

QDA

the difference with LDA is that $\Sigma_{k'} \neq \Sigma_{k''}$ for $k' \neq k''$

$$\delta_k(x) = \log \hat{\pi}_k - \frac{1}{2} (x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) - \frac{1}{2} \log(\hat{\Sigma}_k)$$

Σ_k is different for each class

$\Rightarrow$ the quadratic term (in x) does not disappear when comparing classes

We estimate $\hat{\pi}_k$ and $\hat{\mu}_k$ ($\hat{\pi}_k$ and $\hat{\mu}_k$) as in LDA, but there is a difference with LDA, in the sense that we need to estimate $\hat{\Sigma}_k$ separately

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i: x_i \in k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$\rightarrow$ number of parameters to estimate is much larger
 $pK + K \frac{p(p-1)}{2}$

Note:

- as often in practice, in the juice example the simpler model works better (LDA better than QDA)
- LDA needs the estimation of less parameters than QDA, so we use better the information

Advantages of discriminant analysis

- method constructed for classification;
- simple computation
- good and stable results

Can be an advantage or disadvantage

- clear theoretical distribution assumption

Disadvantage

- hard to select variables
- we need to estimate many parameters (QDA)
- influence of outliers

Regularised discriminant analysis

- we create a sort of compromise between LDA and QDA

$$\hat{\Sigma}_K(\alpha) = \alpha \hat{\Sigma}_K + (1 - \alpha) \hat{\Sigma}$$

where $\alpha \in [0, 1]$ controls "the shrinkage" and can be found via cross-validation

$$\alpha = 0 \rightarrow \text{LDA}$$

$$\alpha = 1 \rightarrow \text{QDA}$$

- we can also shrink $\hat{\Sigma}$ towards $\sigma^2 \mathbf{I}$

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

where $\gamma \in [0, 1]$ has a similar role of α .

- combining the two,

$$\hat{\Sigma}_K(\alpha, \gamma) = \alpha \hat{\Sigma}_K + (1 - \alpha) [\gamma \hat{\Sigma} + (1 - \gamma) \sigma^2 \mathbf{I}]$$

Logistic regression vs LDA

$$\text{LDA} \quad \frac{\log P[G=k|X=x]}{\log P[G=0|X=x]} = \underbrace{\log \frac{\pi_k}{\pi_0} - \frac{1}{2}(\mu_k - \mu_0)^T \Sigma^{-1}(\mu_k - \mu_0)}_{\beta_0} + \underbrace{x^T \Sigma^{-1}(\mu_k - \mu_0)}_{\beta_1}$$

$$= \beta_{0k} + \beta_{1k}^T x$$

which is very similar to the logistic regression.

$$\log \frac{P[G=k|X=x]}{P[G=0|X=x]} = \beta_{0k} + \beta_{1k}^T x$$

So, where is the difference? Consider the joint distribution

$$P[G=k, X=x] = \underbrace{P[G=k|X=x]}_{\text{modelled}} \cdot \underbrace{P[X=x]}_{\text{arbitrary in logistic regression}}$$

$$\frac{e^{\beta_{0k} + \beta_{1k}^T x}}{1 + \sum_{k=1}^K e^{\beta_{0k} + \beta_{1k}^T x}}$$

LDA models it

$$P[X=x] = \sum_{k=1}^K \pi_k \phi(x; \mu_k, \Sigma)$$

↑ additional constraint ϕ