Advantages and disadvantages of Discriminant Analysis

- Advantages
  - method specifically developed for classification;
  - simplicity of calculation
  - quality and stability of the results
  - we can easily add a priori information
  - robustness wrt hypotheses (mainly Gaussianity)
- Disadvantages
  - restrictive assumptions
  - hard to do variable selection
  - large number of parameters to estimate (mainly QDA)
  - sensible to the presence of outliers

Further aspects:

- we can create a compromise between LDA and QDA, by shrinking $\hat{\Sigma}_k$ towards $\hat{\Sigma}$

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

where $\alpha \in [0;1)$ controls the amount of shrinkage (it can be found by cross-validation)

$\alpha = 0 \rightarrow$ LDA

$\alpha = 1 \rightarrow$ QDA

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

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

where $\gamma \in [0;1)$ and has the same meaning of α

- we can also combine the two,

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

Regularized logistic regression

A penalty (e.g., lasso - L_1 ...) can be added to the log-likelihood and so obtain a regularized logistic regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ -\ell(\beta) + \lambda \, \text{penalty}(\beta) \right\}$$

E.g., for lasso, $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} y_i (\beta_0 + \beta^T x) - \log\left(1 + e^{\beta_0 + \beta^T x}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$

Logistic regression versus LDA

LDA $$\log \frac{\Pr[Y=k|X=x]}{\Pr[Y=0|X=x]} = \underbrace{\log \frac{\pi_k}{\pi_0} - \frac{1}{2}(\mu_k+\mu_0)^T \Sigma^{-1}(\mu_k-\mu_0)}_{\alpha_{0k}} + \underbrace{x^T \Sigma^{-1}(\mu_k-\mu_0)}_{\alpha_{1k}}$$

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

which is really similar to

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

Which is then the difference? Consider the joint density of X and Y

$$\Pr[Y=k, X=x] = \Pr[Y=k|X=x] \Pr[X=x]$$

for both LDA and logistic regression $\frac{e^{\beta_{0k}+\beta_k^T x}}{1+\sum_{k=0}^{K-1} e^{\beta_{0k}+\beta_k^T x}}$

there are no assumptions for logistic regression

while LDA models it $\Pr[X=x] = \sum_{k=0}^{K-1} \pi_k \phi(x; \mu_k, \Sigma)$ involves the parameters

in LDA we add an additional constrain

LDA is more precise (more assumptions $\Rightarrow$ less variance)

if X is really Gaussian, there is an improvement in terms of efficiency of 30%

if the assumption is not true (often the case) the results can be strongly influenced by some observations

$\Rightarrow$ logistic regression is the safest choice

- if we have unlabelled data, x_i, we can still use them in LDA (to estimate $\hat{\Sigma}$), while should be thrown away in logistic regression

Linear regression versus LDA in the case of 2 classes

(try exercise 4.2 page 135 of the Elements of Statistical Learning)