Linear Discriminant Analysis

Multi-class classification

$$\boldsymbol{x} = (x_1, \ldots, x_k) \rightsquigarrow P(t|\boldsymbol{x}), \qquad t = 1, 2, \ldots, s$$

- We studied logistic regression.
- The Linear Discriminant Analysis (LDA) is based on Bayesian inference.

- π_t prior probability that an observation belongs to class t

  $f_t(\boldsymbol{x}) := P(\boldsymbol{x}|t)$ likelihood

  Bayes' Theorem:

  $$P(t|\boldsymbol{x}) \propto \pi_t \, f_t(\boldsymbol{x})$$

- LDA assumes

(1) $f_t(\boldsymbol{x})$ is normal, i.e., $f_t(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_t, \Sigma_t)$
(2) all the covariances are the same,
i.e., $\Sigma := \Sigma_1 = \cdots = \Sigma_s$.

- Recall D-dimensional Gaussian distribution:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is the mean and Σ is the covariance.

- We have

$$\begin{aligned}
\ln P(t|\boldsymbol{x}) &= \ln \pi_t + \ln f_t(\boldsymbol{x}) + \text{(constant)} \\
&= \ln \pi_t - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_t)^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_t) + \text{(constant)} \\
&= \ln \pi_t - \frac{1}{2}\boldsymbol{\mu}_t^\top \Sigma^{-1} \boldsymbol{\mu}_t + \boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{\mu}_t + \text{(constant)}.
\end{aligned}$$

- Define the discriminant function by

$$\delta_t(\boldsymbol{x}) := \ln \pi_t - \frac{1}{2}\boldsymbol{\mu}_t^\top \Sigma^{-1}\boldsymbol{\mu}_t + \boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{\mu}_t$$

  for $t = 1, \ldots, s$.

- Given $\boldsymbol{x}$, if $\delta_{t_*}(\boldsymbol{x})$ is the largest,

  observation $\boldsymbol{x}$ belongs to class t_* with largest probability.

  $\rightsquigarrow$ Classification

- In LDA, we use the training data to approximate $\delta_t(\boldsymbol{x})$.

Given $\mathcal{D} = \mathcal{D}_1 \sqcup \mathcal{D}_2 \sqcup \cdots \sqcup \mathcal{D}_s$ (disjoint union),

set

$$N_t := \#(\mathcal{D}_t), \quad t = 1, 2, \ldots, s, \qquad N := N_1 + \cdots + N_s.$$

We make the following estimates:

$$\begin{aligned}
\hat{\pi}_t &= N_t / N, \\
\hat{\boldsymbol{\mu}}_t &= \frac{1}{N_t} \sum_{\boldsymbol{x} \in \mathcal{D}_t} \boldsymbol{x}, \\
\hat{\Sigma} &= \frac{1}{N - s} \sum_{t=1}^{s} \sum_{\boldsymbol{x} \in \mathcal{D}_t} (\boldsymbol{x} - \boldsymbol{\mu}_t)(\boldsymbol{x} - \boldsymbol{\mu}_t)^\top.
\end{aligned}$$

An approximation of the discriminant function is given by

$$\hat{\delta}_t(\boldsymbol{x}) := \ln \hat{\pi}_t - \frac{1}{2} \hat{\boldsymbol{\mu}}_t^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_t + \boldsymbol{x}^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_t.$$

- When $\pi_1 = \pi_2 = \cdots = \pi_s$, the decision boundaries are given by

$$-\frac{1}{2}\hat{\boldsymbol{\mu}}_i^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_i + \boldsymbol{x}^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_i = -\frac{1}{2}\hat{\boldsymbol{\mu}}_j^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_j + \boldsymbol{x}^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_j$$

for $i \neq j$.