

Logistic regression

- linear (Gaussian) regression: $\varepsilon \sim \mathcal{N}(0; \sigma^2)$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} (y - f(x))^2\right\}$$

- logistic regression $\rightarrow$ binomial distribution

$$f_G(g) = \binom{1}{g} \pi(x)^g (1 - \pi(x))^{1-g} \quad \pi(x) = \Pr[G=1|X=x]$$

$$\pi(x) \in [0; 1] \rightarrow \frac{e^{\eta(x)}}{1 + e^{\eta(x)}} \quad \text{logistic function}$$

$$\eta(x) = \beta_0 + \beta^T x \quad \text{linear predictor}$$

$$\hat{\pi} = \frac{e^{\eta}}{1 + e^{\eta}} \quad \text{base (4.18)}$$

$$\hat{\pi} + \hat{\pi} e^{\eta} = e^{\eta}$$

$$e^{\eta} (1 - \hat{\pi}) = \hat{\pi} \quad \text{adds}$$

$$e^{\eta} = \frac{\hat{\pi}}{1 - \hat{\pi}} \quad \text{logit}$$

$$\eta = \log \frac{\hat{\pi}}{1 - \hat{\pi}} \Rightarrow \log \frac{\hat{\pi}}{1 - \hat{\pi}} = \beta_0 + \beta^T x$$

$$\text{book (4.17)}$$

the logistic regression ^{model} is a special case of the Generalised Linear Models (GLM)

$$h(E[Y|x]) = \eta(x)$$

link function, in this case logit*

in R, to fit the logistic regression, we use the command `glm`,
with `family = binomial (link = logit)`

* it is not mandatory to use the logit link, for example an alternative is the probit

Simple example (Brazilian bank example)

G = satisfaction of a customer (low/high)

X = age group (young/old)

$$\pi = \Pr(G_2 \text{ high})$$

		Y = satisfaction		
		high	low	
X = age	old	157	34	191
	young	225	84	309
		382	118	500

$$\hat{\pi}_{x=\text{old}} = \frac{157}{191} \approx 0.822$$

$$\hat{\pi}_{x=\text{young}} = \frac{225}{309} \approx 0.728$$

$$\eta(x) = \beta_0 + \beta_1 x$$

$$\hat{\beta}_0 = 1.5299$$

$$\hat{\beta}_1 = -0.5446 \leftarrow \text{log odds ratio}$$

$$e^{-0.5446} = \text{odds-ratio}$$

$$\text{old: } \eta(x=0) = 1.5299 - 0.5446 \cdot 0 = 1.5299$$

$$\hat{\pi} = \frac{e^{\eta}}{1 + e^{\eta}} \quad \hat{\pi}_{x=\text{old}} = \frac{e^{1.5299}}{1 + e^{1.5299}} \approx 0.822$$

$$\text{young: } \eta(x=1) = 1.5299 - 0.5446 \cdot 1 = 0.9853$$

$$\hat{\pi}_{x=\text{young}} = \frac{e^{0.9853}}{1 + e^{0.9853}} \approx 0.728$$

Why log-odds ratio

$$\hat{\beta}_0 = \log \frac{\hat{\pi}_{\text{old}}}{1 - \hat{\pi}_{\text{old}}}$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \frac{\hat{\pi}_{\text{young}}}{1 - \hat{\pi}_{\text{young}}}$$

$$\hat{\beta}_1 = \log \frac{\hat{\pi}_{\text{young}}}{1 - \hat{\pi}_{\text{young}}} - \log \frac{\hat{\pi}_{\text{old}}}{1 - \hat{\pi}_{\text{old}}}$$

$$= \log \left(\frac{\frac{\hat{\pi}_{\text{young}}}{1 - \hat{\pi}_{\text{young}}}}{\frac{\hat{\pi}_{\text{old}}}{1 - \hat{\pi}_{\text{old}}}} \right) \leftarrow \text{log odds ratio}$$

If x is continuous

$$\hat{\beta}_1 = \hat{\beta}_0 + \hat{\beta}_1(x+1) - \hat{\beta}_0 - \hat{\beta}_1 x$$

$$= \log\left(\frac{\hat{\pi}_{x+1}}{1-\hat{\pi}_{x+1}}\right) - \log\left(\frac{\hat{\pi}_x}{1-\hat{\pi}_x}\right)$$

expected change in terms of log-odds

$$\hat{\beta}_1 < 0$$

$$\hat{\beta}_1 > 0$$

2-test
to see if
 β_1 is
significantly
different
from 0

$$e^{\hat{\beta}_1} = \frac{\frac{\hat{\pi}_{x+1}}{1-\hat{\pi}_{x+1}}}{\frac{\hat{\pi}_x}{1-\hat{\pi}_x}}$$

odds ratio

← expected change in the odds

For classification purposes , $\hat{\pi} > 0.5 \rightarrow \hat{G}=1$
 $\hat{\pi} \leq 0.5 \rightarrow \hat{G}=0$

~~Correspondence of deviance is~~ $-2 \log$ like

Gaussian case : $\sum (y - f(x))^2$

logistic regression : $\log(\hat{\pi}(x)^g (1 - \hat{\pi}(x))^{1-g})$

$$g \log \hat{\pi}(x) + (1-g) \log (1 - \hat{\pi}(x)) \quad \hat{\pi}(x) = \frac{e^{\eta}}{1+e^{\eta}}$$

$$\eta(x) = \beta_0 + \beta^T x$$

$$g \log \frac{e^{\eta}}{1+e^{\eta}} + (1-g) \log \left(\frac{1}{1+e^{\eta}} \right)$$

$$+ g \log(1+e^{\eta})$$

$$g(\beta_0 + \beta^T x) - g \log(1+e^{\eta}) + \log\left(\frac{1}{1+e^{\eta}}\right) - g \log\left(\frac{1}{1+e^{\eta}}\right)$$

$$\log \text{lik} \rightarrow \underline{g(\beta_0 + \beta^T x) - \log(1+e^{\beta_0 + \beta^T x})}$$

penalised logistic regression

$$\max_{\beta_0, \beta} \left\{ \sum_{i=1}^n \left[g_i (\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^p |\beta_j| \right\}$$

↑
Lasso

$$\text{for ridge, } -\lambda \sum_{j=1}^p \beta_j^2$$

$p = \#$ of covariates

Discriminant Analysis

- G categorical response variable (which class my statistical unit belongs to)
- X p -dimensional vector of covariates

Goal: find the probability $\Pr[G=k | X=x]$ ← posterior

The total population is made by K subpopulation with conditional distribution

$$\Pr[X=x | G=k] = f_k(x) \quad \text{density of } X \text{ if } X \in \text{class } k \\ k=1, \dots, K$$

The probability of belonging to the k -th subpopulation is

$$\Pr[G=k] = \pi_k$$

Then, using the Bayes theorem $P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B]}$

$$\Pr[G=k | X=x] = \frac{\Pr[X=x | G=k] \cdot \Pr[G=k]}{\Pr[X=x]}$$

$$\text{where } \Pr[X=x] = \sum_{k=1}^K \Pr[X=x | G=k] \Pr[G=k]$$

$$\text{so } \Pr[G=k | X=x] = \frac{f_k(x) \pi_k}{\sum_{e=1}^K f_e(x) \pi_e}$$

Then we "just" need to estimate π_k and $f_k(x)$

- π_k is the prior probability, so it is natural to use $\frac{n_k}{n}$ $n_k = \# \text{ observations in class } k$
- several options to compute $f_k(x)$ $n = \# \text{ total}$
 - parametric (LDA, QDA)
 - non-parametric (K-NN)

Since we are interested to find the highest $\Pr[G=k | X=x_0]$, we will compare

$$\log \frac{\Pr[G=k | X=x_0]}{\Pr[G=e | X=x_0]} = \log \frac{\pi_k}{\pi_e} + \log \frac{f_k(x)}{f_e(x)}$$

The discriminant function $\delta_k(x)$ in this case is

$$\delta_k(x_0) = \log \pi_k + \log f_k(x)$$

$$Pr[G=k | X=x] = \frac{f_k(x) \pi_k}{\sum_{e=1}^K f_e(x) \pi_e}$$

LDA and QDA uses the Gaussian distribution for $f_k(x)$

$$f_k(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma_k)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right\}$$

In particular, LDA assumes $\Sigma_k = \Sigma$. As a consequence,

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \left(\underbrace{x^T \Sigma^{-1} x}_{\text{same for all classes}} + \mu_k^T \Sigma^{-1} \mu_k - 2 x^T \Sigma^{-1} \mu_k \right)$$

⇒ we do not need to take it into consideration for our classification task

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \left(\mu_k^T \Sigma^{-1} \mu_k - 2 x^T \Sigma^{-1} \mu_k \right) \rightarrow \text{is linear in } x$$

When comparing two classes

$$\log \frac{Pr[G=k | X=x]}{Pr[G=e | X=x]} = \log \frac{\pi_k}{\pi_e} - \frac{1}{2} (\mu_k - \mu_e)^T \Sigma^{-1} (\mu_k - \mu_e) + x^T \Sigma^{-1} (\mu_k - \mu_e)$$

Plugging in the estimates for $\hat{\pi}_k$, $\hat{\pi}_e$, $\hat{\mu}_k$, $\hat{\mu}_e$ and $\hat{\Sigma}$

$$\begin{aligned} \hat{\pi}_k &= \frac{n_k}{n} & \hat{\Sigma} &= \frac{1}{n-K} \sum_{k=1}^K \sum_{i: g_i=k} (x_i - \mu_k)(x_i - \mu_k)^T \\ \hat{\mu}_k &= \frac{1}{n_k} \sum_{i: g_i=k} x_i \\ \hat{\pi}_e &= \frac{n_e}{n}, \quad \hat{\mu}_e = \frac{1}{n_e} \sum_{i: g_i=e} x_i \end{aligned}$$

Note:

- we can have curved boundaries by adding quadratic terms
- with only two classes, there is a correspondence between LDA and classification viz linear regression
- for $K > 2$ there are substantial differences (in particular, LDA ^{suffers less} ~~does not suffer from~~ the masking effect)