CPSC 8420 Advanced Machine Learning
Week 7: Linear Discriminate Analysis

Dr. Kai Liu

February 22, 2022

Motivation

Linear Discriminate Analysis is efficient in binary classification and can be easily modified into multi-classification.

- PCA aims to reduce the dimension but it is for unsupervised learning.
- We need propose a supervised learning (classification) method while reducing data dimension.

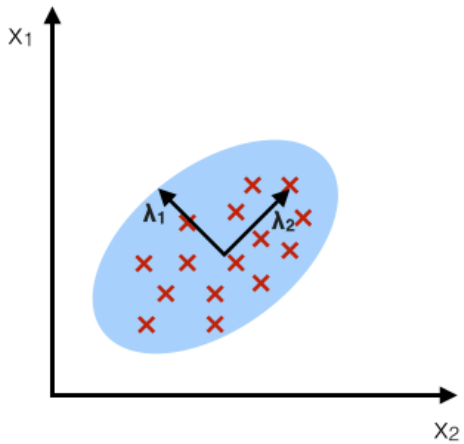
Learning Outcomes

Our goal for today's lecture is to understand:

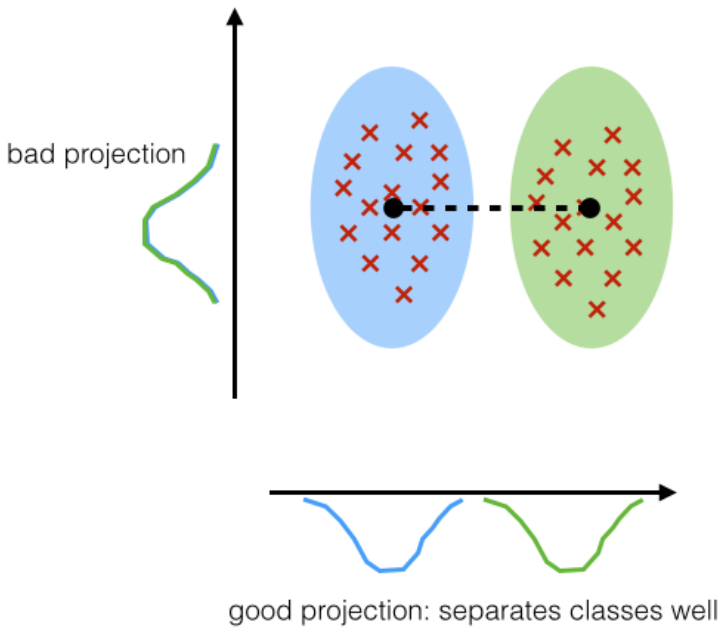
- How to make use of LDA for binary classification
- How to formulate and transfer into an optimization problem
- How to obtain the solution

A Gentle Start

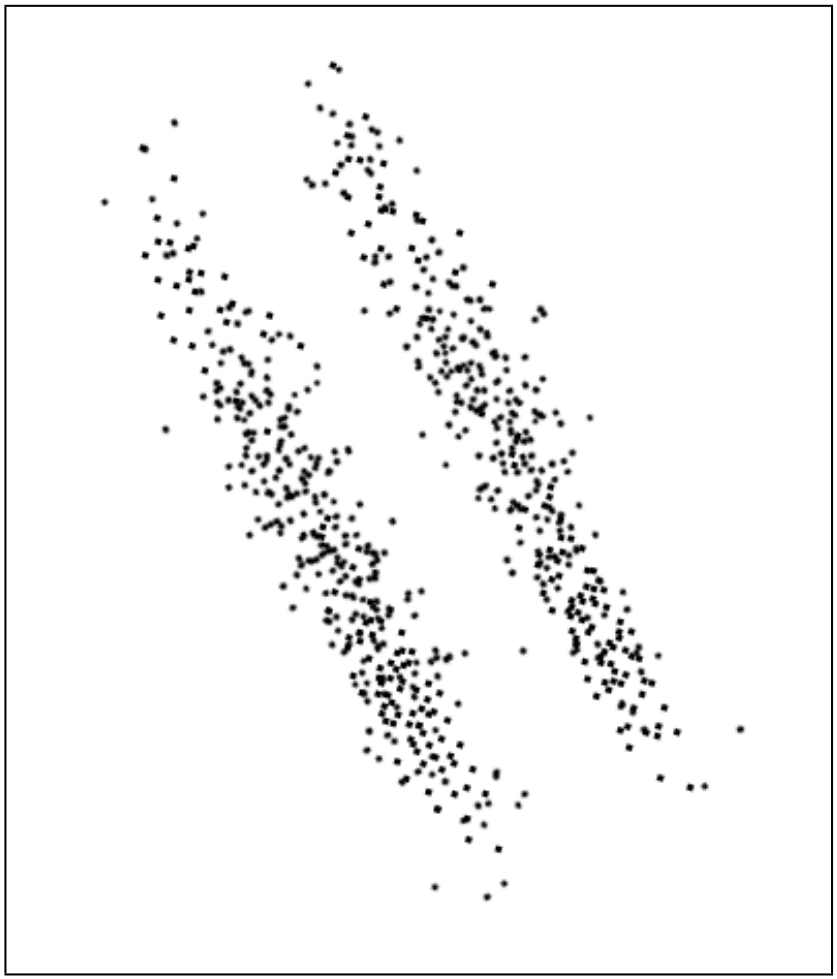
PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



A Gentle Start



A Gentle Start

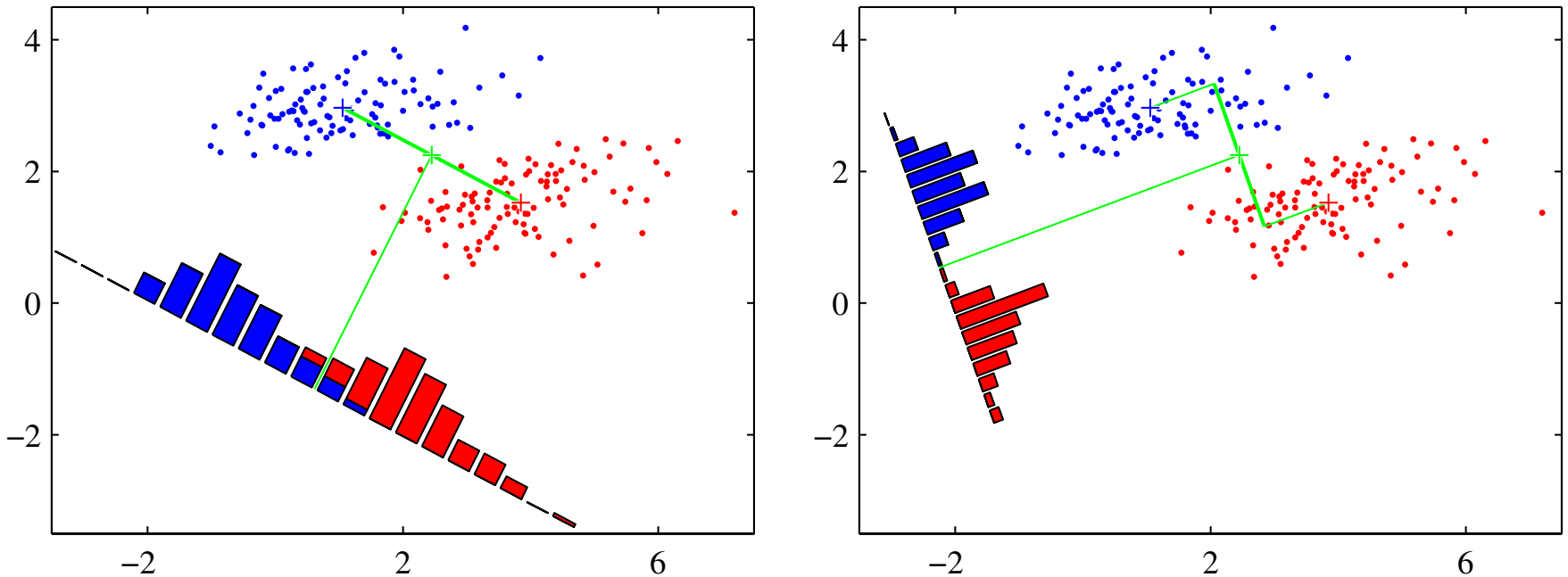


Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

A Gentle Start

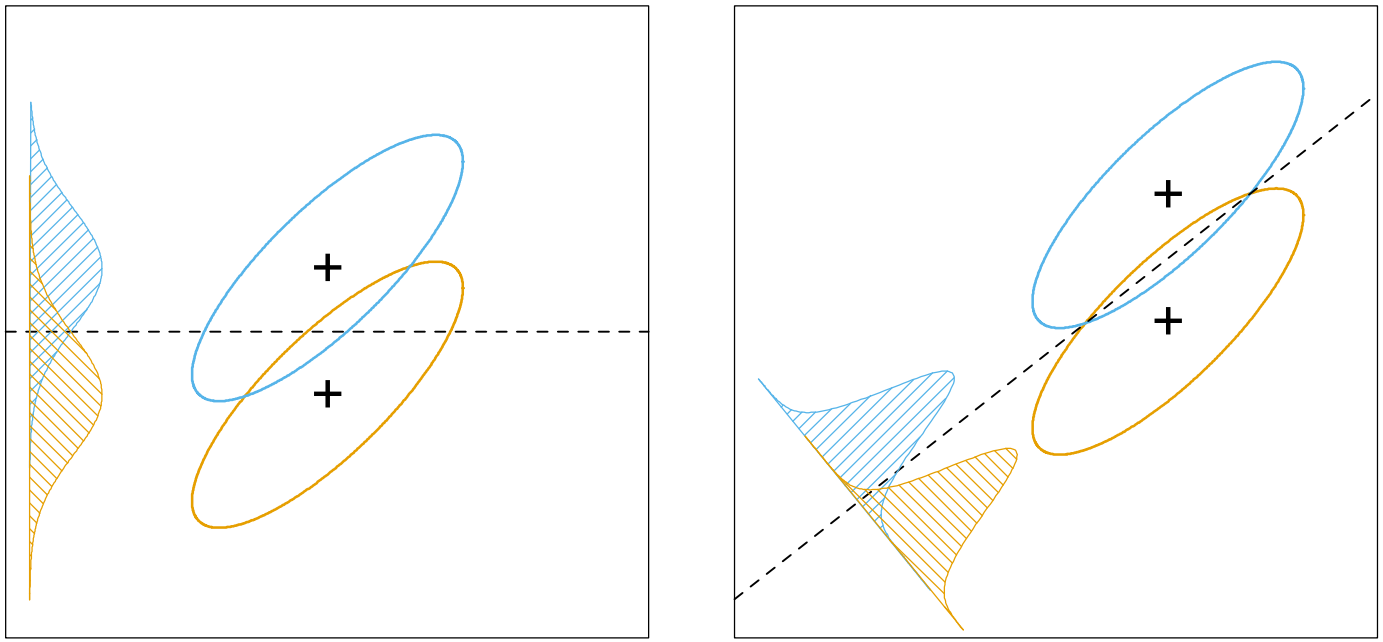
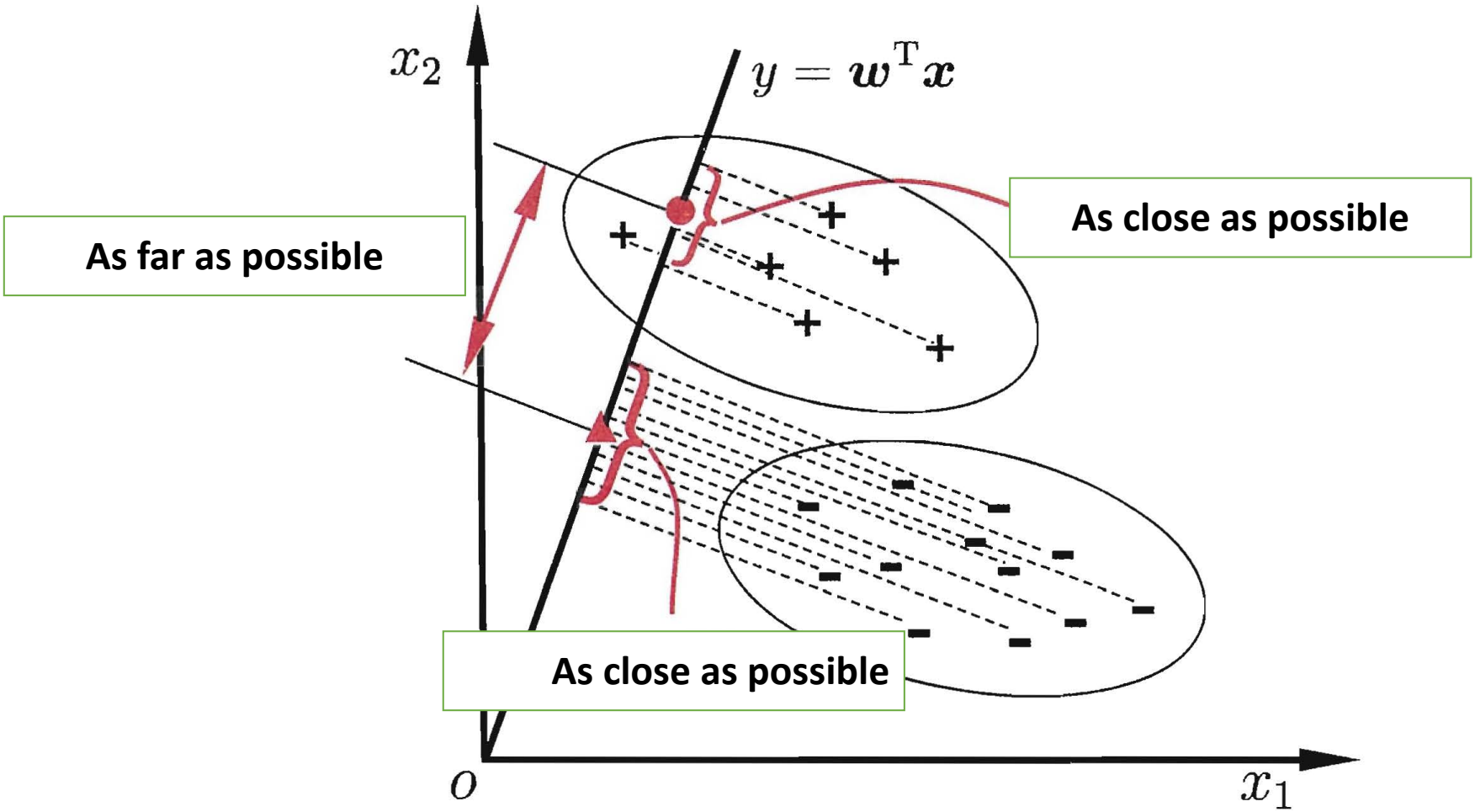


FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

Projection for Classification



Objective

$$
\begin{aligned}
\mathbf{J} &= \frac{\|\mathbf{w}^T\mu_0 - \mathbf{w}^T\mu_1\|_2^2}{\mathbf{w}^T\mathbf{\Sigma}_0\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}_1\mathbf{w}} \\
&= \frac{\mathbf{w}^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T\mathbf{w}}{\mathbf{w}^T(\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)\mathbf{w}}
\end{aligned} \tag{1}
$$

Objective

$$
\begin{aligned}
S_w &= \mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 \\
&= \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^T \\
S_b &= (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T
\end{aligned}
\tag{2}
$$

Objective

$$\mathbf{J} = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \tag{3}$$

which is equivalent to:

$$\min -\mathbf{w}^T S_b \mathbf{w} \quad s.t \quad \mathbf{w}^T S_w \mathbf{w} = 1 \tag{4}$$

Solution

$$
\begin{aligned}
S_b \mathbf{w} &= \lambda S_w \mathbf{w} \\
\Longrightarrow \quad \mathbf{w} &= S_w^{-1}(\mu_0 - \mu_1)
\end{aligned}
\tag{5}
$$

Classification

- Given a new datapoint x, how can we determine whether it belongs to X_0 or X_1?

Classification

- Given a new datapoint x, how can we determine whether it belongs to X_0 or X_1?
- We can compute $\mathbf{w}^T x$, and calculate to which it is closer: $\mathbf{w}^T \mu_0$, $\mathbf{w}^T \mu_1$.

Classification

- Given a new datapoint x, how can we determine whether it belongs to X_0 or X_1?
- We can compute $\mathbf{w}^T x$, and calculate to which it is closer: $\mathbf{w}^T \mu_0$, $\mathbf{w}^T \mu_1$.
- If it is closer to $\mathbf{w}^T \mu_0$, then X_0, otherwise X_1.