Classification: Discriminant analysis

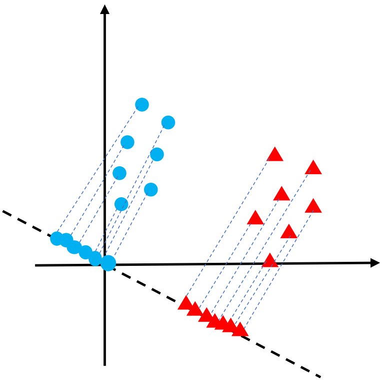
Rui Zhu

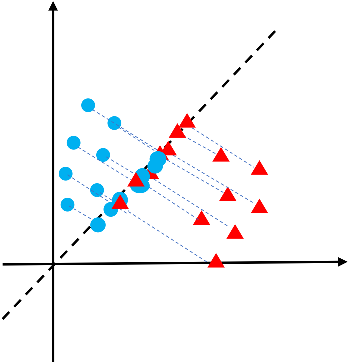
Overview

1. Fisher discriminant analysis (FDA)
2. Linear discriminant analysis (LDA)
3. Quadratic discriminant analysis (QDA)

Fisher discriminant analysis (FDA)

Project the data to a line such that the two classes are well-separated.

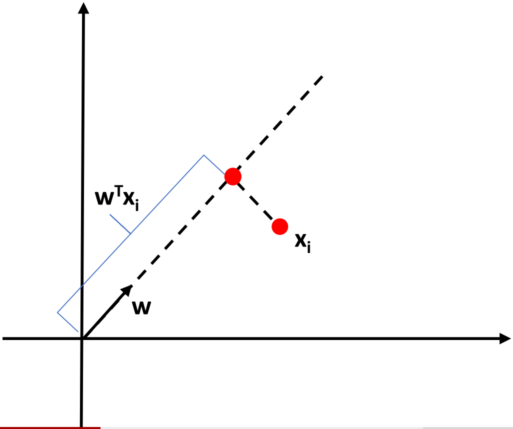




What is projection? How to find a good direction to project?

Fisher discriminant analysis (FDA)

- The projection of $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$ to the direction $\mathbf{w} \in \mathbb{R}^{p\times 1}$ ($||\mathbf{w}|| = 1$) is $\mathbf{w}^T\mathbf{x}_i \in \mathbb{R}^{1\times 1}$: projecting a p-dimensional vector to a one-dimensional subspace/ reduce the dimensions from p to 1.
- We care more about the direction of $\mathbf{w}$, not its length.

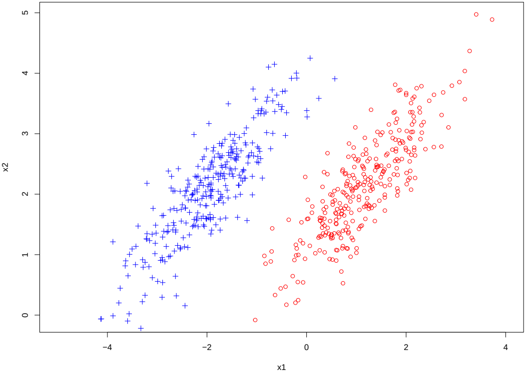


Fisher discriminant analysis (FDA)

Maximise the ratio of between-class scatter and within-class scatter.

- Scatter is similar to variance: measures the spread of data.
- We want the instances from the same class close together while those from different classes separate apart.

Fisher discriminant analysis (FDA)



Fisher discriminant analysis (FDA)

Within-class scatter:

$$S_W = \sum_{c=1}^{2} \sum_{y_i = c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T = S_{W1} + S_{W2} \in \mathbb{R}^{p \times p}$$

Between-class scatter:

$$S_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \in \mathbb{R}^{p \times p}$$

Fisher discriminant analysis (FDA)

Scatter matrices after projection on direction $\mathbf{w} \in \mathbb{R}^{p\times 1}$:

$$S_W^P = \mathbf{w}^T S_W \mathbf{w}$$

$$S_B^P = \mathbf{w}^T S_B \mathbf{w}$$

Fisher's LDA aims to solve the following problem:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Find a direction, $\mathbf{w}$, such that the ratio of between-class scatter and within-class scatter, $\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$, is maximised.

Fisher discriminant analysis (FDA)

Let's have a closer look at the objective function

$$
\begin{aligned}
\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} &= \frac{\mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w}}{\mathbf{w}^T (S_{W1} + S_{W2}) \mathbf{w}} \\
&= \frac{[\mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)][(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w}]}{\mathbf{w}^T S_{W1} \mathbf{w} + \mathbf{w}^T S_{W2} \mathbf{w}} \\
&= \frac{[\mathbf{w}^T \boldsymbol{\mu}_1 - \mathbf{w}^T \boldsymbol{\mu}_2]^2}{\mathbf{w}^T S_{W1} \mathbf{w} + \mathbf{w}^T S_{W2} \mathbf{w}} \\
&= \frac{[\mu_1^P - \mu_2^P]^2}{S_{W1}^P + S_{W2}^P}
\end{aligned}
$$

- Maximise the distance between the projected means of two classes: make the instances from different classes as separate as possible
- Minimise the within-class scatters: make the instances from the same class as close as possible

Fisher discriminant analysis (FDA)

Solving the optimisation problem, we have

$$\mathbf{w} = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in \mathbb{R}^{p\times 1}$$

- $\mathbf{w}$ is the direction of class mean difference, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, normalised by within-class scatter S_W.
- If we just use $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, we fail to consider the scatter of data. This may cause problems in classification if scatter matters.

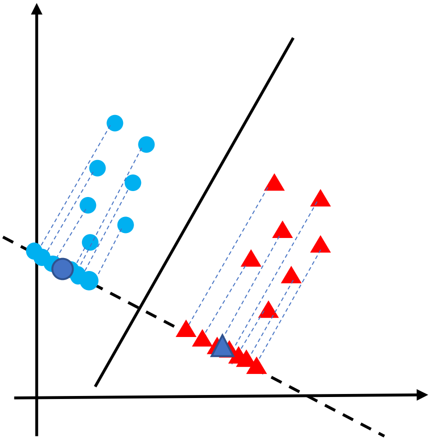
Fisher discriminant analysis (FDA)

After projecting to $\mathbf{w}$, we can classify a test instance to the closest projected class mean, which is equivalent to use the following threshold c

$$c = \mathbf{w}^T \cdot \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

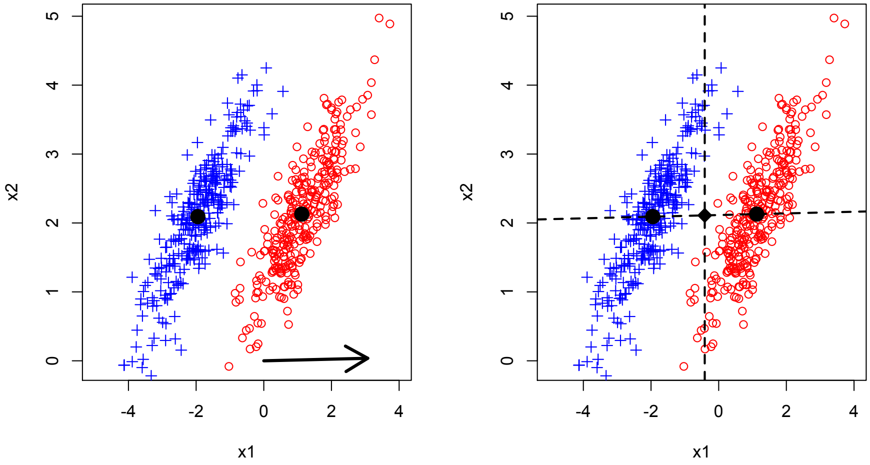
- Here c is the middle point between the projected means $\mathbf{w}^T\boldsymbol{\mu}_1$ and $\mathbf{w}^T\boldsymbol{\mu}_2$.
- We will have a linear classification boundary.

Fisher discriminant analysis (FDA)



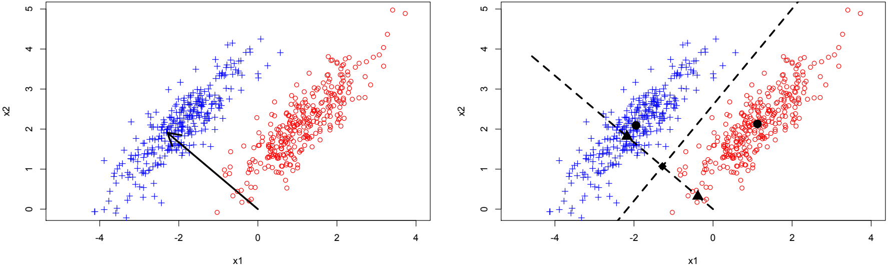
Fisher discriminant analysis (FDA)

If we just use $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$:



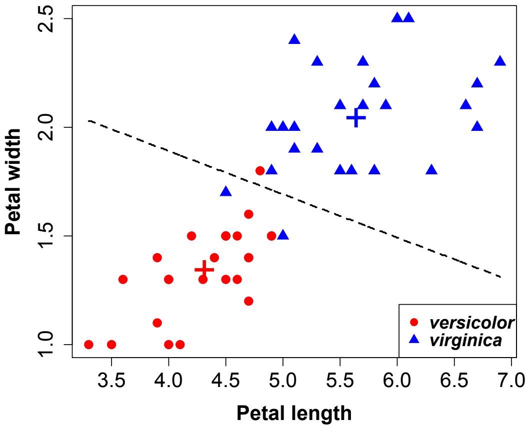
Fisher discriminant analysis (FDA)

Now use $S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in FDA, considering within-class scatter:



Fisher discriminant analysis (FDA)

Classify versicolor and virginica with two features, petal width and petal length



Fisher discriminant analysis (FDA)

Generate to C classes: we can find at most $C-1$ directions, $\mathbf{W} \in \mathbb{R}^{p\times(C-1)}$.

Within-class scatter:

$$S_W = \sum_{c=1}^{C} \sum_{y_i=c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T$$

Between-class scatter:

$$S_B = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T$$

Optimisation problem:

$$\max_{\mathbf{W}} \frac{det(\mathbf{W}^T S_B \mathbf{W})}{det(\mathbf{W}^T S_W \mathbf{W})}$$

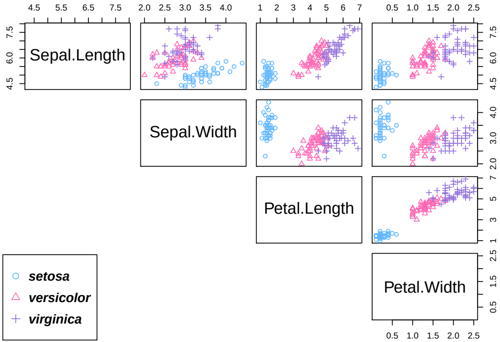
Fisher discriminant analysis (FDA)

- We can project the original data to a $(C-1)$-dimensional subspace using the matrix $\mathbf{W} \in \mathbb{R}^{p\times(C-1)}$

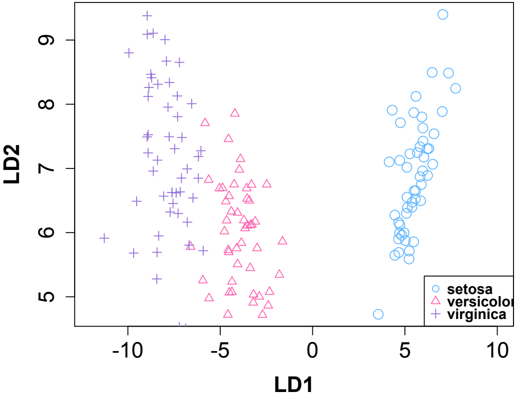
$$\mathbf{x}^T\mathbf{W} \in \mathbb{R}^{1\times(C-1)}$$

- The projected means are $\boldsymbol{\mu}_c^T\mathbf{W}$.
- Given a test instance $\mathbf{x}_t$, we first project it to this subspace $\mathbf{x}_t^T\mathbf{W}$, and then classify it to the class with the nearest projected mean.

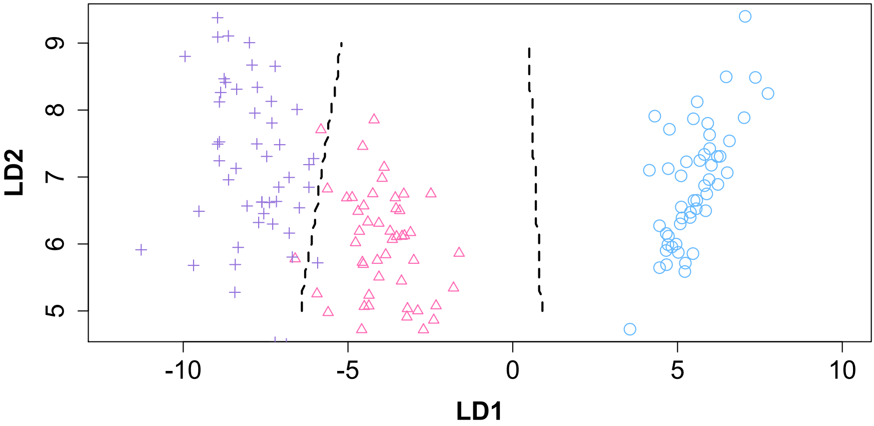
Fisher discriminant analysis (FDA)



Fisher discriminant analysis (FDA)



Fisher discriminant analysis (FDA)



Fisher discriminant analysis (FDA)

- We don't have to use all $C-1$ directions. We can coose the first r ($r \leqslant C-1$) directions for classification: tune r via cross-validation.
- Fisher's LDA can be used as a supervised dimension reduction method.

Using Bayes' Theorem for classification

Suppose there are $K \geqslant 2$ classes. The idea is to

- model each class by a distribution which provides $Pr(X = x|Y = k)$,
- use Bayes' theorem to flip these around to get $Pr(Y = k|X = x)$.

$$f_k(x) \equiv Pr(X = x|Y = k)$$

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- π_k: the prior probability of kth class, $Pr(Y = k)$
- $Pr(Y = k|X = x)$: the posterior probability of $X = x$ belonging to the kth class

Linear discriminant analysis (LDA) for $p = 1$

Now let's model each class by using normal distribution, and we start with the univariate case where $p = 1$, i.e. $N(\mu, \sigma^2)$.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right\}$$

Assuming that $\sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2$, we have

$$Pr(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu_k)^2\right\}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu_l)^2\right\}}.$$

Linear discriminant analysis (LDA) for $p = 1$

Linear discriminant function:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

We assign x to the class with the largest $\delta_k(x)$. For example, if $K = 2$ and $\pi_1 = \pi_2$, we assign x to class 1 if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise. In this case, the Bayes decision boundary corresponds to the point where

$$x = \frac{\mu_1 + \mu_2}{2}.$$

Linear discriminant analysis (LDA) for $p > 1$

Each class is assumed to have a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right).$$

The linear discriminant function becomes:

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k.$$

FDA is equivalent to LDA if the assumptions of LDA are satisfied.

Linear discriminant analysis (LDA)

- Good if the classification boundary is linear.
- Not suitable for $p > N$: it's hard to calculate scatter matrices and S_W^{-1} or $\mathbf{\Sigma}^{-1}$ in this case.

Quadratic discriminant analysis (QDA)

Assumptions:

- Normal distribution
- Each class has its own covariance matrix

The quadratic discriminant function is:

$$\delta_k(x) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| + \log \pi_k$$

Quadratic discriminant analysis (QDA)

- QDA has more parameters to estimate than LDA: QDA assumes different covariance matrices.
- QDA is more flexible than LDA.
- QDA is recommended when there is a large training set, or if the assumption of a common covariance matrix for the K classes is clearly untenable.