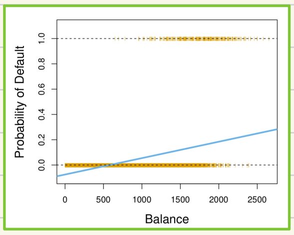
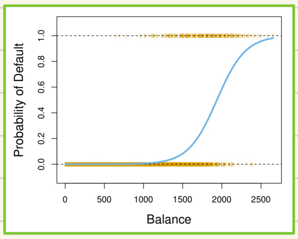
Logistic regression

Why don't we use linear regression model

- We want our estimates to remains b/w $[0,1]$ interval, which provide more meaningful estimates of $Pr(Y|X)$

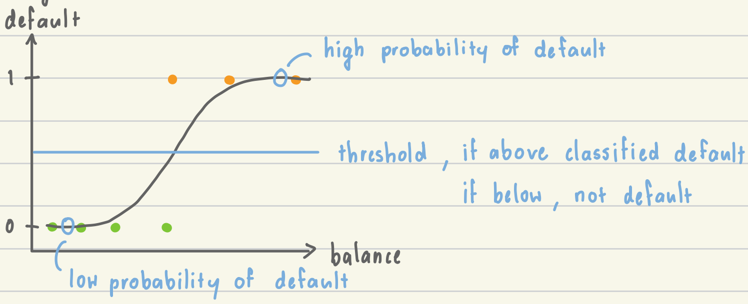


← takes on negative probability
← use least squares estimation



← outputs b/w 0 and 1
← use maximum likelihood estimation

- We are trying to model the probability that Y belongs to a particular category
  e.g. $Pr(\text{default} = \text{Yes} \mid \text{balance})$



Logistic model :

- linear model $p(X) = \beta_0 + \beta_1 X$
- find function that gives output b/w 0 and 1 (logistic function)

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

↳ transform the function

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

logit or log-odds ratio

- interpretation – ↑ X by one unit changes the log odds by β_1 or multiple odds by e^{β_1}

Maximum Likelihood estimation

$$\ell(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}$$

$$\log \prod_{i=1}^{n} f(x) = \sum_{i=1}^{n} \log f(x)$$

$$\log \ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \log [p(x_i)^{y_i}] + \log \{[1 - p(x_i)]^{1-y_i}\}$$

$$\log(x^n) = n \log(x)$$

$$= \sum_{i=1}^{n} y_i \log[p(x_i)] + (1-y_i) \log[1 - p(x_i)]$$

$$\frac{1+\exp(\eta)}{1+\exp(\eta)} - \frac{\exp(\eta)}{1+\exp(\eta)}$$

we have $p(x_i) = \exp(\eta) / 1 + \exp(\eta)$, where $\eta = \beta_0 + \beta_1 x_i$

$$= \sum_{i=1}^{n} y_i \log[\exp(\eta)/1+\exp(\eta)] + (1-y_i) \log \{1 - [\exp(\eta)/1+\exp(\eta)]\}$$

$$= \sum_{i=1}^{n} y_i \{\overbrace{\log[\exp(\eta)]}^{\eta} - \log[1+\exp(\eta)]\} +$$

$$(1-y_i) \log \{[1+\exp(\eta) - \exp(\eta)] / [1+\exp(\eta)]\}$$

$$= \sum_{i=1}^{n} y_i \{\eta - \log[1+\exp(\eta)]\} + (1-y_i) \log\{1/[1+\exp(\eta)]\}$$

$$= \sum_{i=1}^{n} y_i \{\eta - \log[1+\exp(\eta)]\} + (1-y_i) \{\overbrace{\log(1)}^{0} - \log[1+\exp(\eta)]\}$$

$$= \sum_{i=1}^{n} y_i \{\eta - \log[1+\exp(\eta)]\} - (1-y_i) \log[1+\exp(\eta)]$$

$$= \sum_{i=1}^{n} y_i \eta - y_i \log[1+\exp(\eta)] - \log[1+\exp(\eta)] + y_i \log[1+\exp(\eta)]$$

$$= \sum_{i=1}^{n} y_i \eta - \log[1+\exp(\eta)]$$

Note: there is no closed-form solution when setting derivatives w.r.t parameters equal zero

Making prediction:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

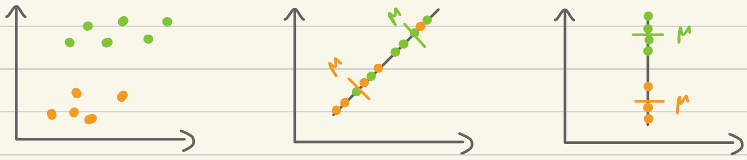
$R^2 = 1 -$ Model deviance / Null model deviance

- watch out for warning, when the data is perfectly separated

Linear Discriminant Analysis

Assumptions :

- distribution of the predictors is a multivariate normal (samples come from normal populations)
- with the same variance - covariance matrix
- like PCA , it produces low - dimensional projection of the data
- However, LDA finds linear combination of predictors that maximise the separation b/w classes



(from BrendiA Github)

- maximise the distance b/w group means and minimise variation within each group

Bayes Theorem :

- The probability of x belong to class k

$$Pr(Y = k \mid X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_k f_k(x)}$$

large = more likely that x belongs to class k

small = unlikely that x belongs to class k

π_k is the prior probability that an observation comes from class k

$f_k(x)$ is the density function for predictor x for class k

LDA with $p = 1$:

$K = 2$, two classes with same prior probability

assign x_0 to class A if $x_0 > \frac{\bar{x}_A + \bar{x}_B}{2}$

- Based on assumption 1 : $f_k(x)$ is a univariate normal

$$f_k(x) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

μ_k is mean of class k

σ_k^2 is variance of class k

- Based on assumption 2 : $\sigma_1^2 = \sigma_2^2 = \quad = \sigma_k^2$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}$$

- A simplification of $p_k(x_0)$ yields the discriminant functions :

$$\delta_k(x_0) = x_0 \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

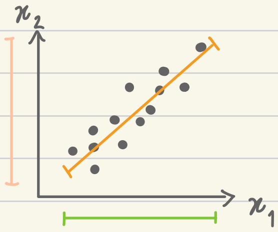
LDA for $p > 1$ (Multivariate LDA) :

- variable X has a multivariate normal distribution with mean μ and variance - covariance Σ , $X \sim N(\mu, \Sigma)$

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

x, μ are p-dimensional vector
Σ is a $p \times p$ variance - covariance matrix

Variance - covariance matrix (from BrendiA Github)



$$\Sigma = \begin{bmatrix} var(x_1) & cov(x_1, x_2) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$$

$$var(x_1) = \frac{1}{n}\sum_{i=1}^{n}(x_1 - \mu_{x_1})^2$$

$$var(x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_2 - \mu_{x_2})^2$$

$$cov(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \mu_{x_1})(x_{i2} - \mu_{x_2})$$

- Discriminant functions for Multivariate LDA is :

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

- Assign observation x_0 to class A if

$$x_0^T \Sigma^{-1}(\mu_A - \mu_B) > \frac{1}{2}(\mu_A + \mu_B)^T \Sigma^{-1}(\mu_A - \mu_B)$$

We need to find μ_k and Σ

- $\bar{x}_k$ or sample mean for μ_k
- pooled variance - covariance for Σ

$$S = \frac{n_1 S_1 + n_2 S_2 + \ldots + n_k S_k}{n_1 + n_2 + \ldots + n_k}$$