5 Classification and Clustering

Classification and clustering are one of the central tasks in machine learning. Given a set of data points, the purpose is to classify the points into subgroups, which express closeness or similarity of the points and which are represented by a cluster head.

5.1 Discriminant Analysis

Suppose that g populations/groups/classes $C_1, \ldots, C_g$ are given, each represented by a p.d.f. $f_i(\mathbf{x})$ on $\mathbb{R}^p$, $i = 1, \ldots, g$.

A discriminant rule divides $\mathbb{R}^p$ into disjoint regions $R_1, \ldots, R_g$, $\cup_{i=1}^p R_i = \mathbb{R}^p$. The rule is defined by:

$$\text{allocate some observation } \mathbf{x} \text{ to } C_i \text{ is } \mathbf{x} \in R_i$$

Often the p.d.f. is completely unknown or the parameters must be estimated from a training set $x_1, \ldots, x_n \in \mathbb{R}^p$ with known class allocation.

5.1.1 Fisher's Linear Discriminant Function

Fix a training set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with known classification. Let $\mathbf{x}$ be some observation. Find a linear discriminant rule $\mathbf{a}^T\mathbf{x}$ such that $\mathbf{x}$ is allocated to some class in an optimal way.

Hence, determine a linear transformation $\mathbf{a} \in \mathbb{R}^p$ such that the ration of the between-groups sum of squares and the within group sum of squares is minimized.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ be samples from g groups $C_1, \ldots, C_g$. Define $\mathbf{X}_l = [\mathbf{x}_j]_{j \in C_l}$ and $n_l = |\{j : 1 \leq j \leq n; j \in C_l\}|$. The average of the training set is given by:

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \in \mathbf{R}^p$$

The average over the group C_l is given by:

$$\overline{\mathbf{x}}_l = \frac{1}{n_l} \sum_{j \in C_l} \mathbf{x}_j \in \mathbf{R}^p.$$

Let $\mathbf{a} \in \mathbb{R}^p$ be the linear discriminant of data; we have:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{X}^T\mathbf{a} \in \mathbb{R}^n, \mathbf{y}_l = (y_j)_{j \in C_l}.$$

Similarly define the general average and between the group average as follows:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \overline{y}_l = \frac{1}{n_l}\sum_{j\in C_l} y_j.$$

Note that:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{l=1}^{g}\sum_{j\in C_l}(y_j - \overline{y}_l + \overline{y}_l - \overline{y})^2 \\
&\overset{(a)}{=} \sum_{l=1}^{g}\left[\sum_{j\in C_l}(y_j - \overline{y}_l)^2 + \sum_{j\in C_l}(\overline{y}_l - \overline{y})^2\right] \\
&= \sum_{l=1}^{g}\sum_{j\in C_l}(y_j - \overline{y}_l)^2 + \sum_{l=1}^{g} n_l(\overline{y}_l - \overline{y})^2
\end{aligned}$$

where (a) follows from a similar argument behind Steiner's rule -Theorem 3.3. $\sum_{l=1}^{g}\sum_{j\in C_l}(y_j - \overline{y}_l)^2$ is the sum of squares within groups and $\sum_{l=1}^{g} n_l(\overline{y}_l - \overline{y})^2$ is the sum of squares between groups.

Let $\mathbf{E}_n$ and $\mathbf{E}_{n_l} = \mathbf{E}_l$, $l = 1, \ldots, g$ bet centering operators. Using matrix notation, we have:

$$\begin{aligned}
\sum_{l=1}^{g}\sum_{j\in C_l}(y_j - \overline{y}_l)^2 &= \sum_{l=1}^{g} \mathbf{y}_l^T \mathbf{E}_l \mathbf{y}_l \\
&= \sum_{l=1}^{g} \mathbf{a}^T \mathbf{X}_l^T \mathbf{E}_l \mathbf{X}_l \mathbf{a} \\
&= \mathbf{a}^T (\sum_{l=1}^{g} \mathbf{X}_l^T \mathbf{E}_l \mathbf{X}_l)\mathbf{a} = \mathbf{a}^T \mathbf{W}\mathbf{a}.
\end{aligned}$$

where $\mathbf{W} = \sum_{l=1}^{g} \mathbf{X}_l^T \mathbf{E}_l \mathbf{X}_l$. Similarly:

$$\begin{aligned}
\sum_{l=1}^{g} n_l(\overline{y}_l - \overline{y})^2 &= \sum_{l=1}^{g} n_l(\mathbf{a}^T(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}}))^2 \\
&= \sum_{l=1}^{g} n_l \mathbf{a}^T(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})^T \mathbf{a} \\
&= \mathbf{a}^T \left(\sum_{l=1}^{g} n_l(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})^T\right)\mathbf{a} = \mathbf{a}^T \mathbf{B}\mathbf{a},
\end{aligned}$$

where $\mathbf{B} = \sum_{l=1}^{g} n_l(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\mathbf{l} - \overline{\mathbf{x}})^T$. Linear discriminant analysis requires:

$$\max_{\mathbf{a}\in\mathbb{R}^p} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \quad (\star)$$

Theorem 5.1. *The maximum value of* $(\star)$ *is attained at the eigenvector of* $\mathbf{W}^{-1}\mathbf{B}$ *corresponding to the largest eigenvalue.*

Proof. Assuming $\mathbf{a} = \mathbf{W}^{-1/2}\mathbf{b}$, we have

$$\max_{\mathbf{a}\in\mathbb{R}^p} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} = \max_{\mathbf{b}\in\mathbb{R}^p} \frac{\mathbf{b}^T\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{b}}{\mathbf{b}^T\mathbf{b}} = \lambda_{\max}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}),$$

where the last part results from Theorem 2.4. Furthermore $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ and $\mathbf{W}^{-1}\mathbf{B}$ have the same eigenvalues, since:

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v} = \lambda\mathbf{v} \iff \mathbf{W}^{-1/2}\mathbf{B}\mathbf{v} = \lambda\mathbf{W}^{1/2}\mathbf{v} \iff \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{W}^{1/2}\mathbf{v} = \lambda\mathbf{W}^{1/2}\mathbf{v}.$$

Therefore the two matrices have the same eigenvalues. Moreover suppose that $\mathbf{v}$ is the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to $\lambda_{\max}$. Then we have:

$$\frac{\mathbf{v}^T\mathbf{B}\mathbf{v}}{\mathbf{v}^T\mathbf{W}\mathbf{v}} = \frac{\mathbf{v}^T\mathbf{B}\mathbf{v}}{\mathbf{v}^T\mathbf{W}(\frac{1}{\lambda_{\max}}\mathbf{W}^{-1}\mathbf{B}\mathbf{v})} = \lambda_{\max}.$$

□

The linear function $\mathbf{a}^T\mathbf{x}$ is called Fisher's linear discriminant function or the first canonical variate. The ratio is invariant with the respect to scaling of $\mathbf{a}$.

Application of the linear discriminant analysis is as follows.

- Given the training set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ with known groups, compute the optimum $\mathbf{a}$ from Theorem 5.1.
- For a new observation $\mathbf{x}$, compute $\mathbf{a}^T\mathbf{x}$.
- Allocate $\mathbf{x}$ to the group with closest value of $\mathbf{a}^T\overline{\mathbf{x}}_\mathbf{l} = \overline{\mathbf{y}}_\mathbf{l}$. Discriminant rule can be formulated as follows:

  Discriminant Rule: Allocate $\mathbf{x}$ to the group l if $|\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_\mathbf{l}| \leq |\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_\mathbf{j}|$ for all $j = 1, \ldots, g$.

Fisher's discriminant function is most important in the special case of $g = 2$, where there are two groups of size n_1 and n_2 with $n = n_1 + n_2$. In this case we have:

$$\begin{aligned}
\mathbf{B} &= n_1(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}})^T + n_2(\overline{\mathbf{x}}_\mathbf{2} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\mathbf{2} - \overline{\mathbf{x}})^T \\
&= n_1(\overline{\mathbf{x}}_\mathbf{1} - \frac{n_1}{n}\overline{\mathbf{x}_\mathbf{1}} - \frac{n_2}{n}\overline{\mathbf{x}}_\mathbf{2})(\overline{\mathbf{x}}_\mathbf{1} - \frac{n_1}{n}\overline{\mathbf{x}_\mathbf{1}} - \frac{n_2}{n}\overline{\mathbf{x}}_\mathbf{2})^T + n_2(\overline{\mathbf{x}}_\mathbf{2} - \frac{n_2}{n}\overline{\mathbf{x}_\mathbf{2}} - \frac{n_1}{n}\overline{\mathbf{x}}_\mathbf{1})(\overline{\mathbf{x}}_\mathbf{2} - \frac{n_2}{n}\overline{\mathbf{x}_\mathbf{2}} - \frac{n_1}{n}\overline{\mathbf{x}}_\mathbf{1})^T \\
&= n_1(\frac{n_2}{n}\overline{\mathbf{x}_\mathbf{1}} - \frac{n_2}{n}\overline{\mathbf{x}}_\mathbf{2})(\frac{n_2}{n}\overline{\mathbf{x}_\mathbf{1}} - \frac{n_2}{n}\overline{\mathbf{x}}_\mathbf{2})^T + n_2(\frac{n_1}{n}\overline{\mathbf{x}_\mathbf{2}} - \frac{n_1}{n}\overline{\mathbf{x}}_\mathbf{1})(\frac{n_1}{n}\overline{\mathbf{x}_\mathbf{2}} - \frac{n_1}{n}\overline{\mathbf{x}}_\mathbf{1})^T \\
&= \frac{n_1 n_2^2}{n^2}(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})^T + \frac{n_2 n_1^2}{n^2}(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})^T \\
&= \frac{n_1 n_2}{n}(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})(\overline{\mathbf{x}_\mathbf{1}} - \overline{\mathbf{x}}_\mathbf{2})^T = \frac{n_1 n_2}{n}\mathbf{d}\mathbf{d}^T,
\end{aligned}$$

where $\mathbf{d} = \overline{\mathbf{x_1}} - \overline{\mathbf{x_2}}$. Therefore $\mathbf{B}$ has rank one and only one eigenvalue that is not equal to 0. Therefore $\mathbf{W}^{-1}\mathbf{B}$ has only one non-zero eigenvalue, which is given by:

$$\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B}) = \frac{n_1 n_2}{n}\mathbf{d}^T\mathbf{W}^{-1}\mathbf{d}.$$

Since $\mathbf{W}$ is nonnegative definite, the above value is nonnegative and therefore is the maximum eigenvalue. Note that $\mathbf{d}$ is an eigenvector of $\mathbf{B}$. We have:

$$\begin{aligned}(\mathbf{W}^{-1}\mathbf{B})\mathbf{W}^{-1}\mathbf{d} &= \mathbf{W}^{-1}(\frac{n_1 n_2}{n}\mathbf{d}\mathbf{d}^T)\mathbf{W}^{-1}\mathbf{d} \\ &= \frac{n_1 n_2}{n}\mathbf{W}^{-1}\mathbf{d}\left(\mathbf{d}^T\mathbf{W}^{-1}\mathbf{d}\right) \\ &= \left(\frac{n_1 n_2}{n}\mathbf{d}^T\mathbf{W}^{-1}\mathbf{d}\right)\mathbf{W}^{-1}\mathbf{d}.\end{aligned}$$

Therefore $\mathbf{W}^{-1}\mathbf{d}$ is an eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to the eigenvalue $\frac{n_1 n_2}{n}\mathbf{d}^T\mathbf{W}^{-1}\mathbf{d}$. Discriminant rule becomes:

- Allocate $\mathbf{x}$ to C_1 if $\mathbf{d}^T\mathbf{W}^{-1}(\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x_1}} + \overline{\mathbf{x_2}})) > 0$.

$\mathbf{a} = \mathbf{W}^{-1}\mathbf{d}$ is normal to the discriminating hyperplane between the classes.

Fischer's approach is distribution free. It is based on the general principle that the between-groups sum of squares is large relative to the within-groups sum of squares. This is measured by the quotient of these two quantities.

5.1.2 Gaussian ML Discriminant Rule

Maximum likelihood rule allocates observation $\mathbf{x}$ to the class C_l which maximizes the likelihood $L_l(\mathbf{x}) = \max_j L_j(\mathbf{x})$. Assume that the class distributions are Gaussian and known as $N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ with $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ fixed and with densities:

$$f_l(\mathbf{u}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_l|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{u} - \boldsymbol{\mu}_l)\right\}, \mathbf{u} \in \mathbb{R}^p.$$

The objective of ML discriminant rule would be to maximize $f_l(x)$ over l given $\mathbf{x}$.

Theorem 5.2. *The ML discriminant allocates* $\mathbf{x}$ *to class* C_l *which maximizes* $f_l(\mathbf{x})$ *over* $l = 1, \ldots, g$.

(a) If $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}$ *for all* l*, then the ML rule allocates* $\mathbf{x}$ *to* C_l *which minimizes the Mahalanobis distance:*

$$(\mathbf{x} - \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_l).$$

(b) If $g = 2$*, and* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$*, then the ML rule allocates* $\mathbf{x}$ *to the class* C_1 *if*

$$\boldsymbol{\alpha}^T(\mathbf{x} - \boldsymbol{\mu}) > 0,$$

where $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ *and* $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$.

Proof. Part (a) follows directly from the definition of ML discriminant rule. The ML discriminant finds the class l such that:

$$l = \arg \max_{1 \leq j \leq g} f_j(\mathbf{x}).$$

Since $\mathbf{\Sigma}$ is fixed for all classes, the maximization of $f_l(\mathbf{x})$ amounts to maximization of exponent which is minimization of the Mahalanobis distance. Part (b) is an exercise. □

Note that the rule (b) is analogue to Fisher's discriminant rule with parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{\Sigma}$ substituting estimates $\overline{\mathbf{x}}_\mathbf{1}$, $\overline{\mathbf{x}}_\mathbf{2}$ and $\mathbf{W}$.

Application in practice: $\mathbf{\Sigma}_l$ and $\boldsymbol{\mu}_l$ are mostly not known. One can estimate these parameters from a training set with known allocations as $\hat{\mathbf{\Sigma}}_\mathbf{l}$ and $\hat{\boldsymbol{\mu}}_l$ for $l = 1, \ldots, g$. Substitute $\mathbf{\Sigma}_l$ and $\boldsymbol{\mu}_l$ by their ML estimates $\hat{\mathbf{\Sigma}}_\mathbf{l}$ and $\hat{\boldsymbol{\mu}}_l$ and compute the ML discriminant rule.