Linear Discriminant Analysis

B. Ravindran

Problem Setting

- $\mathcal{X} \subseteq \mathcal{R}^p$ is the input space
  - $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is a random variable describing the input
- $\mathcal{Y} \subseteq \Gamma$ is the output space with K number of classes.
  - $\mathbf{Y}$ is a random variable describing the output.
- $\Pr(\mathbf{X}, \mathbf{Y}) = \Pr(\mathbf{Y} \mid \mathbf{X}) \Pr(\mathbf{X})$ is the data distribution
  - $\Pr(\mathbf{Y} \mid \mathbf{X} = \bar{x})$ is the predicted output probabilities given an input $\bar{x} \in \mathcal{X}$

Problem Setting

- We are interested in the probability of a class given the data point:

$$\Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x})$$

- If the above probability is known for all K classes, we can predict the label as:

$$\hat{y} \;=\; \arg\max_{k} \; \Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x})$$

- We can write:

$$\Pr(\mathbf{Y} = k \mid \mathbf{X} = \bar{x}) = \frac{\Pr(\bar{x} \mid \mathbf{Y} = k) \Pr(\mathbf{Y} = k)}{\Pr(\mathbf{X} = \bar{x})}$$

$$= \frac{\Pr(\bar{x} \mid \mathbf{Y} = k) \Pr(\mathbf{Y} = k)}{\sum_{k'=1}^{K} \Pr(\bar{x} \mid \mathbf{Y} = k') \Pr(\mathbf{Y} = k')}$$

Some notation:

$$f_k(\bar{x}) \;=\; \Pr(\bar{x} \mid \mathbf{Y} = k) \text{ and } \quad \Pi_k = \Pr(\mathbf{Y} = k)$$

Therefore,

$$\Pr(\mathbf{Y} = k \mid \mathbf{X} = \bar{x}) \;=\; \frac{f_k(\bar{x})\Pi_k}{\sum_{k'=1}^{K} f_{k'}(\bar{x})\Pi_{k'}}$$

Some notation:

$$f_k(\bar{x}) = \Pr(\bar{x} \mid \mathbf{Y} = k) \text{ and } \Pi_k = \Pr(\mathbf{Y} = k)$$

Therefore,

$$\Pr(\mathbf{Y} = k \mid \mathbf{X} = \bar{x}) = \frac{f_k(\bar{x})\Pi_k}{\sum_{k'=1}^{K} f_{k'}(\bar{x})\Pi_{k'}}$$

$$\Pi_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}}}{N}$$

Some Notation:

$$f_k(\bar{x}) \;=\; \Pr(\bar{x} \mid \mathbf{Y} = k) \text{ and } \Pi_k = \Pr(\mathbf{Y} = k)$$

Therefore,

$$\Pr(\mathbf{Y} = k \mid \mathbf{X} = \bar{x}) \;=\; \frac{f_k(\bar{x})\Pi_k}{\sum_{k'=1}^{K} f_{k'}(\bar{x})\Pi_{k'}}$$

Depending upon different assumptions we make on f_k we get different models.

We can assume any probabilistic form on f_k . Some of the commonly used forms are:

- Gaussian
- Mixture of Gaussian
- Non-Parametric
- Naive Bayes

LDA Assumption

We will assume f_k to be

- Gaussian

$$f_k(\bar{x}) = \frac{1}{(2\pi)^{1/p}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\bar{x}-\bar{\mu}_k)^T \Sigma_k^{-1}(\bar{x}-\bar{\mu}_k)\right)$$

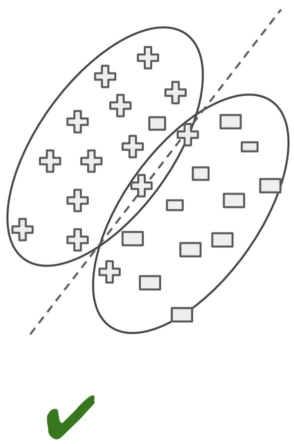
we will also assume, all covariance matrices are <u>equal</u>

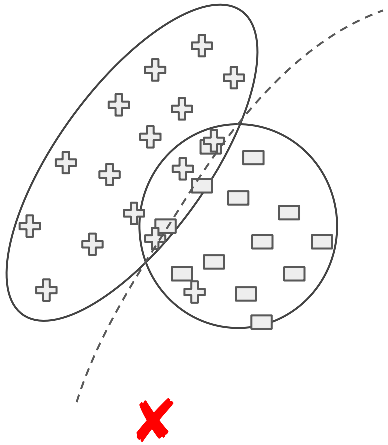
$$\forall k \;\; \Sigma_k = \Sigma$$

LDA Assumption

- The class conditional probability, $\Pr(\bar{x} \mid \mathbf{Y} = k)$, can have distinct means while sharing the same covariance matrix (shape).

<u>Implication:</u>





Class Boundary

- Class boundary between classes "k" and "l" is defined as:

$$\log\left(\frac{\Pr(\mathbf{Y}=k\,|\,\mathbf{X}=\bar{x})}{\Pr(\mathbf{Y}=l\,|\,\mathbf{X}=\bar{x})}\right) = 0$$

$$\implies \log\left(\frac{f_k(\bar{x})\Pi_k}{f_l(\bar{x})\Pi_l}\right) = \log\left(\frac{f_k(\bar{x})}{f_l(\bar{x})}\right) + \log\left(\frac{\Pi_k}{\Pi_l}\right) = 0$$

$$\implies \log\left(\frac{\Pi_k}{\Pi_l}\right) - \frac{1}{2}(\bar{\mu}_k + \bar{\mu}_l)^T\Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) + \bar{x}^T\Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) = 0$$

$$(\because\ \Sigma_k = \Sigma_l = \Sigma)$$

Class Boundary

- The boundary is a linear function in $\bar{x}$

$$\implies \log\left(\frac{\Pi_k}{\Pi_l}\right) - \frac{1}{2}(\bar{\mu}_k + \bar{\mu}_l)^T\Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) + \bar{x}^T\Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) = 0$$

$$\implies \delta_k(\bar{x}) - \delta_l(\bar{x}) = 0$$

$$\text{where, } \delta_k(\bar{x}) = \log\left(\Pi_k\right) - \frac{1}{2}\bar{\mu}_k\Sigma^{-1}\bar{\mu}_k + \bar{x}^T\Sigma^{-1}\bar{\mu}_k$$

Class Prediction

- The data point, $\bar{x}$, belongs to class "k" if

$$\Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x}) \;>\; \Pr(\mathbf{Y} = l \,|\, \mathbf{X} = \bar{x}) \,,\; \forall l \neq k$$

$$\implies \log\left(\frac{\Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x})}{\Pr(\mathbf{Y} = l \,|\, \mathbf{X} = \bar{x})}\right) > 0 \,,\; \forall l \neq k$$

$$\implies \delta_k(\bar{x}) - \delta_l(\bar{x}) > 0 \,,\; \forall l \neq k$$

Class Prediction

- The data point, $\bar{x}$, belongs to class “k” if

$$\Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x}) \;>\; \Pr(\mathbf{Y} = l \,|\, \mathbf{X} = \bar{x})\,,\; \forall\, l \neq k$$

$$\implies \log\left(\frac{\Pr(\mathbf{Y} = k \,|\, \mathbf{X} = \bar{x})}{\Pr(\mathbf{Y} = l \,|\, \mathbf{X} = \bar{x})}\right) > 0\,,\; \forall\, l \neq k$$

$$\implies \delta_k(\bar{x}) - \delta_l(\bar{x}) > 0\,,\; \forall\, l \neq k$$

- Therefore, class prediction for any data point:

$$\hat{y} = \arg\max_k \; \delta_k(\bar{x})$$

Estimating mean and covariance

$$\delta_k(\bar{x}) = \log\left(\Pi_k\right) - \frac{1}{2}\bar{\mu}_k \Sigma^{-1}\bar{\mu}_k \; + \; \bar{x}^T \Sigma^{-1} \bar{\mu}_k$$

$$\Pi_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}}}{N}$$

$$\bar{\mu}_k = \frac{\sum_{i=1}^{N} \; \mathbb{1}_{\{y_i = k\}} \, \bar{x}_k}{\sum_{i=1}^{N} \; \mathbb{1}_{\{y_i = k\}}}$$

$$\Sigma = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}} (\bar{x}_i - \bar{\mu}_k)(\bar{x}_i - \bar{\mu}_k)^T}{N - K}$$

Estimating mean and covariance

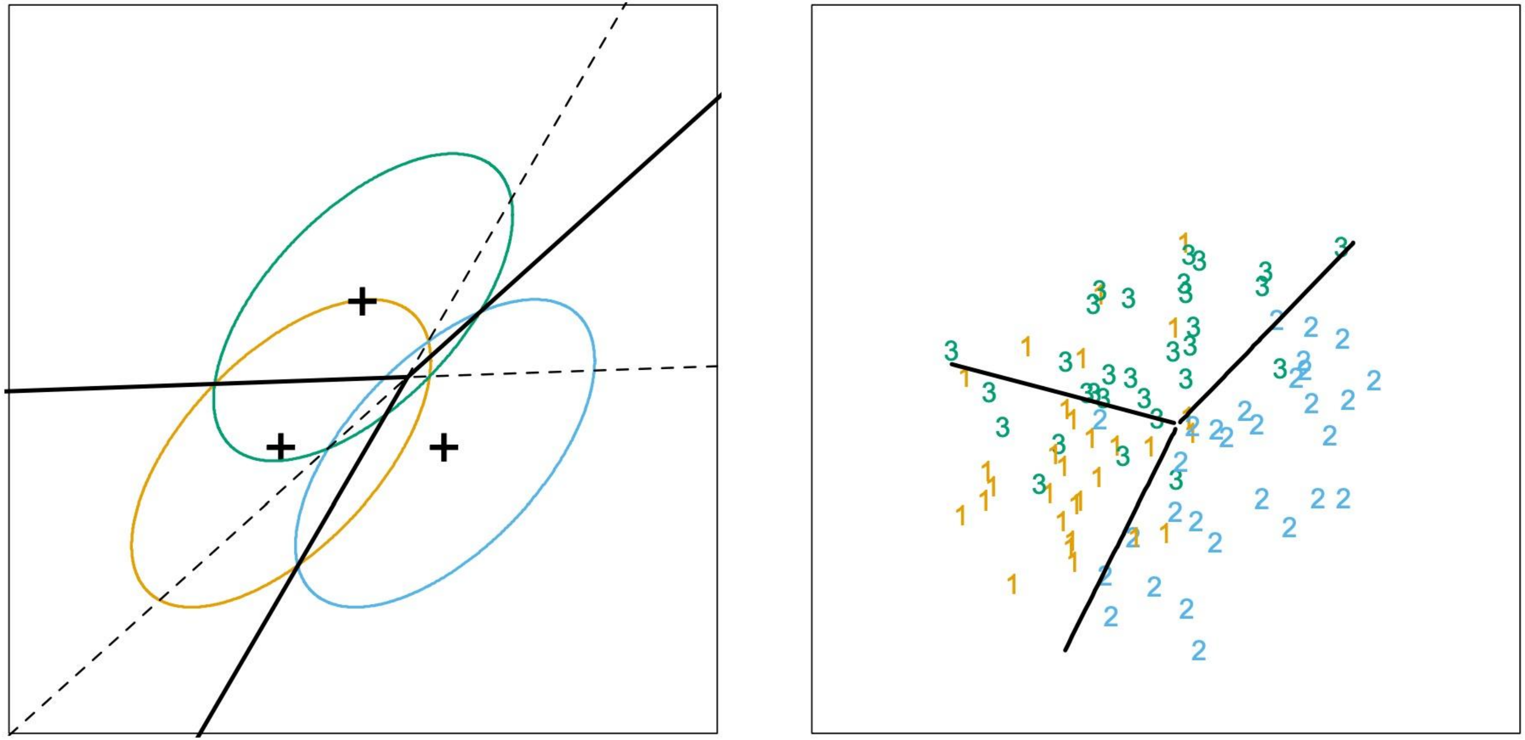
$$\delta_k(\bar{x}) = \log(\Pi_k) - \frac{1}{2}\bar{\mu}_k \Sigma^{-1} \bar{\mu}_k + \bar{x}^T \Sigma^{-1} \bar{\mu}_k$$

$$\Pi_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}}}{N} \qquad \bar{\mu}_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}}\, \bar{x}_k}{\sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}}}$$

$$\Sigma = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{1}_{\{y_i = k\}} (\bar{x}_i - \bar{\mu}_k)(\bar{x}_i - \bar{\mu}_k)^T}{N - K}$$

Pooled Estimate

LDA Example



Alternative View - Feature construction

By choosing a direction in which
means are maximally spread apart.

LDA can also be seen as maximizing the variance between the classes and minimizing the variance within the classes.

Consider a 2 class classification problem, where

$$\hat{y} = \bar{w}^T \bar{x} \qquad \begin{array}{ll} \hat{y} > w_0\,, & class\,1\,(C_1) \\ else, & class\,2\;(C_2) \end{array}$$

$\bar{\mu}_1$ and $\bar{\mu}_2$ are means of class 1 and 2 respectively.

also define projected means as $\mu_1 = \bar{w}^T \bar{\mu}_1$ and $\mu_2 = \bar{w}^T \bar{\mu}_2$

within class variance is defined as

$$s_k^2 = \sum_{i \in C_k} \left(\bar{w}^T \bar{x}_i - \bar{w}^T \bar{\mu}_k \right)^2 \quad , \text{for} \quad k \in \{1,\, 2\}$$

We want to find “**w**” vector such that **within class variance is minimized** and the **class means are maximally separated**.

$$\bar{w}^{\star} = \arg\max_{\bar{w}} J(\bar{w}) := \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

$$= \frac{\bar{w}^T S_B \bar{w}}{\bar{w}^T S_W \bar{w}}$$

$$S_B = (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T$$

$$S_W = \sum_{i \in C_1} (\bar{x}_i - \bar{\mu}_1)(\bar{x}_i - \bar{\mu}_1)^T + \sum_{i \in C_2} (\bar{x}_i - \bar{\mu}_2)(\bar{x}_i - \bar{\mu}_2)^T$$

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \; J(\bar{w}) \; := \; \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to “**w**” and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \; = \; \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \; J(\bar{w}) \; := \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to “**w**” and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \; = \; \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

are some scalar values

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \; J(\bar{w}) \; := \; \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to “**w**” and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \; = \; \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda \, S_B \, \bar{w}$$

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \ J(\bar{w}) \ := \ \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to "**w**" and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \ = \ \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

$$\implies \ S_W \, \bar{w} \ = \ \lambda \, S_B \, \bar{w}$$

$$\implies \ S_W \, \bar{w} \ = \ \lambda \, (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T \bar{w} \qquad \because \ S_B = (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T$$

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \; J(\bar{w}) \; := \; \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to "**w**" and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \; = \; \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda \, S_B \, \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda \, (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T \bar{w} \qquad \because \; S_B = (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T$$

some scalar value

$$\bar{w}^{\star} = \arg\max_{\bar{w}} \; J(\bar{w}) \; := \; \frac{\bar{w}^T \, S_B \, \bar{w}}{\bar{w}^T \, S_W \, \bar{w}}$$

taking derivative of J(**w**) w.r.t. to “**w**” and equating it to zero, we get

$$\left(\bar{w}^T S_B \, \bar{w}\right) S_W \, \bar{w} \; = \; \left(\bar{w}^T S_W \, \bar{w}\right) S_B \, \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda \, S_B \, \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda \, (\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)^T \bar{w}$$

$$\implies \; S_W \, \bar{w} \; = \; \lambda' \, (\bar{\mu}_1 - \bar{\mu}_2) \qquad \boxed{\implies \; \bar{w} \propto S_W^{-1}(\bar{\mu}_1 - \bar{\mu}_2)}$$

Alternative View

Find "**w**" vector such that **within class variance is minimized** and the **class means are maximally separated**.

$$\bar{w}^\star = \arg\max_{\bar{w}} J(\bar{w}) := \frac{\bar{w}^T S_B \bar{w}}{\bar{w}^T S_W \bar{w}}$$

$$\implies \bar{w}^\star \propto S_W^{-1}(\bar{\mu}_1 - \bar{\mu}_2)$$

The form of "**w**" vector is same

Recall, LDA's class boundary is defined as

$$\implies \log\left(\frac{\Pi_k}{\Pi_l}\right) - \frac{1}{2}(\bar{\mu}_k + \bar{\mu}_l)^T \Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) + \bar{x}^T \Sigma^{-1}(\bar{\mu}_k - \bar{\mu}_l) = 0$$

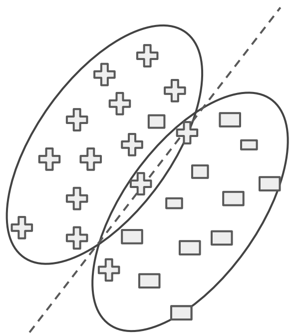
Quadratic Discriminant

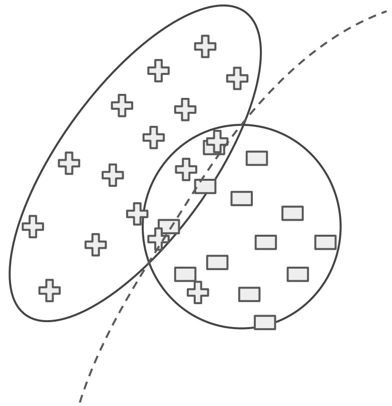
- What if the classes do not have the same covariance matrix?

- The quadratic term in the discriminant does not cancel out

- LDA -> QDA
  - Estimate covariance matrices separately

Quadratic Discriminant





Summary

- Assumes the class conditional density is a Gaussian
  - Same covariance for linear
  - Different covariance for quadratic
- Near ideal if the data comes from Gaussian distributions
  - Good approximation even if the assumption is violated
- It can be viewed as a feature selection method