*T} \ .$$

  - The discriminant coordinates are: $a_l = \mathbf{W}^{-\frac{1}{2}} v_l^*$.

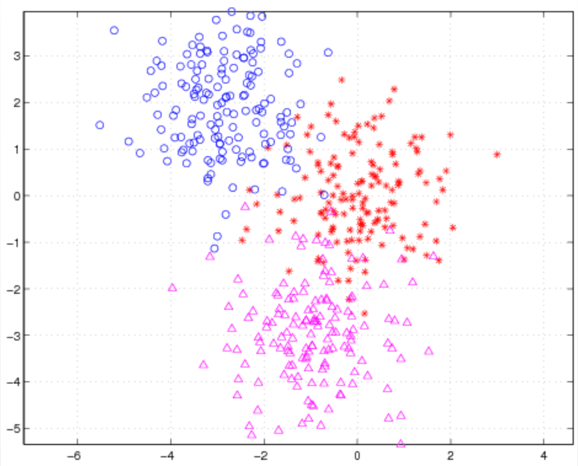
Simulation

- Three classes with equal prior probabilities $1/3$.
- Input is two dimensional.
- The class conditional density of X is a normal distribution.
- The common covariance matrix $\Sigma = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$.
- The three mean vectors are:

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} -3 \\ 2 \end{pmatrix} \quad \mu_3 = \begin{pmatrix} -1 \\ -3 \end{pmatrix}$$

- Total of 450 samples are drawn with 150 in each class for training.
- Another set of 450 samples are drawn with 150 in each class for testing.

The scatter plot of the test data. Red: class 1. Blue: class 2. Magenta: class 3.

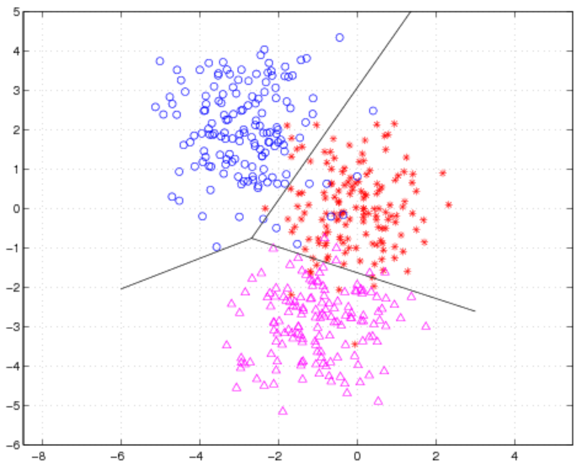


LDA Result

- Priors: $\hat{\pi}_1 = \hat{\pi}_2 = \hat{\pi}_3 = \frac{150}{450} = 0.3333$.
- The three mean vectors are:

$$\hat{\mu}_1 = \begin{pmatrix} -0.0757 \\ -0.0034 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} -2.8310 \\ 1.9847 \end{pmatrix} \quad \hat{\mu}_3 = \begin{pmatrix} -0.9992 \\ -2.9005 \end{pmatrix}$$

- Estimated covariance matrix: $\hat{\Sigma} = \begin{pmatrix} 0.9967 & 0.0020 \\ 0.0020 & 1.0263 \end{pmatrix}$.
- Decision boundaries:
  - Between class 1 (red) and 2 (blue):
  $$5.9480 + 2.7684X_1 - 1.9427X_2 = 0 \,.$$
  - Between class 1 (red) and 3 (magenta):
  $$4.5912 + 0.9209X_1 + 2.8211X_2 = 0 \,.$$
  - Between class 2 (blue) and 3 (magenta):
  $$-1.3568 - 1.8475X_1 + 4.7639X_2 = 0 \,.$$



Classification error rate on the test data set: 7.78%.

Discriminant Coordinates

- Between-class covariance matrix:

$$\mathbf{B} = \begin{pmatrix} 1.3111 & -1.3057 \\ -1.3057 & 4.0235 \end{pmatrix} .$$

- Within-class covariance matrix:

$$\mathbf{W} = \begin{pmatrix} 0.9967 & 0.0020 \\ 0.0020 & 1.0263 \end{pmatrix} .$$

- $\mathbf{W}^{\frac{1}{2}} = \begin{pmatrix} -0.0686 & -1.0108 \\ 0.9960 & -0.0676 \end{pmatrix}.$
- $\mathbf{B}^* = (\mathbf{W}^{-\frac{1}{2}})^T \mathbf{B} \mathbf{W}^{-\frac{1}{2}} = \begin{pmatrix} 3.7361 & 1.4603 \\ 1.4603 & 1.5050 \end{pmatrix}.$

- Eigen-decomposition of $\mathbf{B}^*$:

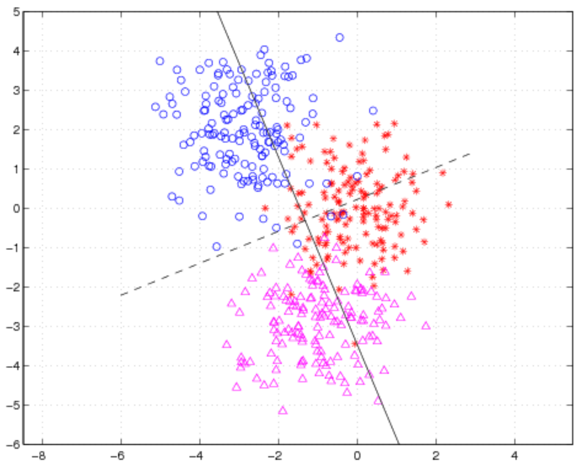
$$\begin{aligned}
\mathbf{B}^* &= \mathbf{V}^*\mathbf{D}_B\mathbf{V}^{*T} \\
\mathbf{V}^* &= \begin{pmatrix} 0.8964 & 0.4432 \\ 0.4432 & -0.8964 \end{pmatrix} \\
\mathbf{D}_B &= \begin{pmatrix} 4.4582 & 0 \\ 0 & 0.7830 \end{pmatrix} .
\end{aligned}$$

- The two discriminant coordinates are:

$$v_1 = \mathbf{W}^{-\frac{1}{2}} v_1^* = \begin{pmatrix} -0.0668 & 0.9994 \\ -0.9848 & -0.0678 \end{pmatrix} \begin{pmatrix} 0.8964 \\ 0.4432 \end{pmatrix} = \begin{pmatrix} 0.3831 \\ -0.9128 \end{pmatrix}$$

$$v_2 = \mathbf{W}^{-\frac{1}{2}} v_2^* = \begin{pmatrix} -0.9255 \\ -0.3757 \end{pmatrix}$$

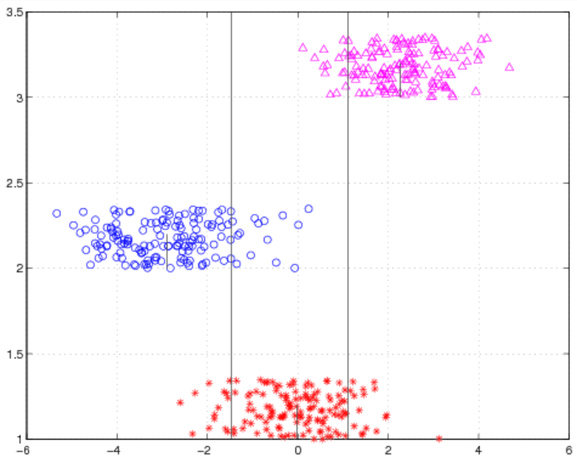
- Project data onto v_1 and classify using only this 1-D data.
- The projected data are $x_i^T v_1$, $i = 1, ..., N$.



Solid line: first DC. Dash line: second DC.

Projection on the First DC

Projection of the training data on the first discriminant coordinate.



- Perform LDA on the projected data.
- The classification rule is:

$$\hat{G}(x) = \begin{cases} 1 & -1.4611 \leq x^T v_1 \leq 1.1195 \\ 2 & x^T v_1 \leq -1.4611 \\ 3 & x^T v_1 \geq 1.1195 \end{cases}$$

- Error rate on the test data: 12.67%.

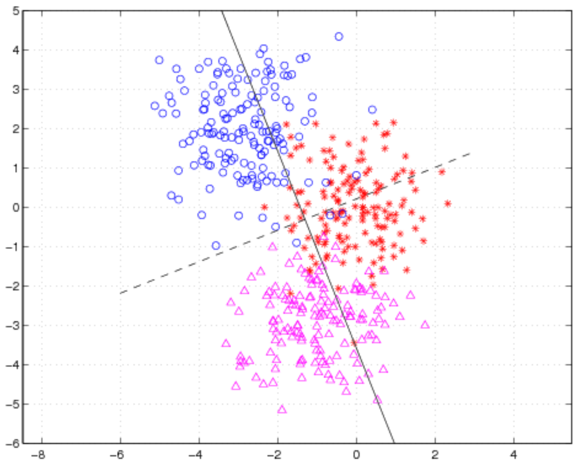
Principal Component Direction

- Find the input matrix of X, or do singular value decomposition of mean removed X, to find the principal component directions.
- Denote the covariance matrix by $\mathbf{T}$:

$$\mathbf{T} = \begin{pmatrix} 2.3062 & -1.3066 \\ -1.3066 & 5.0542 \end{pmatrix} .$$

- Eigen-decomposition of $\mathbf{T} = \mathbf{V}_T \mathbf{D}_T \mathbf{V}_T^T$:

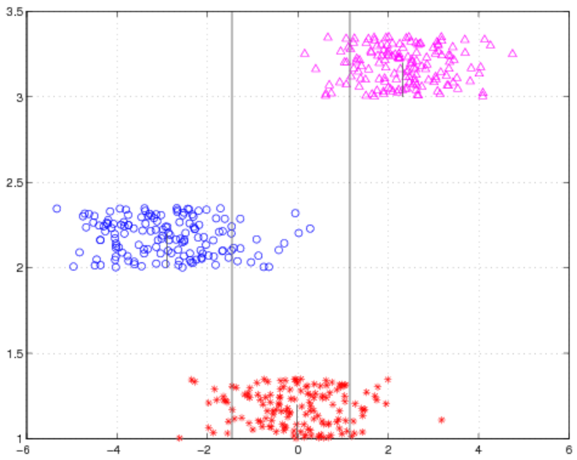
$$\mathbf{V}_T = \begin{pmatrix} 0.3710 & -0.9286 \\ -0.9286 & -0.3710 \end{pmatrix} \mathbf{D}_T = \begin{pmatrix} 5.5762 & 0 \\ 0 & 1.7842 \end{pmatrix}$$



Solid line: first PCD. Dash line: second PCD.

Results Based on the First PC

Projection of data on the first PC. The boundaries between classes are shown.



- Perform LDA on the projected data.
- The classification rule is:

$$\hat{G}(x) = \begin{cases} 1 & -1.4592 \leq x^T v_1 \leq 1.1489 \\ 2 & x^T v_1 \leq -1.4592 \\ 3 & x^T v_1 \geq 1.1489 \end{cases}$$

- Error rate on the test data: 13.11%.

Comparison

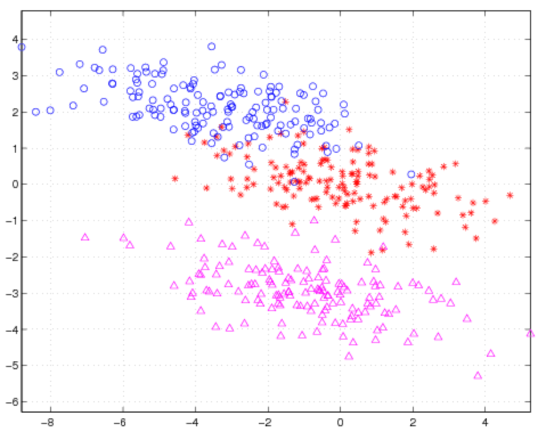
- It is generally true that $\mathbf{T} = \mathbf{B} + \mathbf{W}$.
- For the given example $\mathbf{W} \approx \mathbf{I}$; and the true within-class covariance matrix is $\mathbf{I}$.
- Ideally, for this example, both the discriminant coordinates and the principal component directions are simply the eigenvectors of $\mathbf{B}$.
- In general, discriminant coordinates and principal component directions are different.
- To compute PC directions, class information is not needed; and hence PCs have more flexible applications.
- For classification, DCs tend to be better.

A New Simulation

- Change the common covariance matrix Σ to:

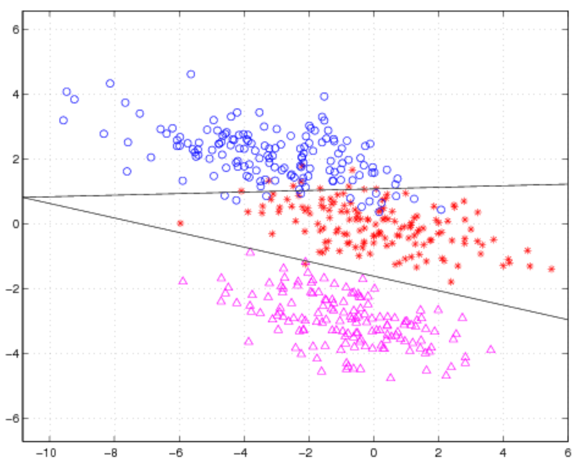
$$\begin{pmatrix} 4.0898 & -0.8121 \\ -0.8121 & 0.5900 \end{pmatrix}$$

- The scatter plot of the test data set.



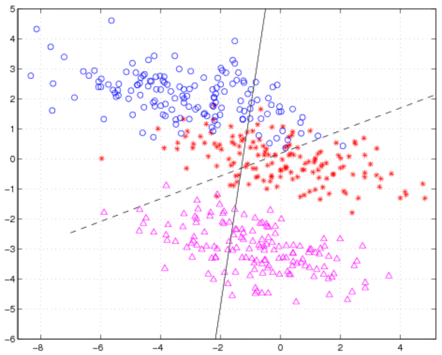
LDA Result

The classification boundaries obtained by LDA. The error rate for the test data is 6%.

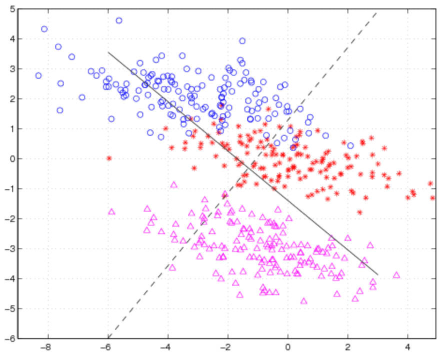


DCs and PC Directions

The solid line indicates the first DC or PC; the dash line indicates the second DC or PC.



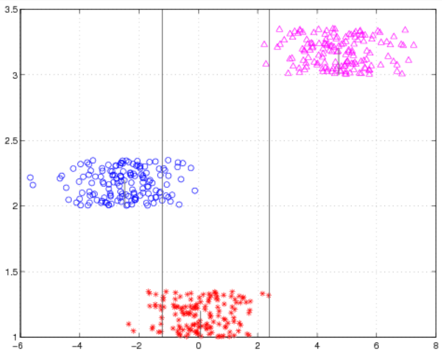
Discriminant coordinates



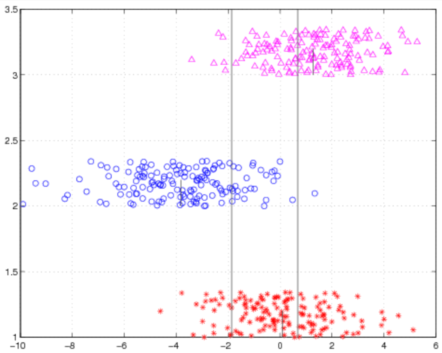
Principal component directions

Projection on 1-D

The LDA results obtained using the projected data onto the first discriminant coordinate and the first principal component direction.



Projection on the first DC (test set error rate: 7.78%)



Projection on the first PCD (test set error rate: 32.44%)