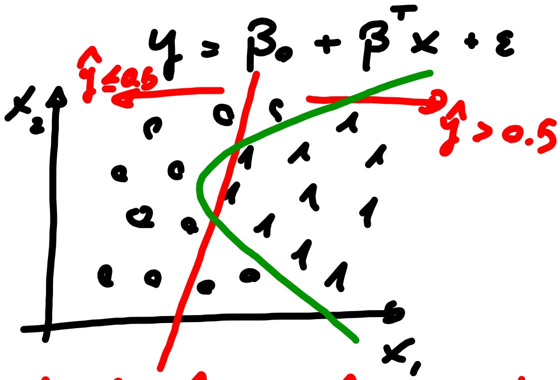
Classification via linear regression

- in logistic regression we modelled $\hat{\pi}_k(x) = \Pr[Y=k \mid X=x]$ or a transformation
- in binary case, $y \in \{0;1\}$ $\hat{y}=1 \iff \hat{\pi}_1(x) > \hat{\pi}_0(x)$
- why do not model directly y?

$$y = \beta_0 + \beta^T x + \varepsilon \qquad y \in \{0;1\}$$



discriminant boundary in $\mathbb{R}^2$ it separates the plane in two regions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = \frac{1}{2}$$

We can extend this procedure by adding non-linear functions of x (e.g., polynomials, quadratic terms) $\hat{\beta}_5 x_1 x_2$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3 x_2 + \hat{\beta}_4 x_2^2 = \frac{1}{2}$$

Note:
- we often get good results, BUT
  - it is not natural to use linear Gaussian regression for classification ($y \in \{0,1\}$, $y \notin \mathbb{R}$, $\varepsilon \not\sim \mathcal{N}(0;\sigma^2)$)
  - all inferential tools do not work (e.g., standard errors)
    - no homoskedasticity
  - $E[\varepsilon] \neq 0$, even if we add the intercept β_0
  - masking effect

Case with several categories $\quad Y \in \{0, \dots, K-1\}$

- codify each class $0, \dots, K-1$ with an indicator function

E.g. $K=3$

$$\underset{n \times K}{Y} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

1^{st} obs is in category 0
2^{nd} and 3^{rd} obs in c. 1
4^{th} " 5^{th} " " " 2

classical multivariate linear regression

$$Y = XB + E$$

X is a $n \times p$ design matrix
B is a $p \times k$ matrix of coefficients

so that

$$\hat{Y} = X \underbrace{(X^T X)^{-1} X^T Y}_{B}$$

new observation x_0^T

$$\hat{Y}^0 = (\hat{Y}^0_0, \hat{Y}^0_1, \hat{Y}^0_2)$$

is assigned to the class of larger $\hat{Y}^0_k$

Note:
- we have that $\sum_{k=0}^{K-1} \hat{Y}^0_k = 1$, but there is no guarantee that $\hat{Y}^0_k$, $k = 0, \dots, K-1$ $\in [0;1]$

$$\Downarrow$$

$\hat{Y}^0_k$ is not a good estimate of $\pi_k(x)$

Important issue : masking effect

(see figures 4.2 and 4.3 on pages 105 and 106 of ESL)

Possible solution: use in case of $K=3$ a quadratic term

in general: use polynomials of degree $K-1$

Discriminant Analysis

Define: Y the categorical variable which tell us to which category an observation belong

X a p-dimensional random variable

Goal: find the probability $\Pr[Y=k \mid X=x]$ $\pi_k(x)$

Suppose that the population is divided in K subpopulations each of them with conditional density

$$p_k(x) = \Pr[X=x \mid Y=k] \quad , \; k=0,\dots,K-1$$

Denote with π_k the probability of belonging to the subpop. k

$$\pi_k = \Pr[Y=k]$$ a priori probability of belonging to class k

Then the marginal density of the whole population is

$$p(x) = \Pr[X=x] = \sum_{k=0}^{K-1} \Pr[X=x \mid Y=k]\Pr[Y=k] = \sum_{k=0}^{K-1} p_k(x)\,\pi_k$$

Then, using the Bayes theorem

$$\Pr[Y=k \mid X=x] = \frac{\Pr[X=x \mid Y=k]\Pr[Y=k]}{\Pr[X=x]} = \frac{p_k(x)\,\pi_k}{\sum_{m=0}^{K-1} p_m(x)\,\pi_m}$$

$$\Pr[X=x] = \sum_{m=0}^{K-1} \Pr[X=x \mid Y=m]\Pr[Y=m]$$

To estimate $\Pr[Y=k \mid X=x]$ we only need to estimate $p_k(x)$ and $\pi_k(x)$, $k=0,\dots,K-1$

- $\pi_k(x)$ is the prior probability, so it is natural to estimate it as $\frac{n_k}{n}$ n_k is # obs belonging to class k, n is total # obs.

Instead, we have several options to estimate $p_k(x)$

- parametric (LDA, QDA)
- non-parametric (KNN)

Since we want to assign a new observation x_0 to the class with the highest posterior probability, it is useful to consider $\Pr[Y=k \mid X=x]$

$$\log \frac{\Pr[Y=k \mid X=x]}{\Pr[Y=m \mid X=x]} = \log \frac{p_k(x)\,\pi_k}{\sum_{l=1}^{K} p_l(x)\,\pi_l} - \log \frac{p_m(x)\,\hat{\pi}_m}{\sum_{l=1}^{K} p_l(x)\,\hat{\pi}_l}$$

$$= \log \frac{\hat{\pi}_k}{\hat{\pi}_m} + \log \frac{p_k(x)}{p_m(x)}$$

and the discriminant function

$$d_k(x_0) = \log \hat{\pi}_k + \log p_k(x_0)$$

we will assign x_0 to the class $\underset{k}{\operatorname{argmax}}\, d_k(x_0)$

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)

LDA and QDA use multivariate Gaussian distribution for $p_k(x)$

$$p_k(x) = \overbrace{\boxed{\frac{1}{(2\pi)^{p/2}\det(\Sigma_k)^{1/2}}}}^{c_k \xrightarrow{\text{LDA}} c} \exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}$$

In particular, LDA assumes $\Sigma_k = \Sigma \quad \forall k = 0, \dots, K-1$

As a consequence

it is a scalar, so $= \mu_k^T \Sigma^{-1} x$

$$d_k(x) = \log \pi_k - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)$$

$$= \log \pi_k - \frac{1}{2}\Big(\underbrace{x\Sigma^{-1}x}_{\text{does not depend on } k} + \mu_k^T\Sigma^{-1}\mu_k - 2x^T\Sigma^{-1}\mu_k\Big)$$

$$= \log \pi_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k$$

which is linear in x $\longrightarrow$ LDA

Compare two classes in terms of log-ratio

$$\log \frac{\Pr[Y=k \mid X=x]}{\Pr[Y=m \mid X=x]} = \log \frac{\cancel{c}\exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\}\pi_k / \cancel{p(x)}}{\cancel{c}\exp\{-\frac{1}{2}(x-\mu_m)^T\Sigma^{-1}(x-\mu_m)\}\pi_m / \cancel{p(x)}}$$

$$= \log \frac{\pi_k}{\pi_m} - \frac{1}{2}\left(\cancel{x^T\Sigma x} - 2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k - \cancel{x^T\Sigma x} + 2x^T\Sigma^{-1}\mu_m + \mu_m^T\Sigma^{-1}\mu_m\right.$$

$$= \log \frac{\pi_k}{\pi_m} - \frac{1}{2}(\mu_k-\mu_m)^T\Sigma^{-1}(\mu_k-\mu_m) + 2x^T\Sigma^{-1}(\mu_k-\mu_m)$$

To compute all these quantities ($d_k(x)$, r_k), we need to plug-in the estimates of π_k, μ_k and Σ

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\Sigma} = \frac{1}{n-K}\sum_{k=0}^{K-1}\sum_{i:y_i=k}(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

parameters to estimate $pK + \frac{p(p+1)}{2}$

μ_k is p-dimensional

To have curved boundaries, we should add quadratic terms

$$x_i = (x_{i1}, x_{i2}) \longrightarrow x_i = (x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2, x_{i1}x_{i2})$$

- with $K=2$, there is no big differenceses in using LDA or linear regression
- with $K>2$, substantial differences (LDA does not suffer from the masking effect issue)

QDA $\rightarrow$ we remove the condition $\Sigma_k = \Sigma \ \forall k$

$$d_k(x) = \log \pi_k - \frac{1}{2}\underbrace{(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}_{} - \frac{1}{2}\log(\det \Sigma_k^{-1})$$

we cannot remove the quadratic term

$x^T \Sigma_k^{-1} x$

QDA

To estimate $d_k(x)$, the same estimates for $\hat{\pi}_k$ and $\hat{\mu}_k$, but

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

#parameters to estimate
$pK + \frac{p(p+1)}{2} K$

Note:
- as often, the simpler model (LDA) performs better than the more complex (QDA)
- LDA needs the estimation of fewer parameters (better use of the information)