ISyE 6416: Computational Statistics
Spring 2017

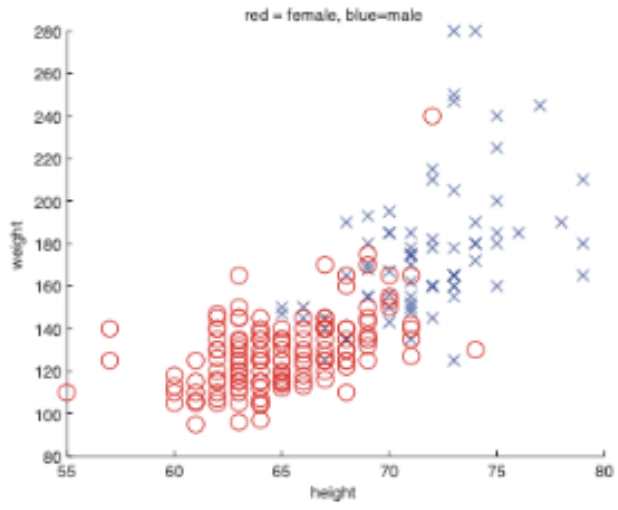
Lecture 5: Discriminant analysis and classification

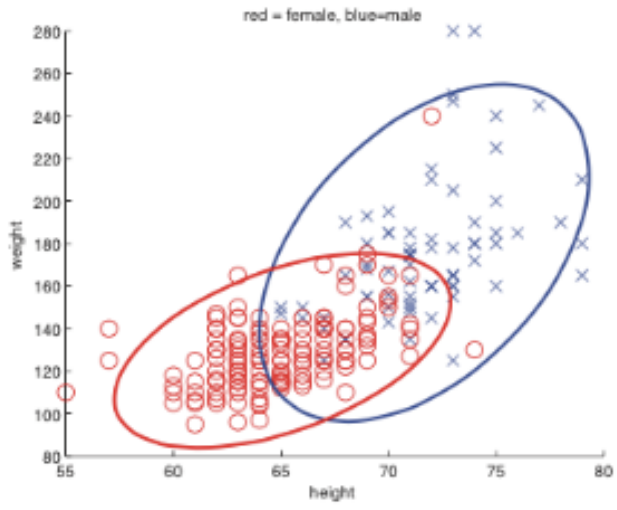
Prof. Yao Xie

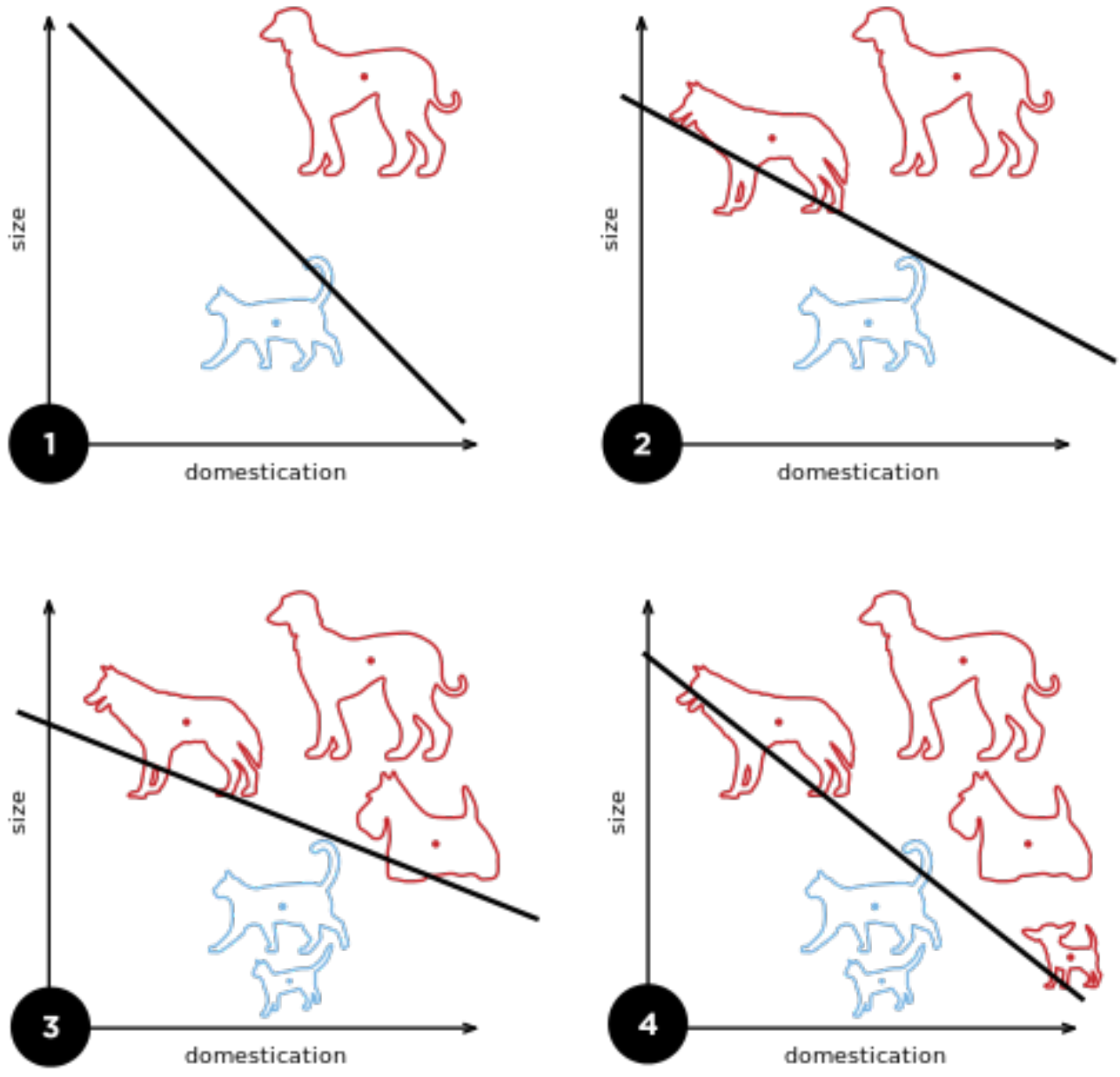
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Classification

- Classification is a predictive task in which the response takes the values across several categories (in the fundamental case, two categories)
- Examples
  - Predicting whether a patient will develop breast cancer or remain healthy, given genetic information
  - Predicting whether or not a user will like a new product, based on user covariates and a history of his/her previous ratings
  - Predicting the voting preference based on voter's social, political, and economical status
- Here we consider supervised classification, i.e., we know the labels for the training data

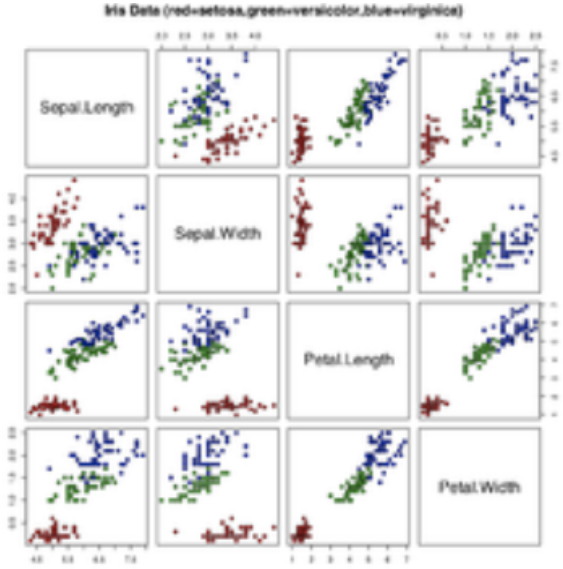

red = female, blue=male
weight
height
280
260
240
220
200
180
160
140
120
100
80
55
60
65
70
75
80


red = female, blue=male
weight
height
280
260
240
220
200
180
160
140
120
100
80
55
60
65
70
75
80


size
1
domestication
size
2
domestication
size
3
domestication
size
4
domestication

Fisher's Irish example

- The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher (1936) as an example of discriminant analysis. [1]
- The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.
- Based on the combinaBon of these four features, Fisher developed a linear discriminant model to disBnguish the species from each other



Classification model-based

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

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- Linear discriminant analysis (LDA) assumes that the data within each class are normally distributed

$$h_j(x) = \mathbb{P}(X = x|C = j) = \mathcal{N}(\mu_i, \Sigma)$$

- Note that the covariance matrix is a common one

$$\Sigma_k = \Sigma, \forall k = 1, \ldots, K$$

- Each class has its own mean $\mu_j \in \mathbb{R}^p$
- Maximum likelihood principle: we find j so that the posterior probability is the largest:

$$\mathbb{P}(C = j|X = x)\pi_j = h_j(x)\pi_j$$

- estimated class given a data (feature vector) x

$$f^{\mathrm{LDA}}(x) = \arg \max_{j=1,\ldots,K} \delta_j(x)$$

- Discriminant function: $\delta_j(x)$, $j = 1, \ldots, K$

$$\delta_j(x) = \underbrace{\mu_j^T \Sigma^{-1}}_{a_j^T} x \underbrace{-\frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j + \log \pi_j}_{b_j}$$

  is an affine function of x for LDA

LDA from data

- In practice, we estimate the model parameters from *training data*
  - $\hat{\pi}_j = n_j/n$ the proportion of observation in class j (n_j: the number of samples labeled as class j)
  - $\hat{\mu}_j = \frac{1}{n_j} \sum_{y=j} x_i$: the centroid of class j
  - $\hat{\Sigma} = \frac{1}{n-K} \sum_{j=1}^{K} \sum_{y_i=j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$ the pooled sample covariance matrix

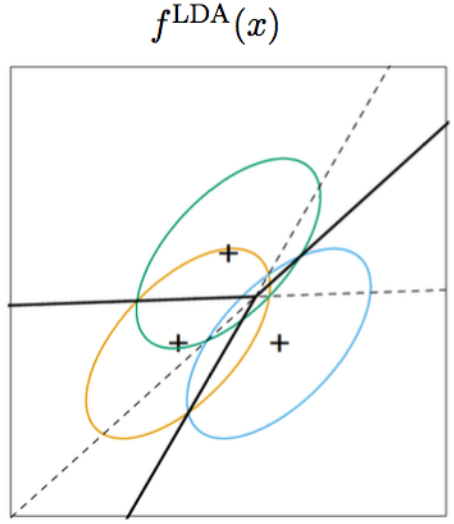
LDA decision boundaries

- Due to our decision rule: find j so that the posterior probability is the largest:

$$\mathbb{P}(C = j|X = x)\pi_j = h_j(x)\pi_j$$

- The decision boundary between classes j and k is the set of all $x \in \mathbb{R}^p$ such that $\delta_j(x) = \delta_k(x)$, i.e.,

$$a_j - a_k + (b_j - b_k)^T x = 0$$



LDA computations and sphering

- The decision boundaries for LDA are useful for graphical purposes, but to classify a new data x_0, we simply compute $\hat{\delta}_j(x_0)$ for each $j = 1, \ldots, K$
- Note that LDA's discriminant function can be written as

$$\frac{1}{2}(x - \hat{\mu}_j)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_j) - \log \hat{\pi}_j$$

- It helps to perform eigendecomposition of the sample covariance matrix $\hat{\Sigma} = UDU^T$, computation can be simplied

$$(x - \hat{\mu}_j)^T \Sigma^{-1}(x - \hat{\mu}_j) = \| \underbrace{D^{-1/2}U^T x}_{\tilde{x}} - \underbrace{D^{-1/2}U^T \hat{\mu}_j}_{\tilde{\mu}_j} \|_2^2$$

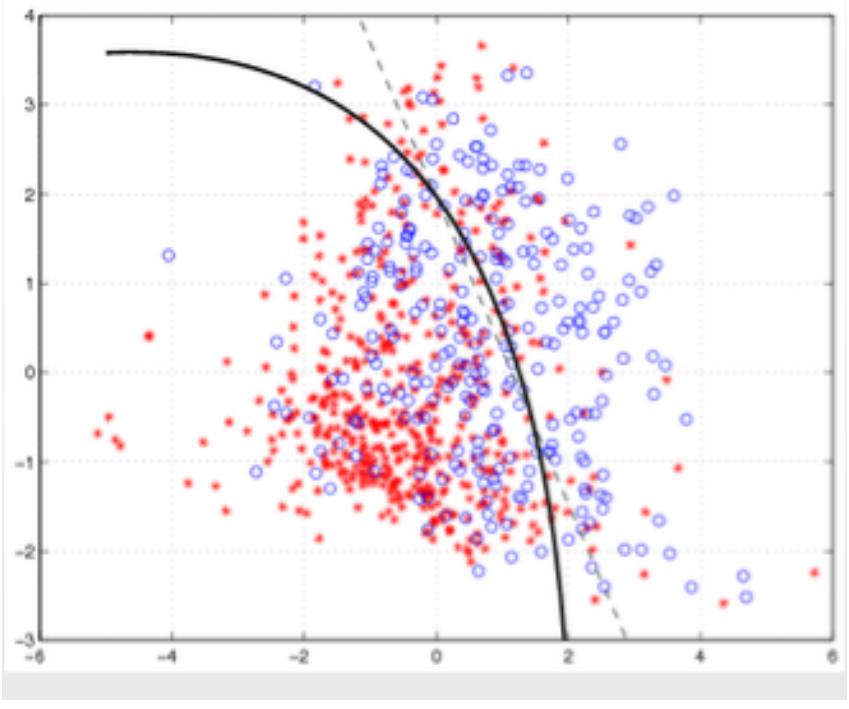
  which is just the Euclidean distance between $\tilde{x}$ and $\tilde{\mu}_j$
- compute the discriminant function $\hat{\delta}_j = \frac{1}{2}\|\tilde{x} - \tilde{\mu}_j\|_2^2 - \log \hat{\pi}_j$, and then find the **nearest centroid**

Quadratic discriminant analysis

- Estimate covariance matrix Σ_k separately for each class k
- Quadratic discriminant function

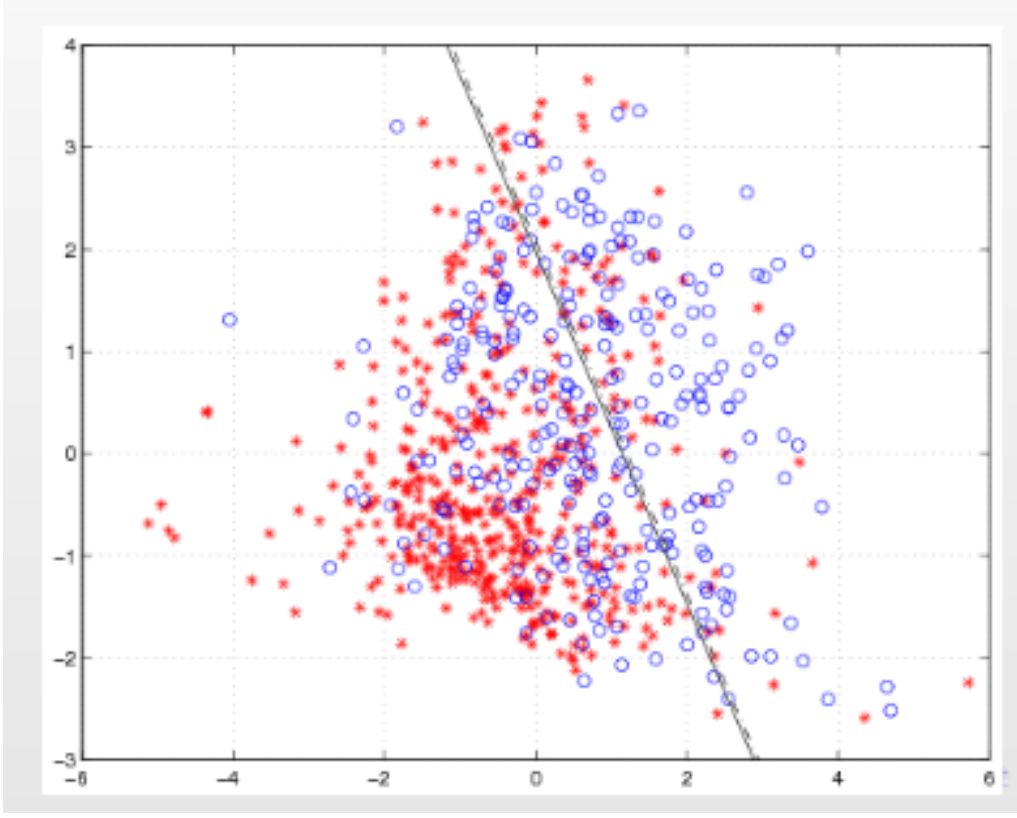
$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

- QDA fits data better than LDA, but has more parameters to estimate.



Diabetes dataset

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.



Diabetes dataset

- Two input variables computed from the principle components of the original 8 variables
- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$

$$\hat{\Sigma} = \begin{bmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1.6634 \end{bmatrix}$$

- LDA Decision rule

$$f^{\mathrm{LDA}}(x) = \begin{cases} 1 & 1.1443 - x_1 - 0.5802 x_2 \geq 0 \\ 2 & \text{otherwise.} \end{cases}$$

Within training data classification error rate: 28.26%

QDA for diabetes dataset

- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{bmatrix}$$

Within training data classification error rate: 29.04%.

LDA on expanded basis

- Expand input space to include X_1X_2, X_1^2, X_2^2
- Input is five dimensional $X = (X_1, X_2, X_1X_2, X_1^2, X_2^2)$

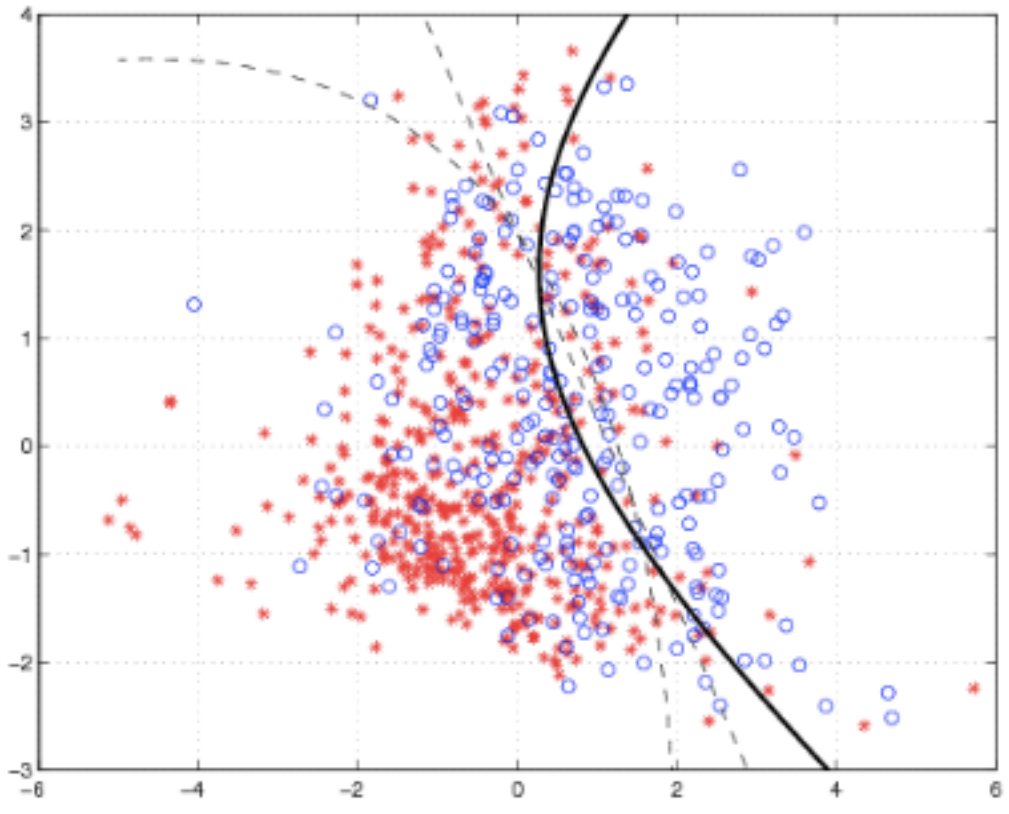
$$\hat{\mu}_1 = \begin{bmatrix} 0.403 \\ 0.193 \\ 0.032 \\ 1.836 \\ 1.630 \end{bmatrix}, \quad \hat{\mu}_2 = \begin{bmatrix} 0.752 \\ 0.361 \\ 0.059 \\ 2.568 \\ 1.912 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 1.7925 & -0.1461 & -0.6254 & 0.3548 & 0.5215 \\ -0.1461 & 1.6634 & 0.6073 & 0.7421 & 1.2193 \\ -0.6254 & 0.6073 & 3.5751 & 1.1118 & 0.5044 \\ 0.3548 & 0.7421 & 1.1118 & 12.3355 & 0.0957 \\ 0.5215 & 1.2193 & -0.5044 & -0.0957 & 4.4650 \end{bmatrix}$$

Classification boundary

$$0.651 - 0.728x_1 - 0.552x_2 - 0.006x_1x_2 - 0.071x_1^2 + 0.170x_2^2 = 0$$

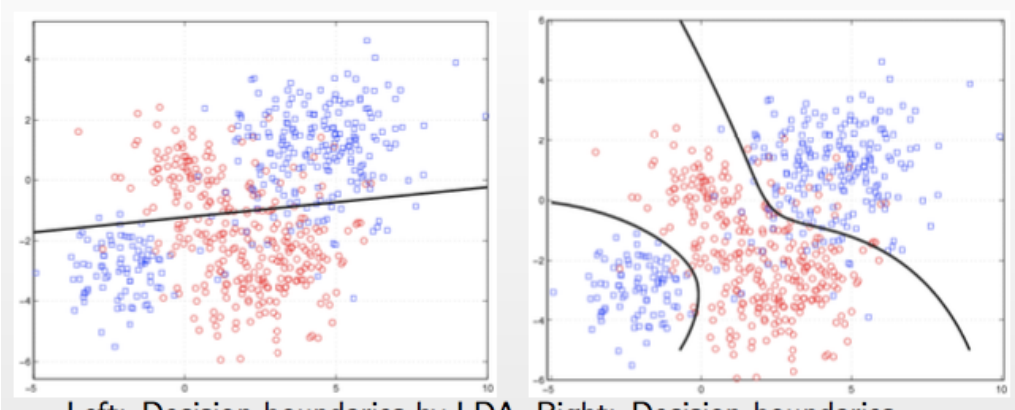
- Classification boundaries obtained by LDA using the expanded input space $X_1, X_2, X_1X_2, X_1^2, X_2^2$.



- Within training data classification error rate: 26.82%. Lower than those by LDA and QDA with the original input.

Mixture discriminant analysis

- A single Gaussian to model a class, as in LDA, can be quite restricted
- A method for classification (supervised) based on mixture models
- Extension of linear discriminant analysis
- a mixture of normals is used to obtain a density estimation for each class



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