Machine Learning and Data Mining

Linear Discriminant Analysis

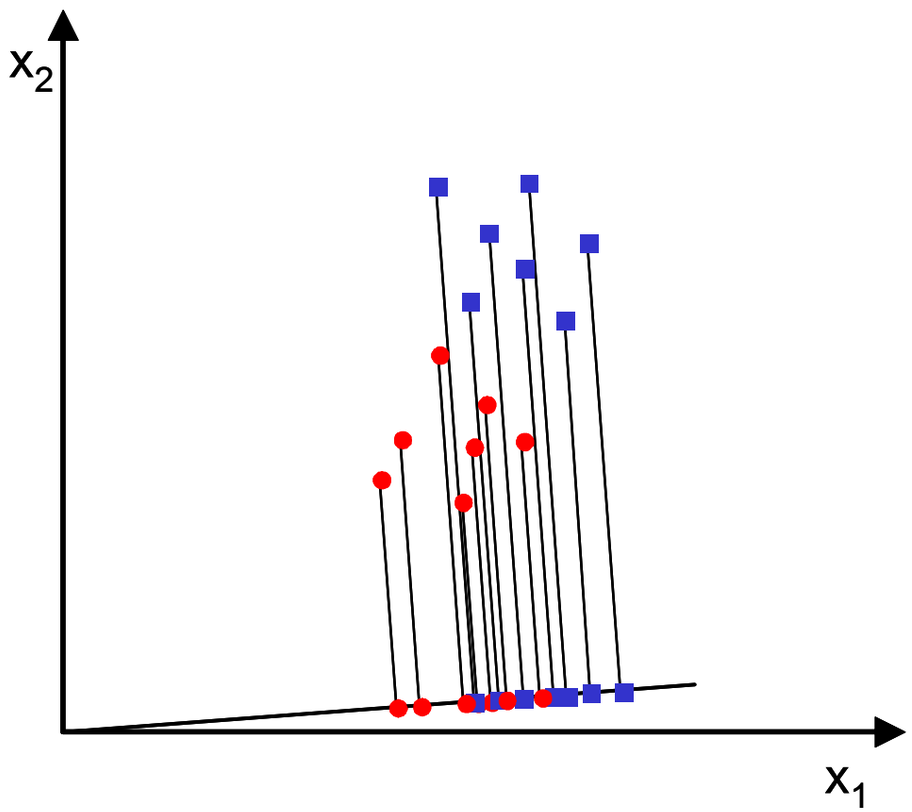
Xiuxia Du, Ph.D.
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte

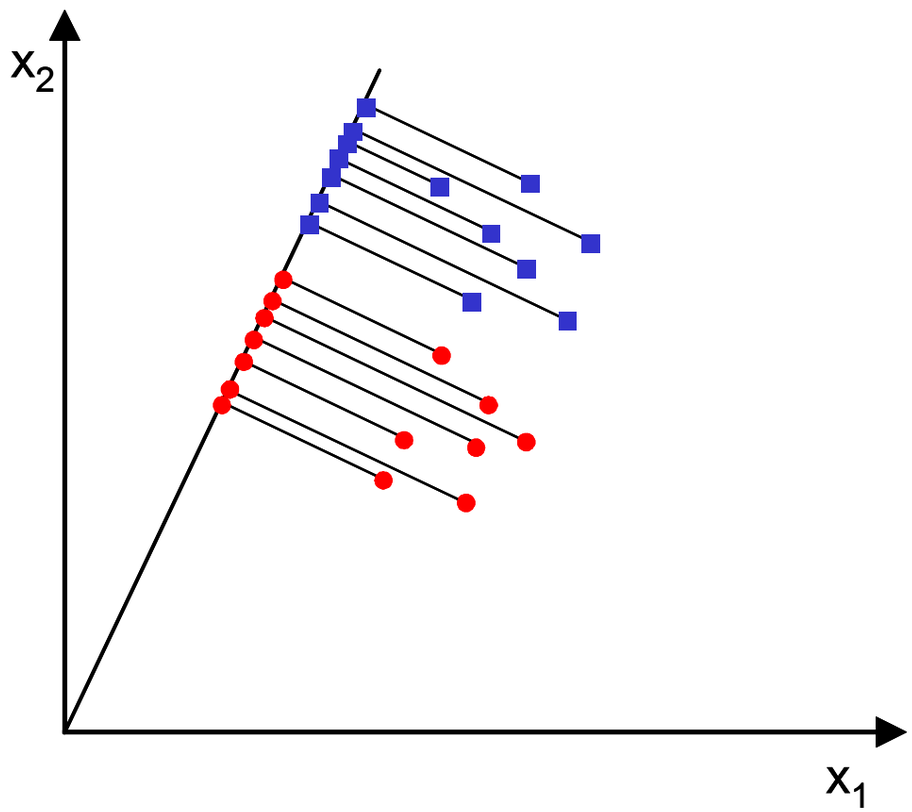
Linear discriminant analysis

Two-classes

Introduction

- The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible.





http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

Problem formulation

- Let's assume we have a set of d-dimensional samples $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$, N_1 of which belong to class c_1, and N_2 to class c_2.
- We seek to obtain a scaler y by projecting the samples X onto a line

$$y = W^T X$$

  where both W and X are column vectors.
- Of all the possible lines, we would like to select the one that maximizes the separability of the scalars.

Measure of separation

- In order to find a good projection vector, we need to define a measure of separation between the projections.
- The mean vector of each class in X and Y feature space is

$$\mu_i = \frac{1}{N_i} \sum_{x \in c_i} x$$

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in c_i} y = \frac{1}{N_i} \sum_{x \in c_i} W^T X = W^T \mu_i$$

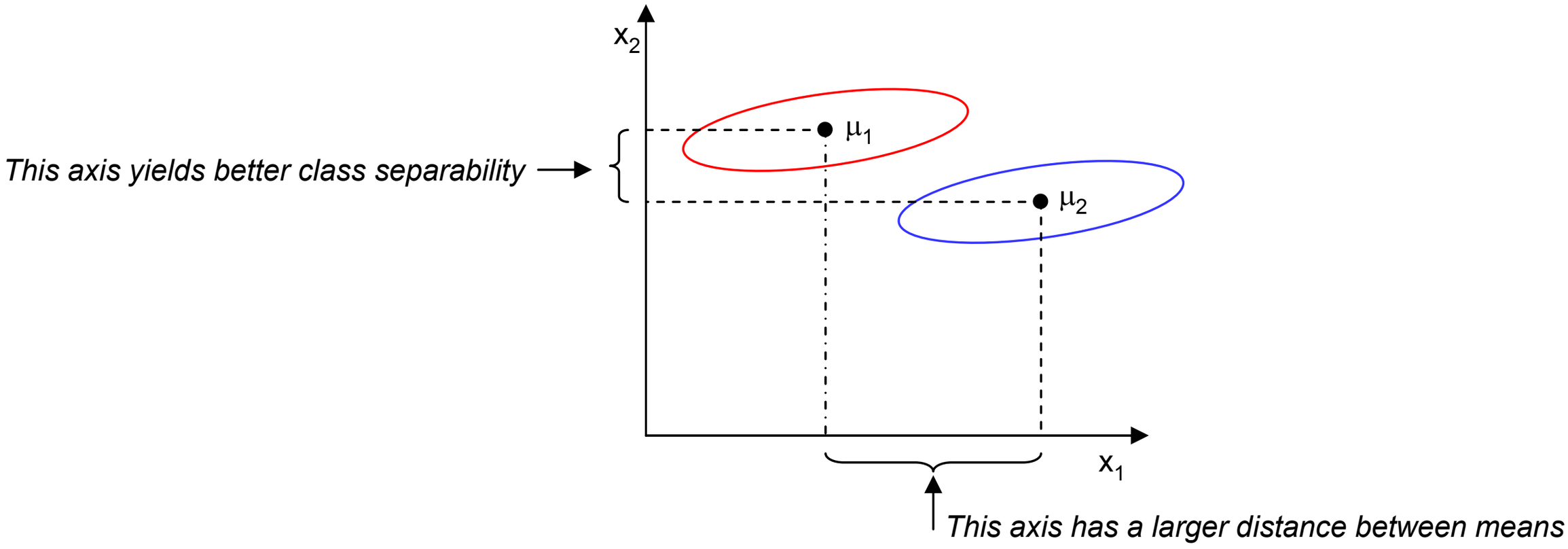
  for $i = 1, 2$.

- We could then choose the distance between the projected means as our objective function

$$J(W) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |W^T(\mu_1 - \mu_2)|$$

Distance itself is not enough

- However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes.



Fisher's linear discriminant

- To solve this issue, Fisher proposed to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter.

- For each class we define the scatter, an equivalent of the variance, as

$$\tilde{s}_i^2 = \sum_{y \in c_i} (y - \tilde{\mu}_i)^2 \quad \text{for } i = 1, 2$$

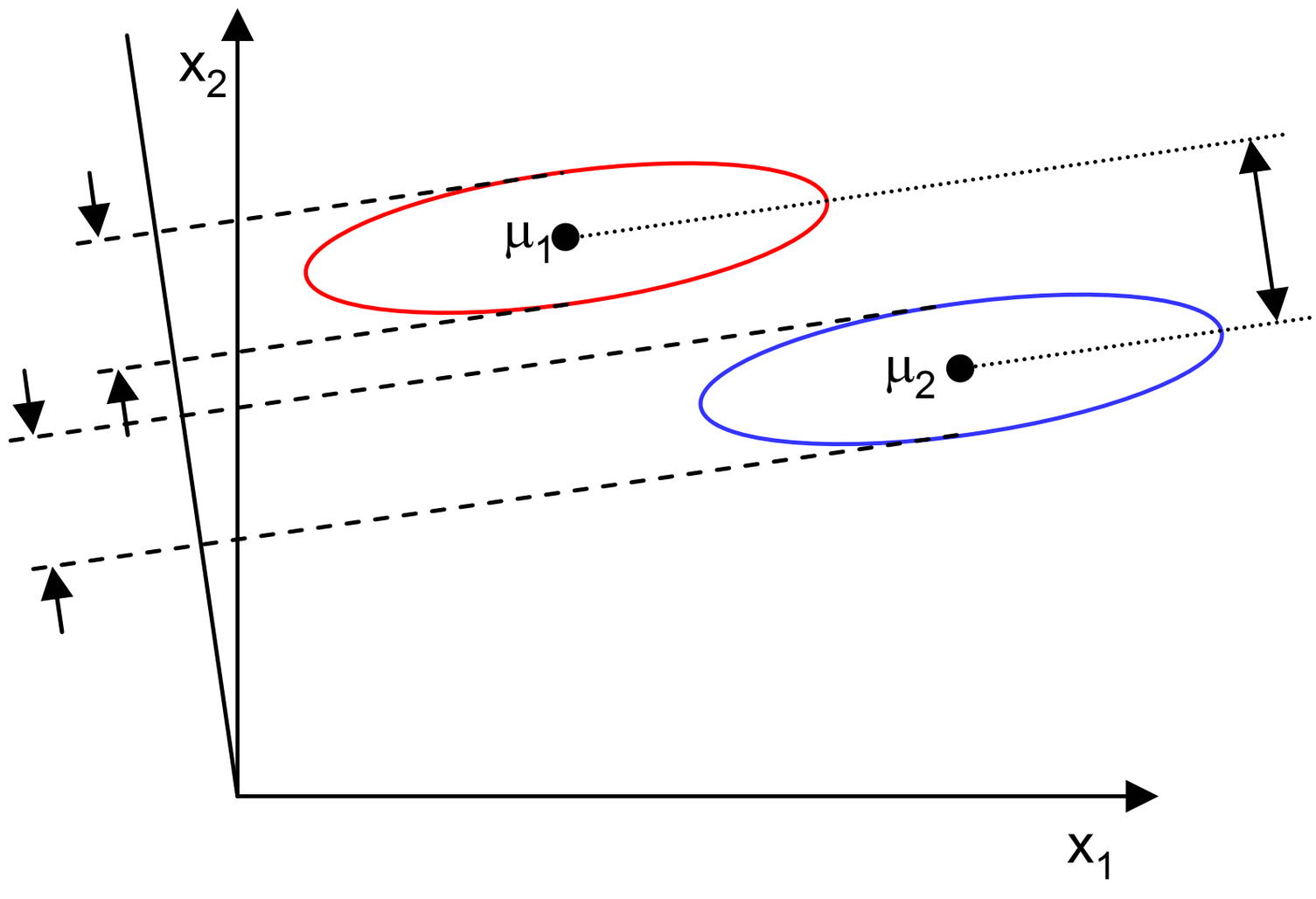
- Then we can calculate $(\tilde{s}_1^2 + \tilde{s}_2^2)$, called the within-class scatter of the projected samples.

- The Fisher linear discriminant is then defined as the linear function $W^T X$ that maximizes the criterion function

$$J(W) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Need both distance and scatter

- Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible.



Scatter matrix in feature space

- In order to find the optimum projection W, we need to express $J(W)$ as an explicit function of W.
- We define a measure of the scatter in multivariate feature space X, which are scatter matrices

$$S_i = \sum_{x \in c_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_1 + S_2 = S_{within}$$

where the dimension of S_1 and S_2 is d by d and S_{within} is called the within-class scatter matrix.

Scatter matrix in projection space

- The scatter of the projection Y can then be expressed as a function of the scatter matrix in feature space X:

$$\begin{aligned}\tilde{s}_i^2 &= \sum_{y \in c_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in c_i} (W^T x - W^T \mu_i)^2 \\ &= \sum_{x \in c_i} W^T (x - \mu_i)(x - \mu_i)^T W = W^T S_i W\end{aligned}$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = W^T S_{within} W$$

Between-class scatter and cost function

- Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (W^T\mu_1 - W^T\mu_2)^2 = W^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T W = W^T S_{between} W$$

- The matrix $S_{between}$ is called the between-class scatter.
- Note that, since $S_{between}$ is the outer product of two vectors, its rank is at most one.

- Now, we can express the Fisher criterion in terms of S_{within} and $S_{between}$ as

$$J(W) = \frac{W^T S_{between} W}{W^T S_{within} W}$$

Solution to LDA

- To find the maximum of $J(W)$, we take the derivative and equate the derivative to zero

$$\frac{d}{dW}\left[J(W)\right] = \frac{d}{dW}\left[\frac{W^T S_{between} W}{W^T S_{within} W}\right] = 0 \Rightarrow S_{within}^{-1} S_{between} W - JW = 0$$

- Solving the generalized eigenvalue problem yields

$$W^* = \text{argmax}_W \left\{\frac{W^T S_{between} W}{W^T S_{within} W}\right\} = S_{within}^{-1}(\mu_1 - \mu_2)$$

- This is known as Fisher's Linear Discriminant, although it is not a discriminant but rather a specific choice for the projection of the data down to one dimension.

Linear discriminant analysis

Multiple classes

Formulation

- Fisher's LDA generalizes gracefully for k-class problems.

- Instead of one projection y, we now seek $k-1$ projections $y_1, y_2, \cdots, y_{k-1}$ by means of $k-1$ projection vectors w_i arranged by columns into a projection matrix $W = [w_1|w_2|\cdots|w_{k-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

where y is of dimension $(k-1) \times 1$, W is of dimension $k \times (k-1)$, and x is of dimension $k \times 1$.

Within-class scatter

- The within-class scatter generalizes as

$$S_{within} = \sum_{i=1}^{k} S_i$$

where

$$\begin{aligned} S_i &= \sum_{x \in c_i} (x - \mu_i)(x - \mu_i)^T \\ \mu_i &= \frac{1}{N_i} \sum_{x \in c_i} x \end{aligned}$$

where x and μ_i are d-dim column vectors and S_i is of dimension $d \times d$ for $i = 1, 2, \cdots, k$.

Between-class scatter

- The between-class scatter becomes

$$S_{between} = \sum_{i=1}^{k} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

where $S_{between}$ is of dimension $d \times d$ and

$$\mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N}\sum_{i=1}^{k} N_i \mu_i$$

- Matrix $S_{total} = S_{between} + S_{within}$ is called the total scatter.

Scatter matrices in the projection space

- Similarly, we define the mean vector and scatter matrices for the projected samples as

$$\begin{aligned}
\tilde{\mu}_i &= \frac{1}{N_i}\sum_{y\in c_i} y \\
\tilde{S}_{within} &= \sum_{i=1}^{k}\sum_{y\in c_i}(y-\tilde{\mu}_i)(y-\tilde{\mu}_i)^T \\
\tilde{\mu} &= \frac{1}{N}\sum_{\forall y} y \\
\tilde{S}_{between} &= \sum_{i=1}^{k} N_i(\tilde{\mu}_i-\tilde{\mu})(\tilde{\mu}_i-\tilde{\mu})^T
\end{aligned}$$

where $\tilde{\mu}_i$ and $\tilde{\mu}$ are of dimension $(k-1)\times 1$. $\tilde{S}_{within}$ and $\tilde{S}_{between}$ are of dimension $(k-1)\times(k-1)$.

Cost function

- From our derivation for the two-class problem, we can write

$$\begin{aligned} \tilde{S}_{within} &= W^T S_{within} W \\ \tilde{S}_{between} &= W^T S_{between} W \end{aligned}$$

- Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scaler (it has $k-1$ dimensions), we use the determinant of the scatter matrices to obtain a scaler objective function

$$J(W) = \frac{|\tilde{S}_{between}|}{|\tilde{S}_{within}|} = \frac{|W^T S_{between} W|}{|W^T S_{within} W|}$$

Solution

- We will now seek the projection matrix W^* that maximizes this ratio.
- It can be shown that the optimal projection matrix W^* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem

$$\begin{aligned} W^* &= [W_1^*|W_2^*|\cdots|W_{k-1}^*] \\ &= \text{argmax}\frac{|W^T S_{between} W|}{|W^T S_{within} W|} \Rightarrow (S_{between} - \lambda_i S_{within}) W_i^* = 0 \end{aligned}$$

- Therefore, the projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_{within}^{-1} S_{between}$.