Logistic regression (Bernoulli outcome)

$(Y_i \mid p_i \, \mathcal{B}) \stackrel{I}{\sim} \text{Bernoulli}(p_i)$ (1)

(1 if diabetes, 0 not) $(i = 1, \ldots, n)$ ← 768

$y = (y_1, \ldots, y_n)$

$(\mathcal{L}:) \quad \log\left(\frac{p_i}{1 - p_i}\right) = \sum_{j=1}^{k} \beta_j x_{ij}$ (2)

$\underset{k \times 1}{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$

$p(\beta \mid \mathcal{L} \, \mathcal{B})$ ← prior for β

likelihood function

$\ell(\beta \mid y \, X \, \mathcal{L} \, \mathcal{B}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$

$= \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i)$

$\underset{n \times k}{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$

log-likelihood function

$\ell\ell(\beta \mid y \, X \, \mathcal{L} \, \mathcal{B}) = \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]$

some algebra reveals that

$\log\left(\frac{p_i}{1 - p_i}\right) = L_i \leftrightarrow p_i = \frac{1}{1 + \exp(-L_i)}$

Bayes' Theorem

$$p(\beta \mid y, X, \alpha, \beta) =$$

$$c \; p(\beta \mid \alpha, \beta) \cdot \ell(\beta \mid y, X, \alpha, \beta)$$

log rule

log posterior

$$\log p(\beta \mid y, X, \alpha, \beta) = c + \log \text{likelihood } \ell(\beta \mid y, X, \alpha, \beta) + \log p(\beta \mid \alpha, \beta)$$

log prior

Consider, eg., a Normal prior for β

$$p(\beta \mid \alpha, \beta) \overset{N}{\sim} N_k[\mu_\beta, \Sigma_\beta]$$

$$p(\beta \mid \alpha, \beta) = c \exp\left[-\tfrac{1}{2}(\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta)\right]$$

take $\Sigma_\beta = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_k^2 \end{bmatrix}$ and $\mu_\beta = 0$

Then

$$\log p(\beta \mid \tau, X, y) = -\frac{1}{\tau} \sum_{j=1}^{k} \left(\frac{\beta_j}{\sigma_j}\right)^2$$

for some $\tau > 0$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\|\beta\|^2 = \beta^T \beta = \sum_{j=1}^{k} \beta_j^2$$

log posterior

$$\log p(\beta \mid \tau, X, y) = \ell(\beta \mid \tau, X, y)$$

log likelihood

$$-\frac{1}{\tau} \sum_{j=1}^{k} \left(\frac{\beta_j}{\sigma_j}\right)^2$$

← log prior

Consider estimating β not by maximizing the log likelihood but by maximizing the log posterior: result is called $\hat{\beta}_{MAP}$

maximum a posteriori →

but look what the prior is doing: it's

shrinking the $\hat{\beta}_{MLE}$ vector toward $\underset{\sim}{0}$

Why might this be a good idea?

imagine k (# predictors) growing with n (# observations) fixed

turns out (makes sense)

that $\hat{\underset{\sim}{\beta}}_{MLE}$ becomes more & more

unstable, in the sense that

$\hat{SE}\left[(\hat{\beta}_{MLE})_j\right]$ ↑ sharply

$$\left[\begin{array}{c|c} y & \hat{y} \\ \hline y_1 & \hat{y}_1 \\ \vdots & \vdots \\ y_n & \hat{y}_n \end{array}\right]$$

general prediction problem

a natural way to measure how good $\hat{y}$ is in predicting y is with $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

In terms of random variables, the square of RMSE is $MSE(\Sigma, \hat{\Sigma}) = E\left[(\Sigma - \hat{\Sigma})^2\right]$

---

fact: $MSE = \left[bias(\hat{\Sigma})\right]^2 + V(\hat{\Sigma})$

Excellent idea: if we deliberately bias $\underset{\sim}{\beta}_{MLE}$ by shrinking it toward $\underset{\sim}{0}$, the variance of the resulting $\underset{\sim}{\hat{\beta}}_{MAP}$ may go down more than the $\left[bias(\underset{\sim}{\hat{\beta}}_{MAP})\right]^2$ goes up, yielding a reduction in (R)MSE

This turns out to work better & better as $k \uparrow n$