第9章 统计模型应用

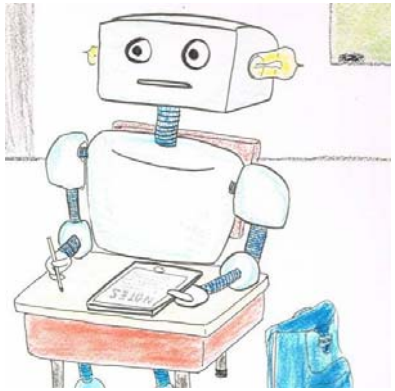
感谢清华大学自动化系
江瑞教授提供PPT

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

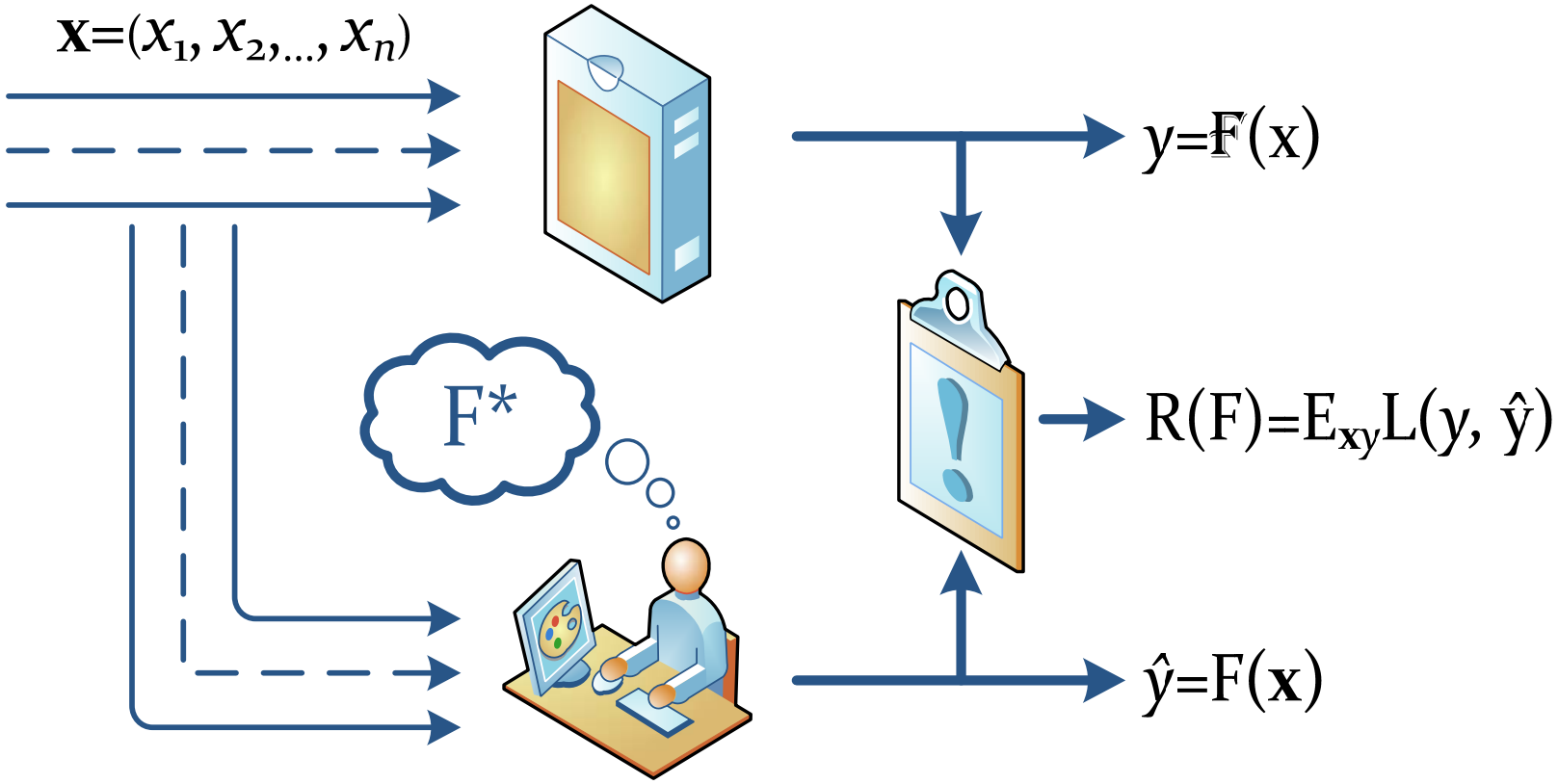
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



Binary Classification

Supervised learning



Continuous response — Regression

$$\mathrm{E}\, Y_i = f(\mathbf{x}_i)$$

Linear regression

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

Polynomial regression

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_i^j + \varepsilon_i$$

Other regression models

$\ldots$

Categorical response — Classification

$$P(Y_i = c_l) = f(\mathbf{x}_i)$$

Binary classification

$$P(Y_i = 1) = f(\mathbf{x}_i)$$

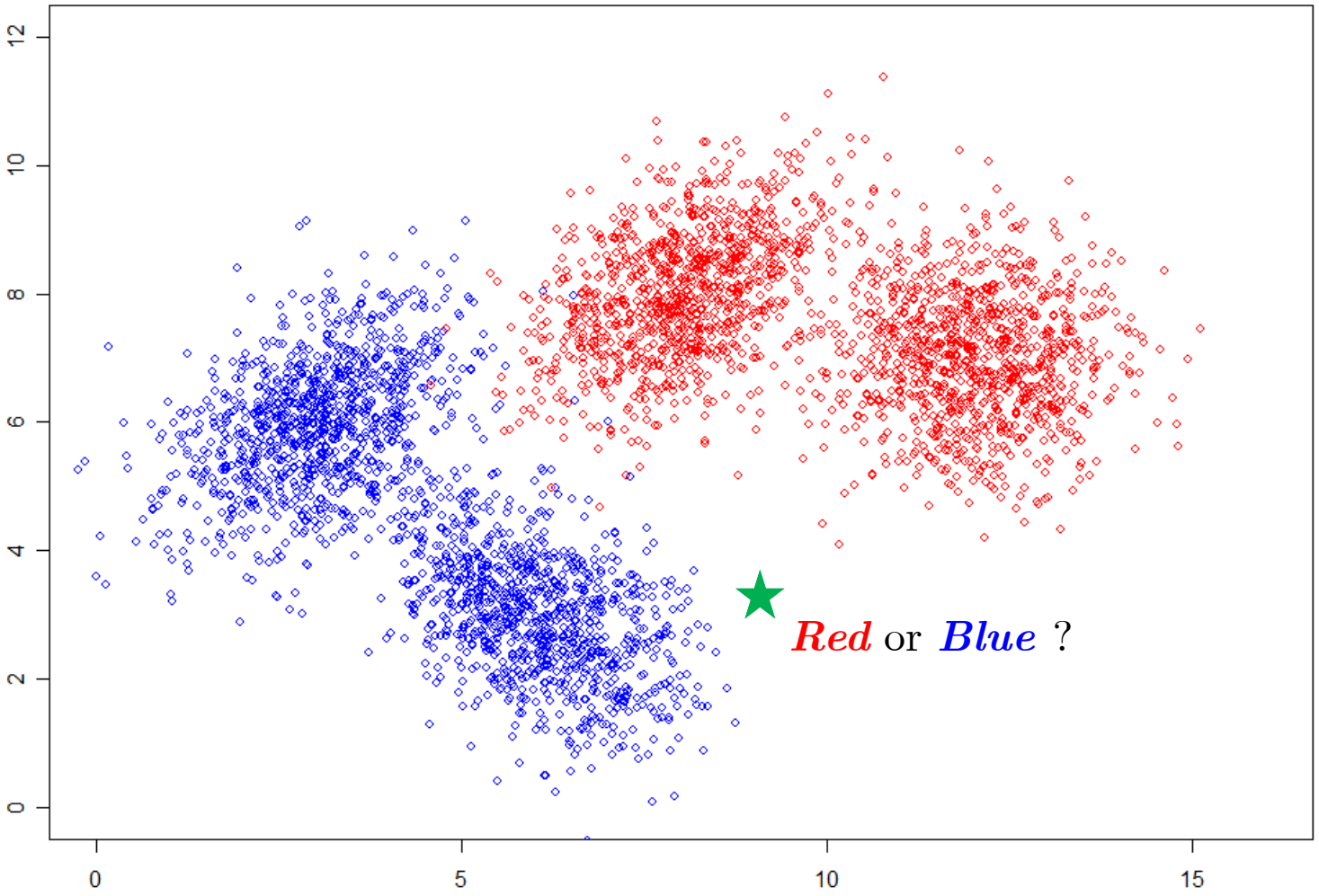
$$P(Y_i = 0) = g(\mathbf{x}_i)$$

$$\frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{f(\mathbf{x}_i)}{g(\mathbf{x}_i)} = h(\mathbf{x}_i)$$

Multiple class classification

$$l \geq 3$$

Binary classification



Classification via density estimation

Suppose that

red dots $(c = 1)$ come from a distribution given by $f_r(x, y)$ and

blue dots $(c = 0)$ come from a distribution given by $f_b(x, y)$.

Let (1) $Z \sim \text{Bernoulli}\ (\lambda)$;

(2) $(X, Y) \mid z = 1 \sim f_r(x, y) \Rightarrow p(x, y \mid z = 1) = f_r(x, y)$;

(3) $(X, Y) \mid z = 0 \sim f_b(x, y) \Rightarrow p(x, y \mid z = 0) = f_b(x, y)$.

$$\text{Then } p(z = 1 \mid x, y) = \frac{p(x, y \mid z = 1)p(z = 1)}{p(x, y \mid z = 1)p(z = 1) + p(x, y \mid z = 0)p(z = 0)}$$

$$= \frac{\lambda f_r(x, y)}{\lambda f_r(x, y) + (1 - \lambda) f_b(x, y)}$$

$$p(z = 0 \mid x, y) = \frac{(1 - \lambda) f_b(x, y)}{\lambda f_r(x, y) + (1 - \lambda) f_b(x, y)}$$

Independent density — Naïve Bayesian

If

$$f_r(x,y) = f_r(\mathbf{x}) = \prod_{i=1}^{n} f_{ri}(x_i)$$

$$f_b(x,y) = f_b(\mathbf{x}) = \prod_{i=1}^{n} f_{bi}(x_i)$$

Then

$$\log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)} = \log \frac{\lambda f_r(x,y)}{(1-\lambda) f_b(x,y)}$$

$$= \log \prod_{i=1}^{n} \frac{\lambda_r f_{ri}(x_i)}{\lambda_b f_{bi}(x_i)}$$

$$= \log \frac{\lambda_r}{\lambda_b} + \sum_{i=1}^{n} \log \frac{f_{ri}(x_i)}{f_{bi}(x_i)}$$

Gaussian density

Bivariate normal distribution

$$f(x,y \mid \mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$

$$\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

$$\Rightarrow \quad f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right],$$

where $\mathbf{x} = (x,y); \boldsymbol{\mu} = (\mu_X, \mu_Y)$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}$$

Equal covariance matrix — *L*inear *D*iscriminant *A*nalysis

If $$f_r(x,y) = f_r(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_r \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r)\right]$$

$$f_b(x,y) = f_b(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_b \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_b)^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x}-\boldsymbol{\mu}_b)\right]$$

and $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}$

Then

$$\log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)}$$

$$= \mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_b) - \frac{1}{2}(\boldsymbol{\mu}_r + \boldsymbol{\mu}_b)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_b) + \log \frac{\lambda}{1-\lambda}$$

$$= \left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}_r \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r + \log \lambda_r\right) - \left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_b - \frac{1}{2}\boldsymbol{\mu}_b \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_b + \log \lambda_b\right)$$

Linear discriminant function

$$\delta_l(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l - \frac{1}{2} \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l + \log \lambda_l$$

If $\delta_1(\mathbf{x}_i) \geq \delta_0(\mathbf{x}_i)$, y_i belongs to class 1.

If $\delta_1(\mathbf{x}_i) < \delta_0(\mathbf{x}_i)$, y_i belongs to class 0.

$\delta_1(\mathbf{x}_i) = \delta_0(\mathbf{x}_i)$ represents the decision boundary. In two dimension case, the boundary is a line. In high dimension case, the boundary is a hyperplane.

Unequal covariance matrix — *Q*uadratic *D*iscriminant *A*nalysis

If $$f_r(x,y) = f_r(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_r \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r)\right]$$

$$f_b(x,y) = f_b(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_b \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_b)^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x}-\boldsymbol{\mu}_b)\right]$$

and $$\boldsymbol{\Sigma}_r \neq \boldsymbol{\Sigma}_b$$

Then $$\log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)}$$

$$= \left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r) - \frac{1}{2}\log \boldsymbol{\Sigma}_r + \log \lambda_r\right)$$

$$- \left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{x}-\boldsymbol{\mu}_b) - \frac{1}{2}\log \boldsymbol{\Sigma}_b + \log \lambda_b\right)$$

Quadratic discriminant function

$$\delta_l(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) - \frac{1}{2}\log\boldsymbol{\Sigma}_l + \log\lambda_l$$

If $\delta_1(\mathbf{x}_i) \geq \delta_0(\mathbf{x}_i)$, y_i belongs to class 1.

If $\delta_1(\mathbf{x}_i) < \delta_0(\mathbf{x}_i)$, y_i belongs to class 0.

$\delta_1(\mathbf{x}_i) = \delta_0(\mathbf{x}_i)$ represents the decision boundary. In two dimension case, the boundary is a parabola. In high dimension case, the boundary is a hyperplane.

Gaussian mixture density

If

$$f_r(x,y) = f_r(\mathbf{x}) = \sum_{i=1}^{m_r} \frac{\delta_{ri}}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_{ri} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{ri})^T \boldsymbol{\Sigma}_{ri}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{ri})\right]$$

$$f_b(x,y) = f_b(\mathbf{x}) = \sum_{i=1}^{m_b} \frac{\delta_{ri}}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_{bi} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{bi})^T \boldsymbol{\Sigma}_{bi}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{bi})\right]$$

Then

$$\log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)} = \log \frac{\lambda f_r(x,y)}{(1-\lambda) f_b(x,y)} = \log \frac{\lambda}{1-\lambda} + \log \frac{f_r(x,y)}{f_b(x,y)}$$

Logistic density — Logistic regression

$$\frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)} = \frac{\lambda f_r(x,y)}{(1-\lambda) f_b(x,y)}$$

Since

$$\log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)} = \log \frac{\lambda}{1-\lambda} + \log \frac{f_r(x,y)}{f_b(x,y)}$$

If

$$f_r(x,y) = \frac{\exp(\alpha + \beta_x x + \beta_y y)}{1 + \exp(\alpha + \beta_x x + \beta_y y)}$$

$$f_b(x,y) = \frac{1}{1 + \exp(\alpha + \beta_x x + \beta_y y)}$$

Then

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log \frac{p(z=1 \mid x,y)}{p(z=0 \mid x,y)} = \beta_0 + \beta_x x + \beta_y y$$

$$\frac{\exp(\bullet)}{1+\exp(\bullet)}$$

Logistic Regression

统计学方法及其应用 > Inferential statistics > Logistic regression

The simplified situation

In one dimentional case, the problem is simplified as

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

For a series of observation $(x_i, y_i), i = 1, \ldots n,$ we have for each i

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i,$$

where $\pi_i = P(Y_i = 1)$.

This means that the response variable Y_i has a Bernoulli (π_i) distribution, and the log-odds $\pi_i / (1-\pi_i)$ follows a linear relation with the predictor variable x_i.

In other words,

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

Properties of the logistic function

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

For fixed α and β,

$$\pi(x) > 0,$$
$$\pi(x) < 1,$$
$$\pi(x) \to 0, \text{ when } \alpha + \beta x \to -\infty$$
$$\pi(x) \to 1, \text{ when } \alpha + \beta x \to \infty$$

Therefore, $\pi(x)$ is not suitable for modeling a probability that is either 0 or 1.

Properties of the logistic function

For fixed α and β

$$\begin{aligned}\frac{d}{dx}\pi(x) &= \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\beta - \frac{e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})^2}e^{\alpha+\beta x}\beta \\ &= \beta\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)\left(1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) \\ &= \beta\pi(x)(1-\pi(x))\end{aligned}$$

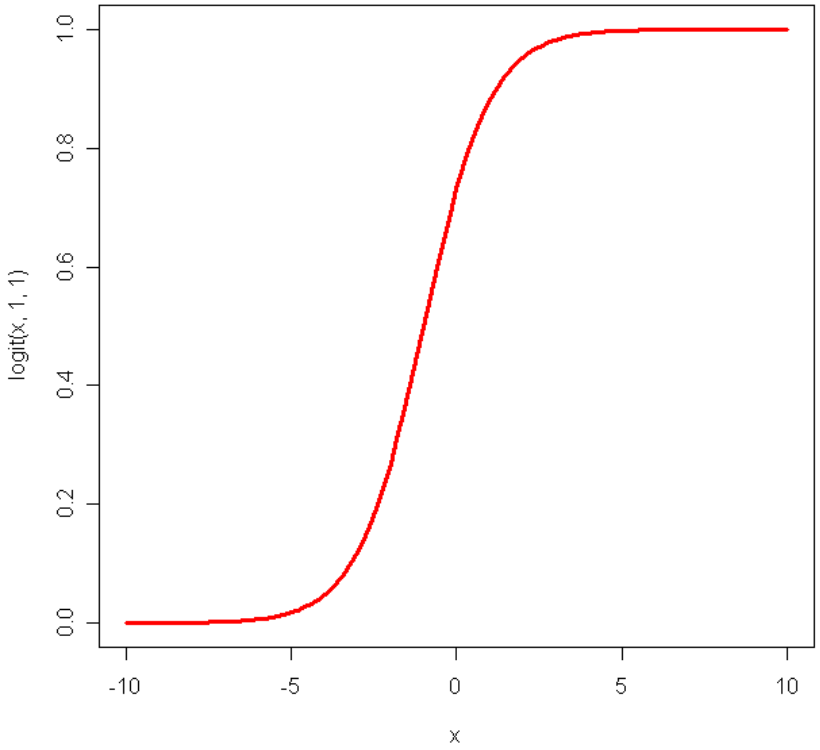
Because $0 < \pi(x) < 1,\ \pi(x)(1-\pi(x)) > 0,$

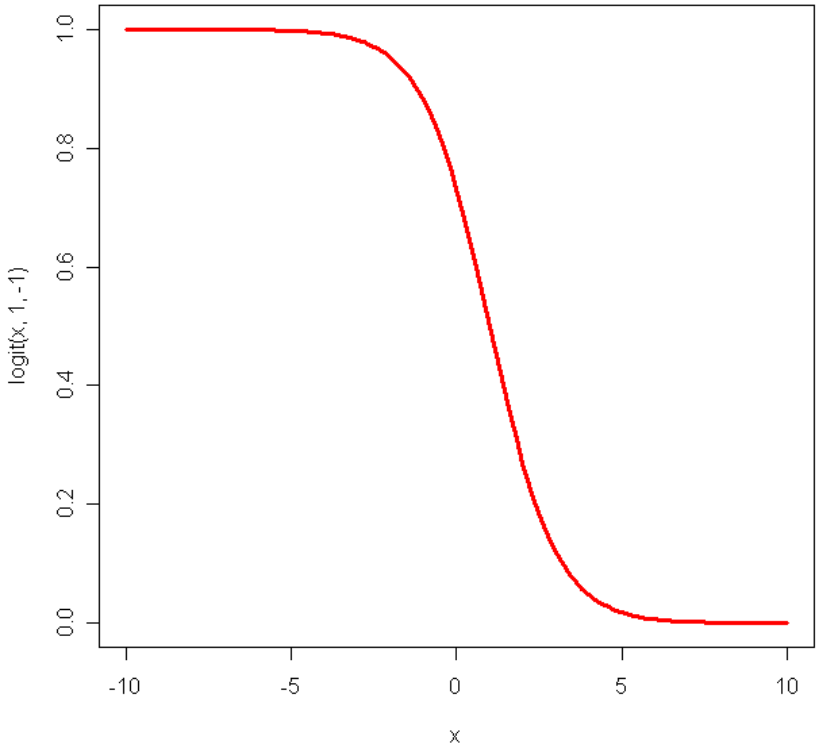
therefore, the sign of $\dfrac{d}{dx}\pi(x)$ depends on β

$\beta > 0,\ \pi(x) > 0 \Rightarrow \pi(x)$ is a strictly increacing function of x

$\beta < 0,\ \pi(x) < 0 \Rightarrow \pi(x)$ is a strictly decreacing function of x

Illustration of the logistic function





Change in odds

For fixed α and β,

$$\log\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = a + \beta(x+1) - (a+\beta x) = \beta$$

Therefore,

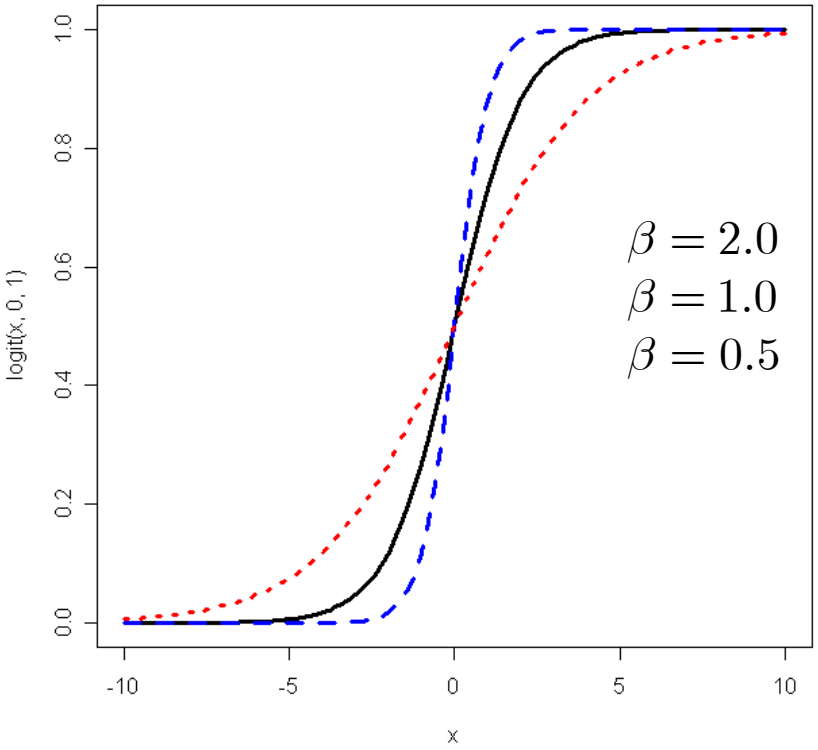
β is the change in the log-odds of success corresponding to one unit increament in x.

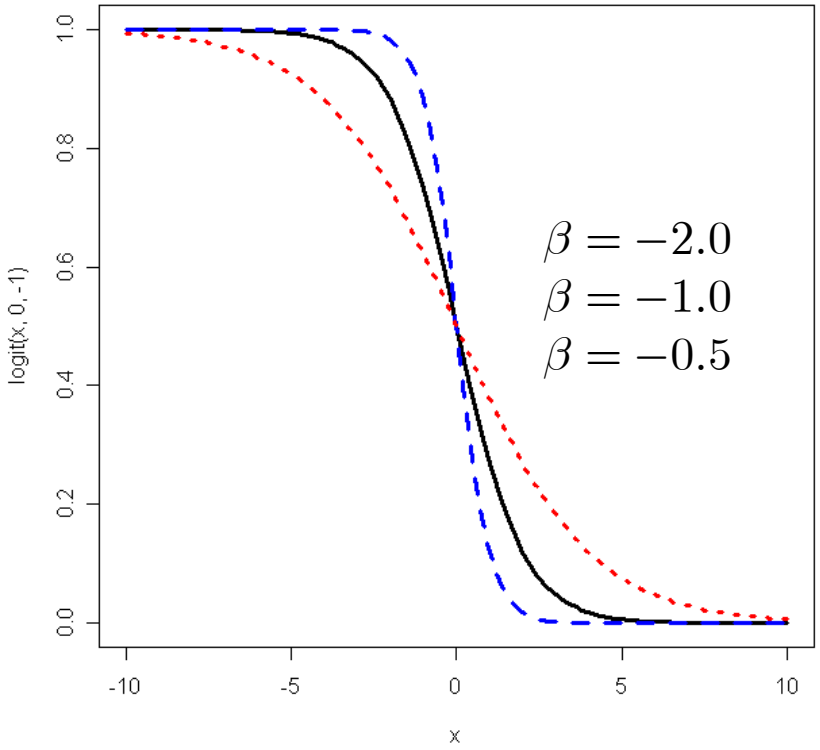
$$e^{\beta} = \frac{\pi(x+1)\,/\,(1-\pi(x+1))}{\pi(x)\,/\,(1-\pi(x))}$$

Therefore,

e^{β} is the odds ratio comparing the odds of success at $x+1$ to that at x. This ratio keeps as a constant when β is fixed.

Illustration of change in odds





How to determine the parameters?

For the parameters α and β, the likelihood function is

$$L(\alpha,\beta \mid \mathbf{y}) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Let

$$F_i = F_i(\alpha + \beta x_i) = \pi(x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

$$L(\alpha,\beta \mid \mathbf{y}) = \prod_{i=1}^{n} F_i^{y_i} (1 - F_i)^{1-y_i}$$

$$l(\alpha,\beta \mid \mathbf{y}) = \log L(\alpha,\beta \mid \mathbf{y}) = \sum_{i=1}^{n} \left[\log(1 - F_i) + y_i \log\left(\frac{F_i}{1 - F_i} \right) \right]$$

Maximum likelihood estimation

$$l(\alpha,\beta \mid \mathbf{y}) = \sum_{i=1}^{n} \left[\log(1-F_i) + y_i \log\left(\frac{F_i}{1-F_i}\right) \right]$$

Let $$f_i = \frac{\partial}{\partial\alpha} F_i = \frac{\partial}{\partial\alpha} \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} + \frac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2} = F_i(1-F_i),$$

then $$\frac{\partial}{\partial\alpha} \log(1-F_i) = \frac{1}{1-F_i} \frac{\partial}{\partial\alpha}(1-F_i) = -F_i$$

$$\frac{\partial}{\partial\alpha} \log\left(\frac{F_i}{1-F_i}\right) = \frac{1-F_i}{F_i} \frac{\partial}{\partial\alpha} \frac{F_i}{1-F_i} = \frac{f_i}{F_i(1-F_i)} = 1$$

$$\frac{\partial}{\partial\alpha} l(\alpha,\beta \mid \mathbf{y}) = \sum_{i=1}^{n} (y_i - F_i)$$

Maximum likelihood estimation

$$l(\alpha,\beta \mid \mathbf{y}) = \sum_{i=1}^{n} \left[\log(1-F_i) + y_i \log\left(\frac{F_i}{1-F_i}\right) \right]$$

Let $$f_i = \frac{\partial}{\partial\beta} F_i = \frac{\partial}{\partial\beta} \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} x_i + \frac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2} x_i = x_i F_i (1-F_i),$$

Then $$\frac{\partial}{\partial\beta} \log(1-F_i) = \frac{1}{1-F_i} \frac{\partial}{\partial\beta}(1-F_i) = -F_i x_i$$

$$\frac{\partial}{\partial\beta} \log\left(\frac{F_i}{1-F_i}\right) = \frac{1-F_i}{F_i} \frac{\partial}{\partial\beta} \frac{F_i}{1-F_i} = x_i$$

$$\frac{\partial}{\partial\beta} l(\alpha,\beta \mid \mathbf{y}) = \sum_{i=1}^{n} (y_i - F_i) x_i$$

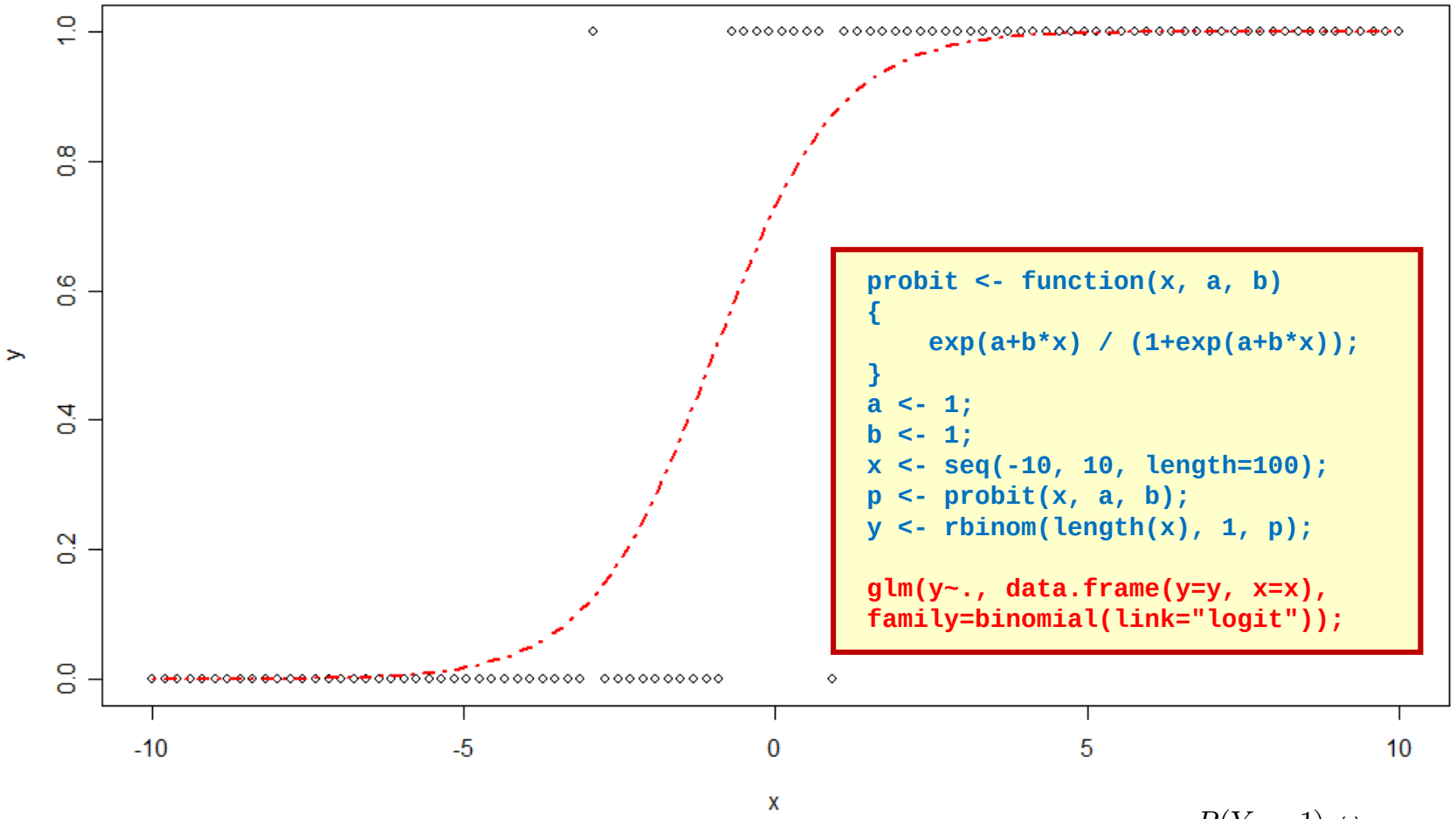
Maximum likelihood estimation

$$\begin{cases} \dfrac{\partial}{\partial \alpha} l(\alpha,\beta \mid \mathbf{y}) = \displaystyle\sum_{i=1}^{n} (y_i - F_i) = 0 \\ \dfrac{\partial}{\partial \beta} l(\alpha,\beta \mid \mathbf{y}) = \displaystyle\sum_{i=1}^{n} (y_i - F_i) x_i = 0 \end{cases}$$

Nonlinear equations, must be solved numerically by using methods such as

Newton - Raphson

An example




```
probit <- function(x, a, b)
{
    exp(a+b*x) / (1+exp(a+b*x));
}
a <- 1;
b <- 1;
x <- seq(-10, 10, length=100);
p <- probit(x, a, b);
y <- rbinom(length(x), 1, p);

glm(y~., data.frame(y=y, x=x),
family=binomial(link="logit"));
```

$$P(Y_i = 1) \Leftrightarrow x_i = -\frac{\alpha}{\beta}$$

Significance of the parameters

For the intercept

$$H_0 : \alpha = 0 \quad \text{versus} \quad H_1 : \alpha \neq 0$$

χ^2 test

$$-2\log\lambda = -2\left[\log L(0,\hat{\beta}) - \log L(\hat{\alpha},\hat{\beta})\right] \sim \chi_1^2, \text{ asymptotically}$$

p-value is $P(\chi_1^2 \geq -2\log\lambda)$.

For the slope

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

χ^2 test

$$-2\log\lambda = -2\left[\log L(\hat{\alpha},0) - \log L(\hat{\alpha},\hat{\beta})\right] \sim \chi_1^2, \text{ asymptotically}$$

p-value is $P(\chi_1^2 \geq -2\log\lambda)$.

Multiple logistic regression

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Let $y_1, \ldots, y_n$ be the class label.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the predictor variable, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$.

Significance of the parameters

For the parameter

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

Wald test

$$Z = \frac{\widehat{\beta}_j}{\text{se}(\widehat{\beta}_j)} \sim \text{N}(0,1), \text{ asymptotically}$$

p-value is $2P(Z \geq z_{\widehat{\beta}_j})$.

For the slope

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

χ^2 test

$$-2\log\lambda = -2\left[\log L(\beta_j = 0) - \log L(\widehat{\beta}_j)\right] \sim \chi^2_1, \text{ asymptotically}$$

p-value is $P(\chi^2_1 \geq -2\log\lambda)$.

Regularization — ridge and lasso

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \lambda \sum_{j=1}^{k} \beta_j^2$$

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \lambda \sum_{j=1}^{k} |\, \beta_j \,|$$

Let $y_1, \ldots, y_n$ be the class label.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the predictor variable, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$.

Classification

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right)=\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}$$

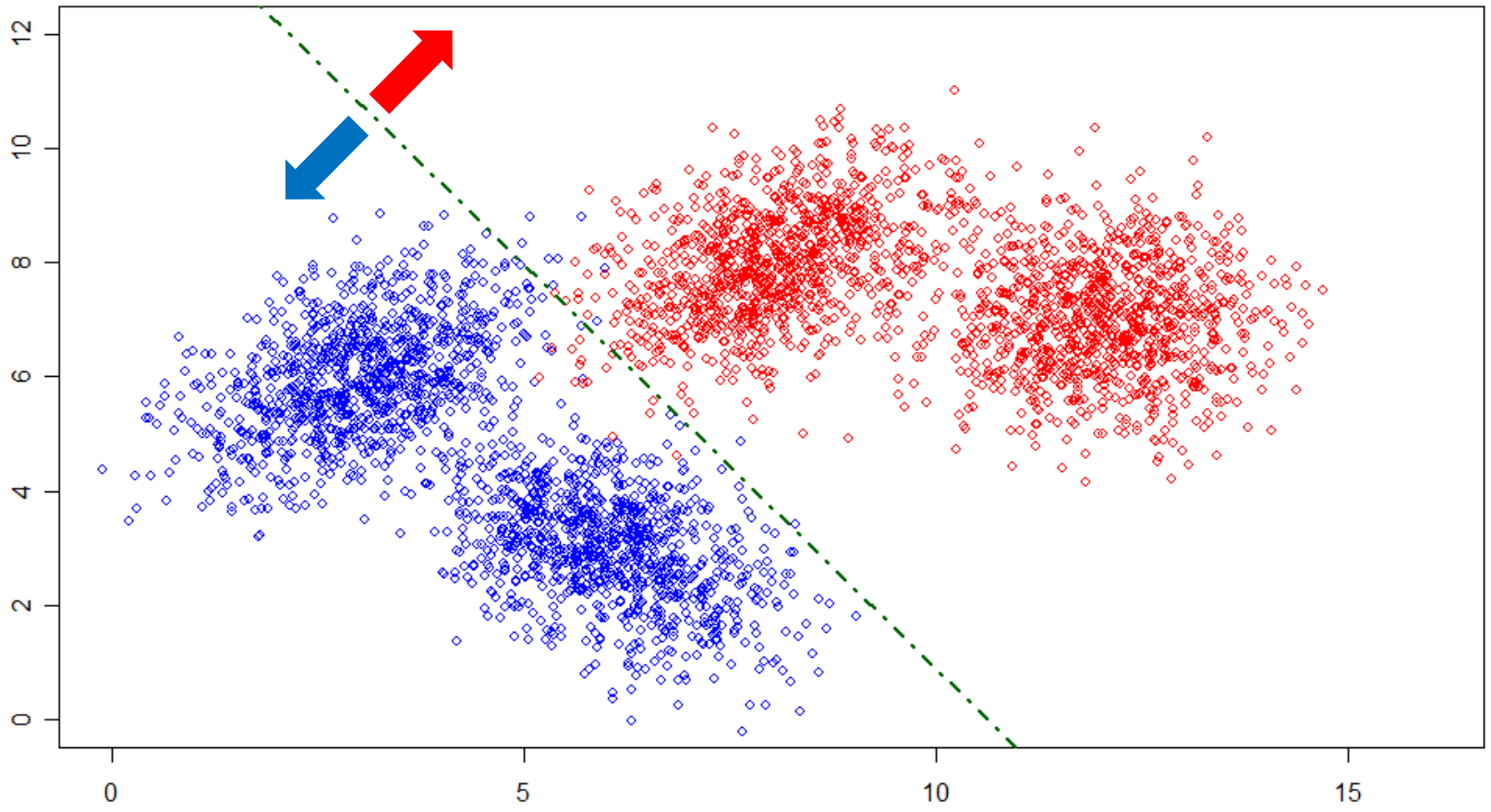
$$(1)\ \beta_0+\sum_{j=1}^{k}\beta_j x_{ij}\geq 0,\ \text{claim class 1.}$$

$$(2)\ \beta_0+\sum_{j=1}^{k}\beta_j x_{ij}< 0,\ \text{claim class 0.}$$

$\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}=0$ represents the decision boundary. In two dimension case, the boundary is a line. In high dimension case, the boundary is a hyperplane.

Hard decision

Decision boundary



Prediction

$$P(Y_i = 1) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}$$

$P(Y_i = 1)$ is the probability that Y_i belongs to class 1.
Therefore we have a kind of **soft decision**.

Soft decision

Generalized linear model

$$Y_i \rightleftharpoons \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}$$

Random component : the response variable Y_i,
independent,
not identically distributed,
from the same exponential family.

Systematic component : $\beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}$

Link function : $g(\mu_i) = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}, \mu_i = \mathrm{E}\, Y_i$

Logistic regression

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Random component : the response variable Y_i,
independent,
not identically distributed,
from the Bernoulli family.

Systematic component : $\beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}$

Link function : $\mathrm{E}\, Y_i = P(Y_i = 1) = \dfrac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}$

Probit regression

$$\Phi^{-1}\left[P(Y_i = 1)\right] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Random component : the response variable Y_i, independent, not identically distributed, from the Bernoulli family.

Systematic component : $\beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}$

Link function : $\mathrm{E}\, Y_i = P(Y_i = 1) = \Phi\left(\beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}\right)$

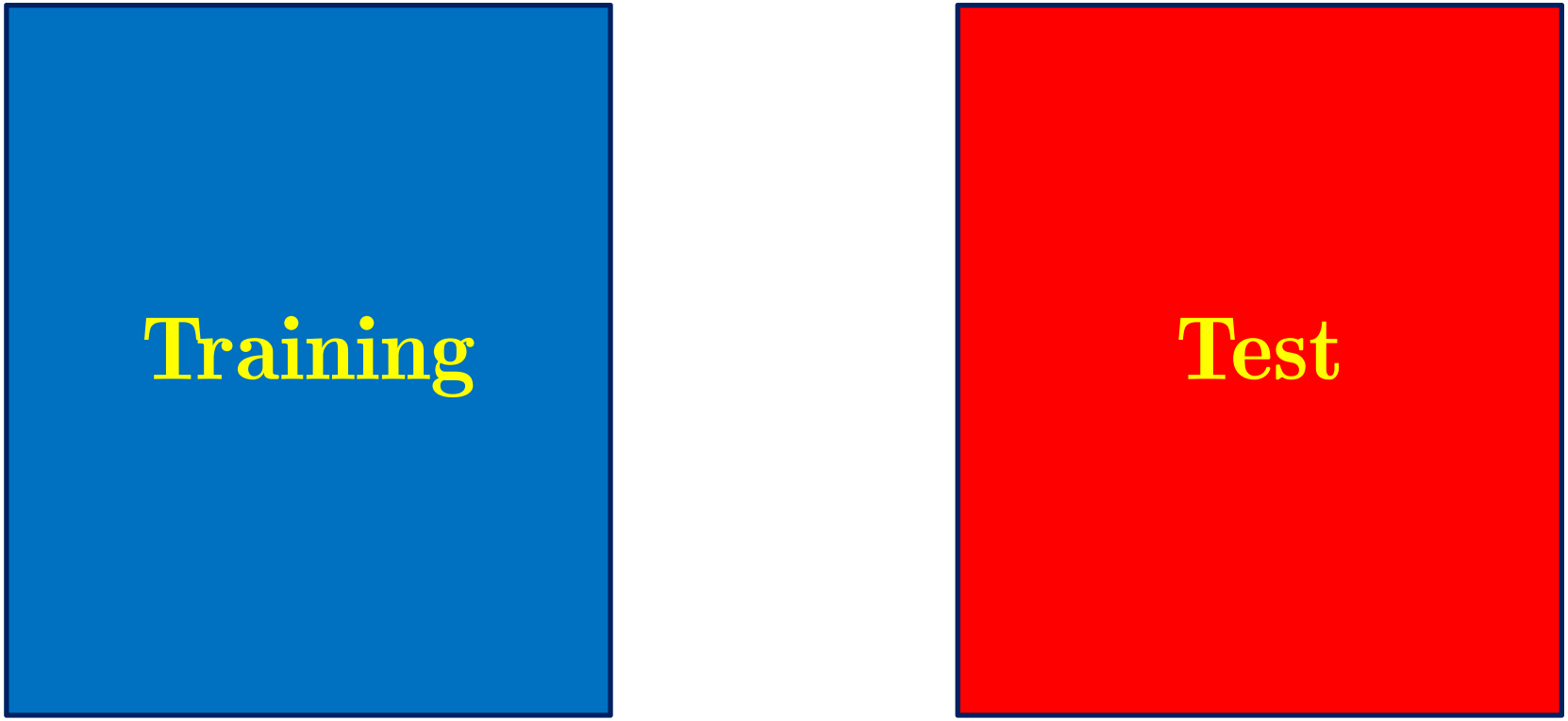
GLM in R

```
glm(    formula,                          y~.; y~x1+x2+x3.
        family = gaussian,                binomial(link="logit")
        data,                             binomial(link="probit")
        weights,
        subset,
        na.action,
        start = NULL,
        etastart,
        mustart,
        offset,
        control = glm.control(...),
        model = TRUE,
        method = "glm.fit",
        x = FALSE,
        y = TRUE,
        contrasts = NULL,
        ...)
```

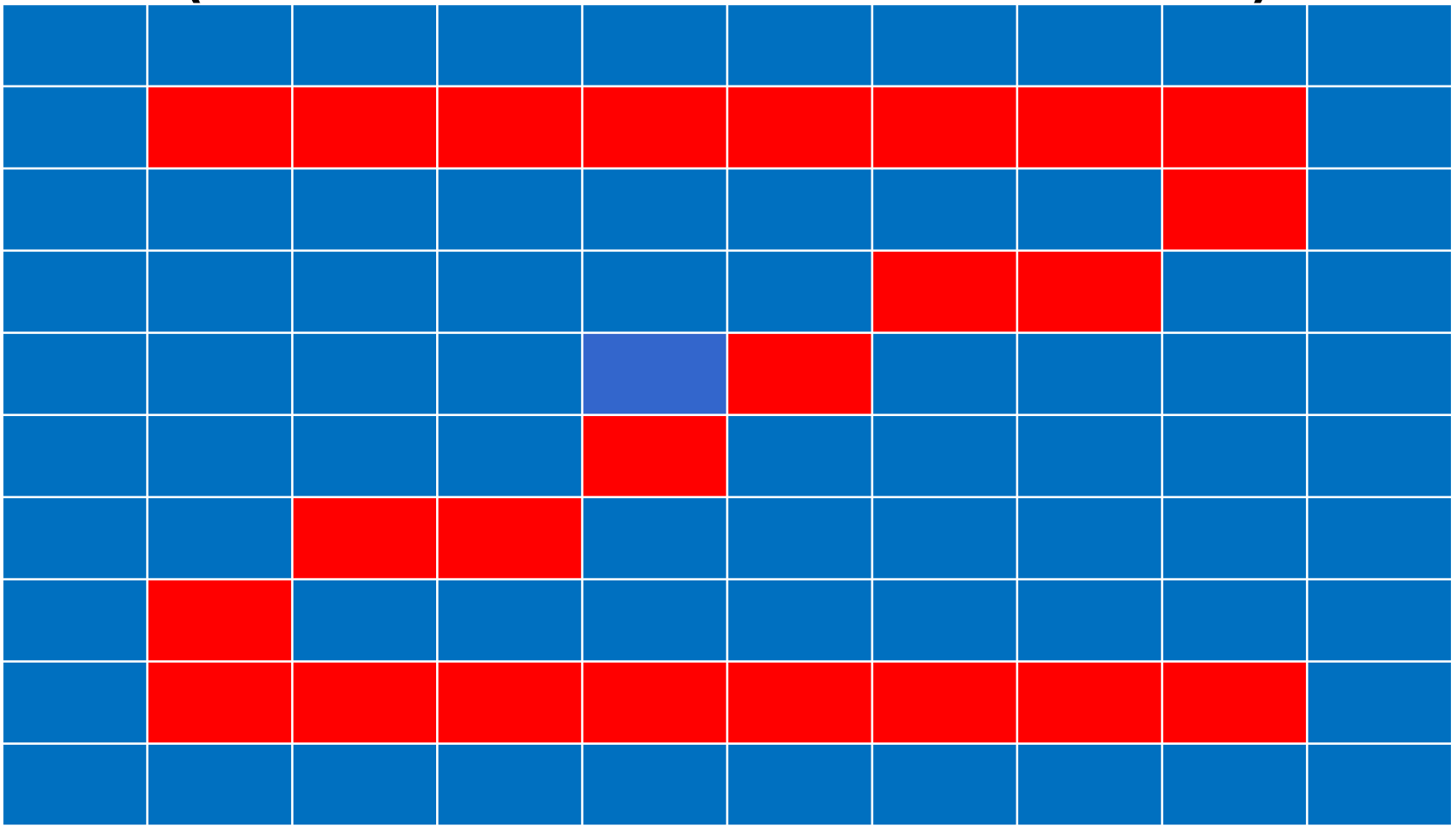
Validation

统计学方法及其应用 > 统计学应用 > 预测

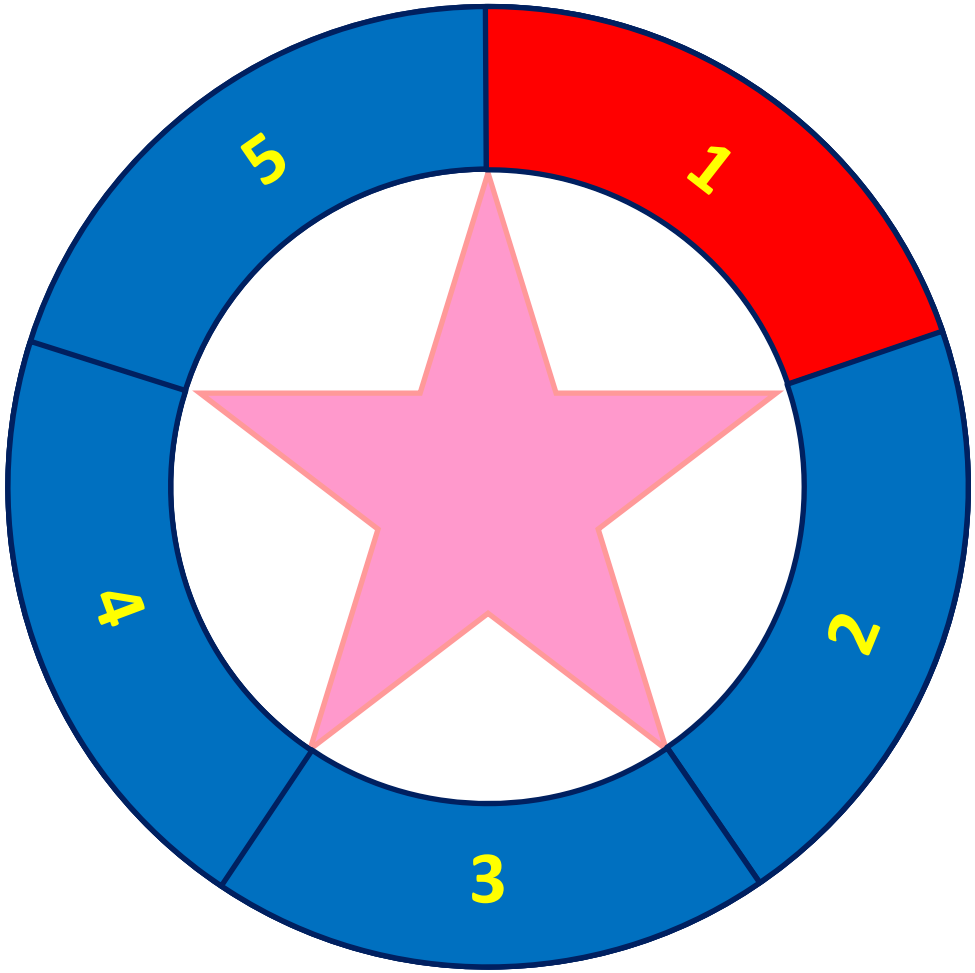
Independent test set



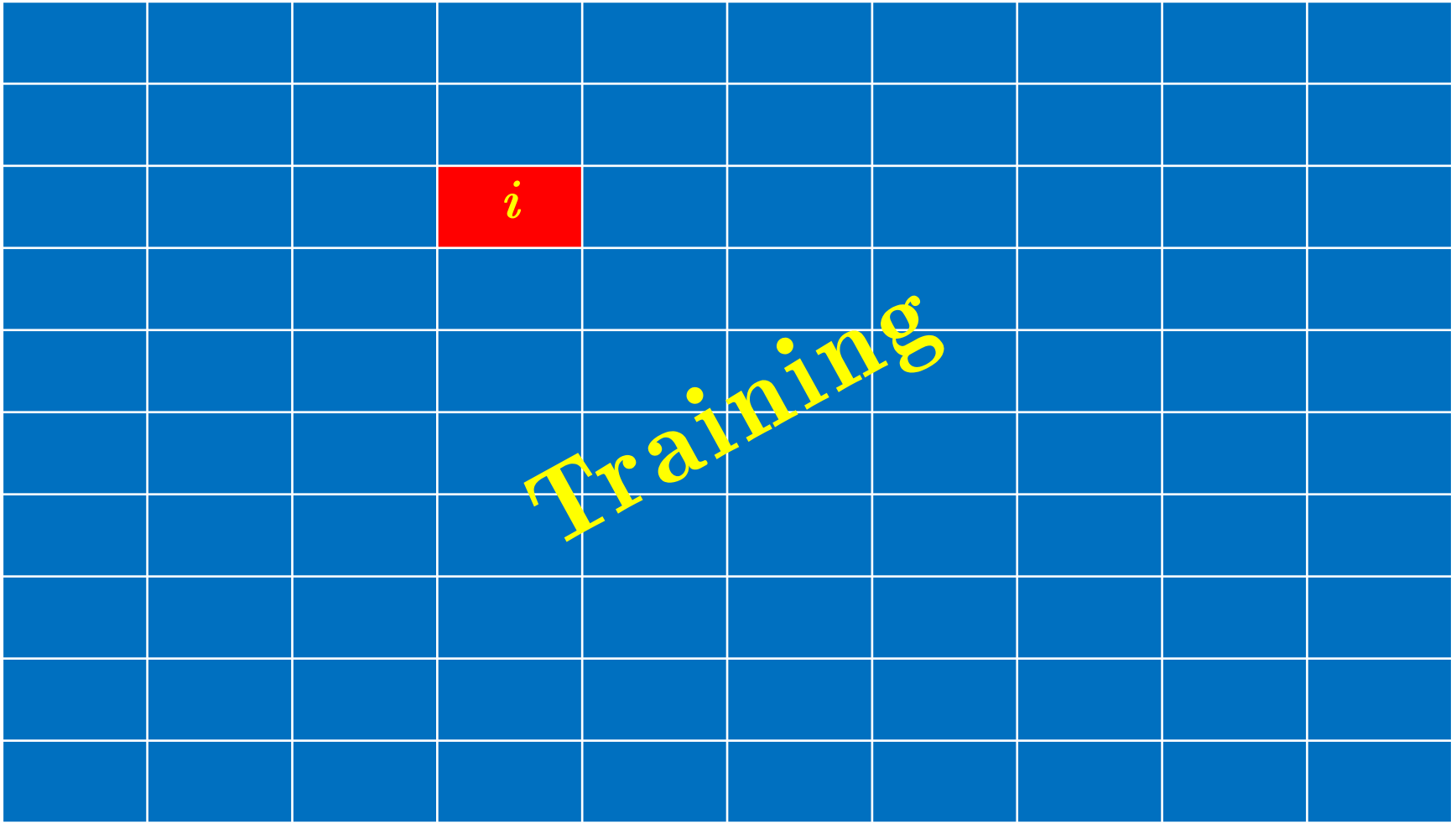
Random sub-sampling
(Monte Carlo cross-validation)



Multi-fold cross-validation



Leave-one-out cross-validation



Error rate

- Confusion matrix

Binary classification		Truth	
		Positive	Negative
Decision	Claim Positive ($\lambda < c$)	**True Positive (TP)**	**False Positive (FP)**
	Claim Negative ($\lambda \geq c$)	**False Negative (FN)**	**True Negative (TN)**

- Empirical ratios (on the basis of a large number of decisions)
  - **True positive rate = TP / (TP+FN) = Sensitivity**
  - **False positive rate = FP / (TN+FP) = 1- Specificity**
  - **True negative rate = TN / (TN+FP) = Specificity**
  - **False negative rate = FN / (TP+FN) = 1- Sensitivity**

Sensitivity and specificity

- Confusion matrix

Binary classification		Truth	
		Positive	Negative
Decision	Claim Positive ($\lambda < c$)	**Sensitivity**	**1-Specificity**
	Claim Negative ($\lambda \geq c$)	**1-Sensitivity**	**Specificity**

Evaluation criteria

- Classification accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

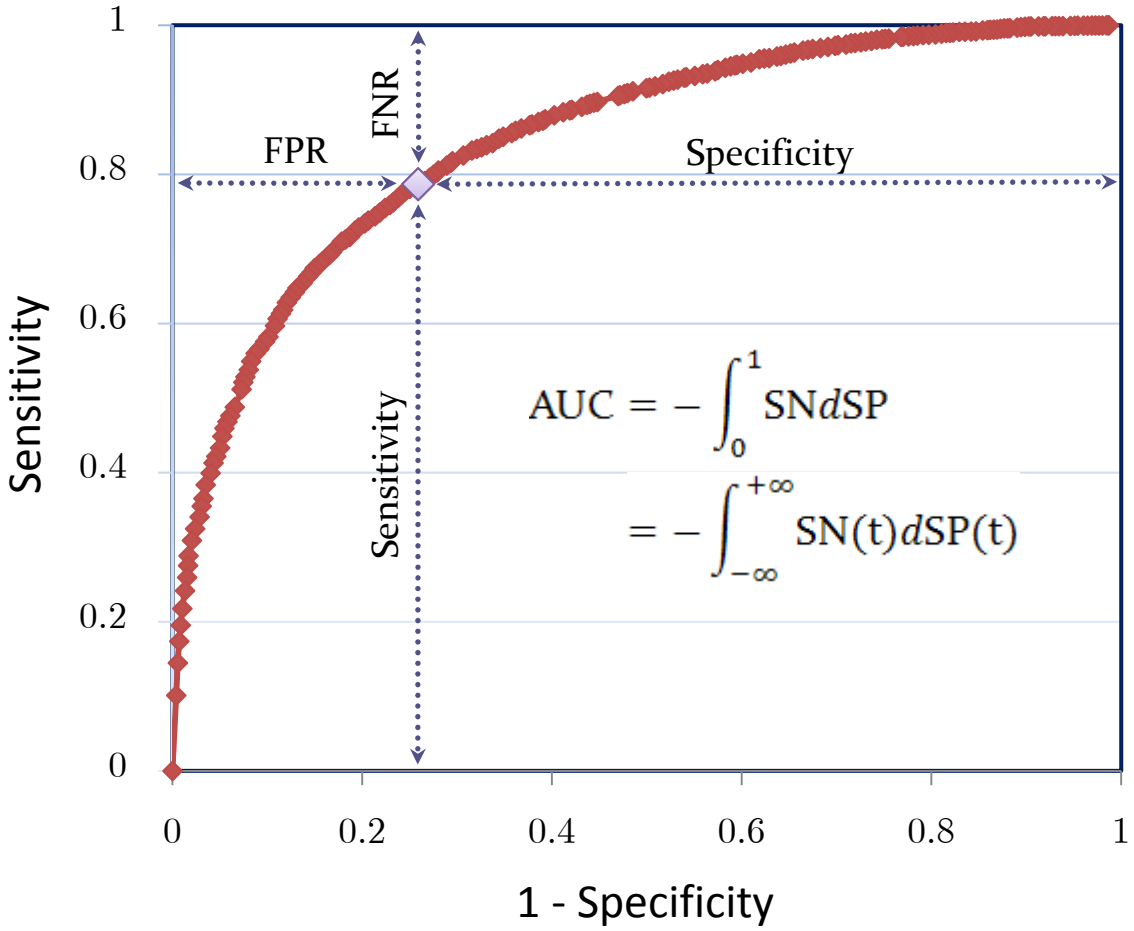
- Balanced error rate

$$BER = \frac{1}{2}\left(\frac{FP}{FP + TN} + \frac{FN}{FN + TP}\right)$$

- Matthew's correlation coefficient

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(FP + TN)(TN + FN)(FN + TP)}}$$

Receiver ***O***perating ***C***haracteristic curve and ***A***rea ***U***nder this ***C***urve



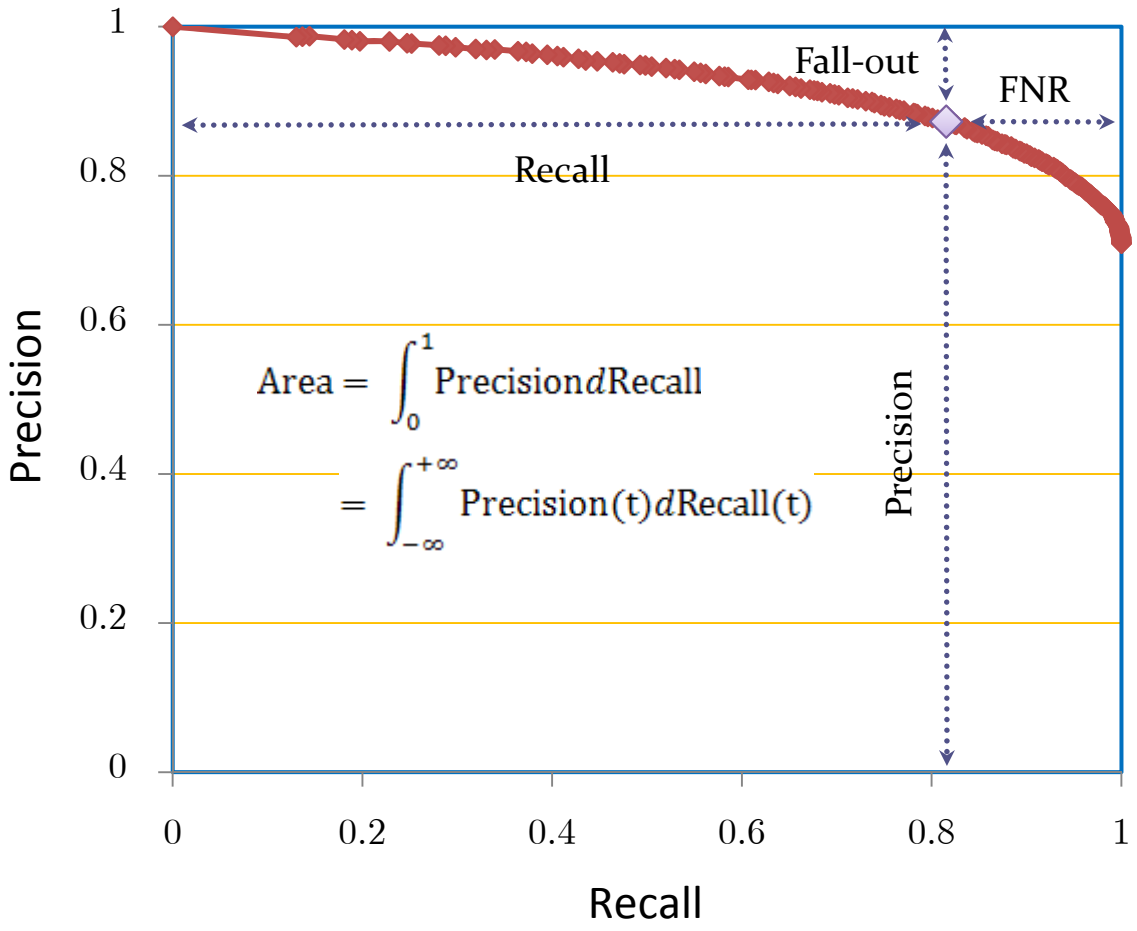
Error rate

- Confusion matrix

Binary classification		Truth	
		Positive	Negative
Decision	Claim Positive ($\lambda < c$)	**True Positive (TP)**	**False Positive (FP)**
	Claim Negative ($\lambda \geq c$)	**False Negative (FN)**	**True Negative (TN)**

- Empirical ratios
  - Recall = TP / (TP+FN)
  - Precision = TP / (TP+FP)
  - Fall-out = FP / (TP+FP)
  - F1 measure = $2 \times \text{Recall} \times \text{Precision}$ / (Recall + Precision)

Precision-Recall curve



R program — Confusion matrix

```
confusion.matrix <- function(prob, class, cut=0.5, ratio=F)
{
        CM <- matrix(rep(0, 4), nr=2, nc=2, byrow=T);
        CM[1,1] <- sum(prob >  cut & class == 1);
        CM[1,2] <- sum(prob >  cut & class != 1);
        CM[2,1] <- sum(prob <= cut & class == 1);
        CM[2,2] <- sum(prob <= cut & class != 1);
        if(ratio==T){
                CM[,1] <- CM[,1] / sum(CM[,1]);
                CM[,2] <- CM[,2] / sum(CM[,2]);
        }
        CM;
}
```

R program — Evaluation criteria

```
evaluation.criteria <- function(CM)
{
        TP  <- CM[1,1];
        FP  <- CM[1,2];
        FN  <- CM[2,1];
        TN  <- CM[2,2];

        TPR <- TP / (TP + FN);
        FPR <- FP / (FP + TN);
        FNR <- FN / (FN + TP);
        TNR <- TN / (TN + FP);

        ACC <- (TP+TN) / (TP+FP+FN+TN);
        MCC <- (TP*TN - FP*FN) / sqrt((TP+FP)*(FP+TN)*(TN+FN)*(FN+TP));
        BER <- (FPR + FNR) / 2;

        data.frame(
                TP  = TP,  FP  = FP,  FN  = FN,  TN  = TN,
                TPR = TPR, FPR = FPR, FNR = FNR, TNR = TNR,
                SN  = TPR, SP  = TNR,
                ACC = ACC, MCC = MCC, BER = BER);
}
```

R program — ROC curve

```
roc.curve <- function(prob, class)
{
        cut <- sort(prob, decreasing=T);
        roc <- matrix(NA, nr=length(prob), nc=2, byrow=T);

        for(i in 1:length(cut)){
                cm <- confusion.matrix   (prob, class, cut[i]);
                ec <- evaluation.criteria(cm);
                roc[i,1] = ec$FPR;
                roc[i,2] = ec$TPR;
        }

        roc;
}
```

R program — AUC score

```
auc.score <- function(roc)
{
        auc = 0.0;
        for(i in 2:dim(roc)[1]){
                dsp <- roc[i,1] - roc[i-1,1];
                dsn <- roc[i,2] + roc[i-1,2];
                auc <- auc + dsp*dsn/2;
        }

        auc;
}
```

Information matrix

Information matrix

$$I(\theta_1,\theta_2)=\begin{pmatrix} -\dfrac{\partial^2}{\partial\theta_1^2}\log L(\theta_1,\theta_2\mid\mathbf{y}) & -\dfrac{\partial^2}{\partial\theta_1\theta_2}\log L(\theta_1,\theta_2\mid\mathbf{y}) \\ -\dfrac{\partial^2}{\partial\theta_1\theta_2}\log L(\theta_1,\theta_2\mid\mathbf{y}) & -\dfrac{\partial^2}{\partial\theta_2^2}\log L(\theta_1,\theta_2\mid\mathbf{y}) \end{pmatrix}$$

In logistic regression

$$I(\alpha,\beta) = \begin{pmatrix} -\dfrac{\partial^2}{\partial\alpha^2}\log L(\alpha,\beta\mid\mathbf{y}) & -\dfrac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha,\beta\mid\mathbf{y}) \\ -\dfrac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha,\beta\mid\mathbf{y}) & -\dfrac{\partial^2}{\partial\beta^2}\log L(\alpha,\beta\mid\mathbf{y}) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} F_i(1-F_i) & \sum_{i=1}^{n} x_i F_i(1-F_i) \\ \sum_{i=1}^{n} x_i F_i(1-F_i) & \sum_{i=1}^{n} x_i^2 F_i(1-F_i) \end{pmatrix}$$

Inverse of the information matrix

Inverse of the information matrix

$$I^{-1}(\theta_1,\theta_2) = \frac{1}{|\, I(\theta_1,\theta_2)\,|}\begin{pmatrix} -\dfrac{\partial^2}{\partial\theta_2^2}\log L(\theta_1,\theta_2 \mid \mathbf{y}) & \dfrac{\partial^2}{\partial\theta_1\theta_2}\log L(\theta_1,\theta_2 \mid \mathbf{y}) \\ \dfrac{\partial^2}{\partial\theta_1\theta_2}\log L(\theta_1,\theta_2 \mid \mathbf{y}) & -\dfrac{\partial^2}{\partial\theta_1^2}\log L(\theta_1,\theta_2 \mid \mathbf{y}) \end{pmatrix}$$

In logistic regression

$$\begin{aligned} I^{-1}(\alpha,\beta) &= \frac{1}{|\, I(\alpha,\beta)\,|}\begin{pmatrix} -\dfrac{\partial^2}{\partial\beta^2}\log L(\alpha,\beta \mid \mathbf{y}) & \dfrac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha,\beta \mid \mathbf{y}) \\ \dfrac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha,\beta \mid \mathbf{y}) & -\dfrac{\partial^2}{\partial\alpha^2}\log L(\alpha,\beta \mid \mathbf{y}) \end{pmatrix} \\ &= \frac{1}{|\, I(\alpha,\beta)\,|}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 F_i(1-F_i) & -\sum_{i=1}^{n} x_i F_i(1-F_i) \\ -\sum_{i=1}^{n} x_i F_i(1-F_i) & \sum_{i=1}^{n} F_i(1-F_i) \end{pmatrix} \end{aligned}$$

Standard error of the parameters

Inverse of the logistic regression information matrix

$$I^{-1}(\hat{\alpha},\hat{\beta}) = \frac{1}{|\, I(\hat{\alpha},\hat{\beta})\,|}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 F_i(1-F_i) & -\sum_{i=1}^{n} x_i F_i(1-F_i) \\ -\sum_{i=1}^{n} x_i F_i(1-F_i) & \sum_{i=1}^{n} F_i(1-F_i) \end{pmatrix}$$

$$[\mathrm{se}(\hat{\alpha})]^2 = \frac{1}{|\, I(\hat{\alpha},\hat{\beta})\,|}\sum_{i=1}^{n} x_i^2 F_i(1-F_i)$$

$$[\mathrm{se}(\hat{\beta})]^2 = \frac{1}{|\, I(\hat{\alpha},\hat{\beta})\,|}\sum_{i=1}^{n} F_i(1-F_i)$$

Significance of the parameters

For the intercept

$$H_0 : \alpha = 0 \quad \text{versus} \quad H_1 : \alpha \neq 0$$

Wald test

$$Z = \frac{\hat{\alpha}}{\text{se}(\hat{\alpha})} \sim \text{N}(0,1), \text{ asymptotically}$$

p-value is $2P(Z \geq z_{\hat{\alpha}})$.

For the slope

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

Wald test

$$Z = \frac{\hat{\beta}}{\text{se}(\hat{\beta})} \sim \text{N}(0,1), \text{ asymptotically}$$

p-value is $2P(Z \geq z_{\hat{\beta}})$.