# 1 Bayes' rule-ian approach

Suppose that

- $p_g(X) = \mathbb{P}(X|Y = g)$ is the likelihood of the covariates given the class labels
- $\pi_g = \mathbb{P}(Y = g)$ is the prior

Then

$$\mathbb{P}(Y = g|X) = \frac{p_g(X)\pi_g}{\sum_{g\in\mathcal{G}} p_g(X)\pi_g} \propto p_g(X)\pi_g$$

is the Bayes rule.

# 2 Discriminant analysis

Suppose that

$$p_g(X) \propto |\Sigma|^{-1/2} e^{-(X-\mu_g)^\top \Sigma^{-1}(X-\mu_g)/2}$$

Then the log-odds between two classes $g, g'$ is:

$$\begin{aligned}\log\left(\frac{\mathbb{P}(Y = g|X)}{\mathbb{P}(Y = g'|X)}\right) &= \log\frac{p_g(X)}{p_{g'}(X)} + \log\frac{\pi_g}{\pi_{g'}} \\ &= \log\frac{\pi_g}{\pi_{g'}} - (\mu_g + \mu_{g'})^\top \Sigma^{-1}(\mu_g - \mu_{g'})/2 \\ &\quad + X^\top \Sigma^{-1}(\mu_g - \mu_{g'})\end{aligned}$$

This is linear in $X$, and hence has a linear decision boundary

## 2.1 Types of discriminant analysis

The linear discriminant function is (proportional to) the log posterior:

$$\delta_g(X) = \log\pi_g + X^\top \Sigma^{-1}\mu_g - \mu_g^\top \Sigma^{-1}\mu_g/2$$

and we assign $g(X) = \operatorname{argmin}_g \delta_g(X)$

## 2.2 Linear/regularized discriminant analysis

Now, we must estimate $\mu_g$ and $\Sigma$. If we...

- use the intuitive estimators $\hat{\mu}_g = \overline{X}_g$ (sample mean of all $X$ s.t. $Y = g$) and

$$\hat{\Sigma} = \frac{1}{n-G} \sum_{g \in \mathcal{G}} \sum_{i \in g} (X_i - \hat{\mu}_g)(X_i - \hat{\mu}_g)^\top$$

  then we have produced linear discriminant analysis (LDA)

- regularize these 'plug-in' estimates, we can form regularized discriminant analysis (Friedman (1989)). This could be (for $\lambda \in [0, 1]$):

$$\hat{\Sigma}_\lambda = \lambda \hat{\Sigma} + (1 - \lambda)\hat{\sigma}^2 I$$

# 3 LDA intuition

Intuitively, assigning observations to the nearest $\overline{X}_g$ (but ignoring the covariance) would amount to

$$\begin{aligned}
\tilde{g}(X) &= \mathrm{argmin}_g \, ||X - \overline{X}_g||_2^2 \\
&= \mathrm{argmin}_g \, X^\top X - 2X^\top \overline{X}_g + \overline{X}_g^\top \overline{X}_g \\
&= \mathrm{argmin}_g \, -X^\top \overline{X}_g + \frac{1}{2}\overline{X}_g^\top \overline{X}_g \\
&\text{compare this to:} \\
\hat{g} &= \mathrm{argmin}_g \, \underbrace{X^\top \hat{\Sigma}_\lambda^{-1} \overline{X}_g - \frac{1}{2}\overline{X}_g^\top \hat{\Sigma}_\lambda^{-1} \overline{X}_g}_{likelihood} + \underbrace{\log(\hat{\pi}_g)}_{prior}
\end{aligned}$$

The difference is we weight the distance by $\hat{\Sigma}_\lambda^{-1}$ and weight the class assignment by fraction of observations in each class.

## 3.1 Performance of LDA

The quality of the classifier produced by LDA depends on two things:

- The sample size $n$

  This determines how accurate the $\hat{\pi}_g$, $\hat{\mu}_g$, and $\hat{\Sigma}$ are

- How wrong the LDA assumptions are

  That is: $X|Y = g$ is a Gaussian with mean $\mu_g$ and variance $\Sigma$

## 3.2 The LDA variance assumption

The assumption: $\Sigma_g = \Sigma$ provides two benefits:

- Allows for estimation when $n$ isn't large compared with $Gp(p + 1)/2$

- Lowers the variance of the procedure (but produces bias)

However, when $n$ is large compared with $Gp(p+1)/2$

Then the induced bias can outweigh the variance

(This is hard to determine. Usually compare the prediction error on test set)

We relax the assumption and let $X|Y=g$ have

- mean $\mu_g$
- variance $\Sigma_g$

This makes the decision boundary quadratic