Lecture 02 线性方法

汪小我

清华大学自动化系

统计学习 Statistical learning

$$Y = f(X) + \varepsilon$$

$$X = [X_1, X_2, \cdots, X_p]^T$$

- X是p维变量，也称为输入变量（input variable），特征（feature），预测因子(predictor)
- Y称为输出变量（output variable），响应（response）
- ε是随机误差项

训练样本: $(\mathbf{x}^{(i)}, \mathrm{y}^{(i)})$, $i=1,\ldots,n$

大写字母代表随机变量
小写字母表示具体的样本取值

统计学习的目标是预测函数关系$f()$

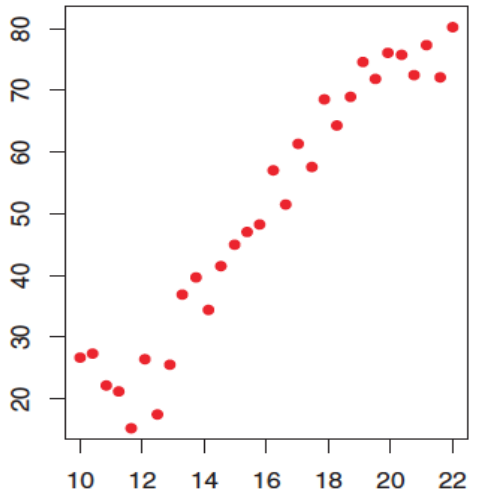
符号约定

- n : 训练样本数量
- $\mathbf{x}$: 特征向量取值
- p/d : 特征向量维度
- y : 样本取值、标签
- $(\mathbf{x},y)$ 表示训练样本
- 第i个训练样本$(\mathbf{x}^{(i)},y^{(i)})$, $i=1,\ldots,n$

为什么要求解 f？

- 1. 预测问题(prediction)

$$\hat{Y} = \hat{f}(X)$$

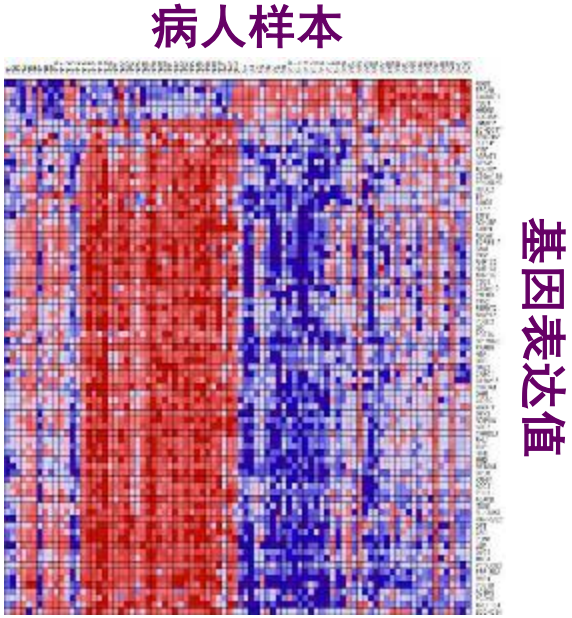


给定新样本的属性X，预测其标签（y的取值）

为什么要求解 f？

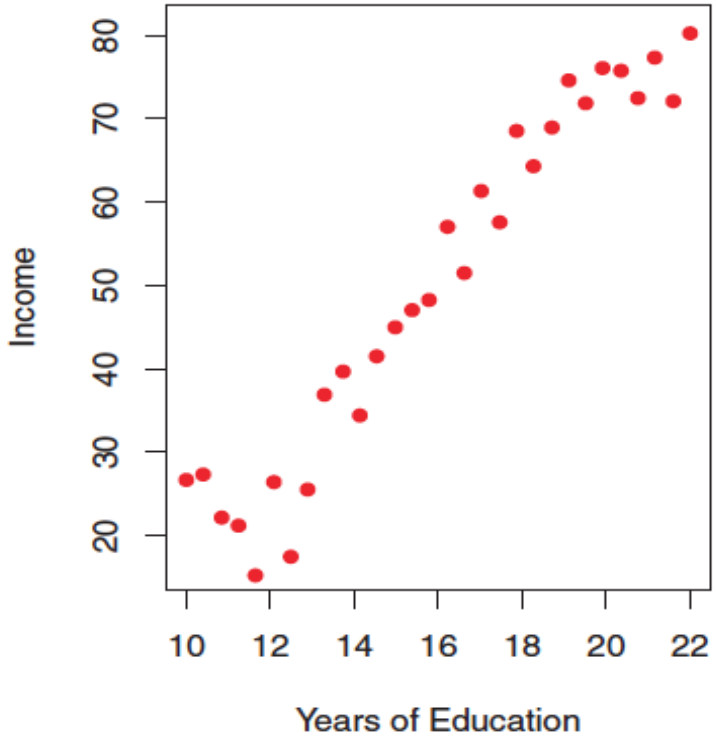
- 2. 推理问题（Inference）
  - *X*的哪分量对*Y*有影响（association）？
  - 这些分量与*Y*之间满足什么样的函数关系？
  - 这些分量与*Y*之间的关系能否用一个统一的模型描述

例如寻找癌症的致病基因



线性模型

回归问题（regression）



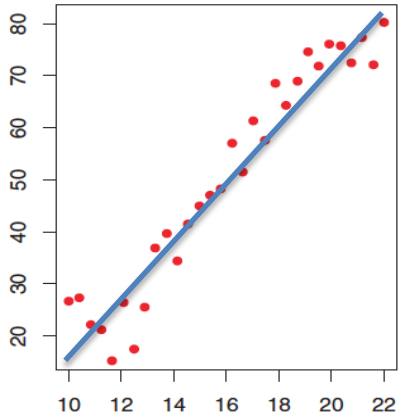
Gareth James et al., *An Introduction to Statistical Learning with Applications in R*, Springer, 2015

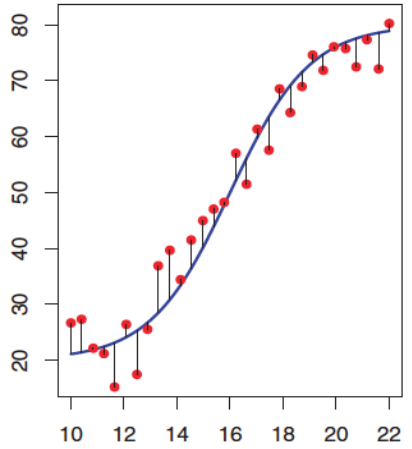
关于预测模型的选择

$$\hat{Y} = \hat{f}(X)$$

- *Y*的预测值与真实值之间的差异依赖于两个方面

1. Reducible error
2. Irreducible error





$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

关于预测模型的选择

- 参数化方法（Parametric）

  分两步完成：

  – 假设模型形式 (e.g. 线性 or 非线性模型)

  – 学习模型参数 (fit parameters)

- 非参数方法 (non-parametric)

  – 不对 f 的具体形式进行约束和假设

  – 优点：灵活

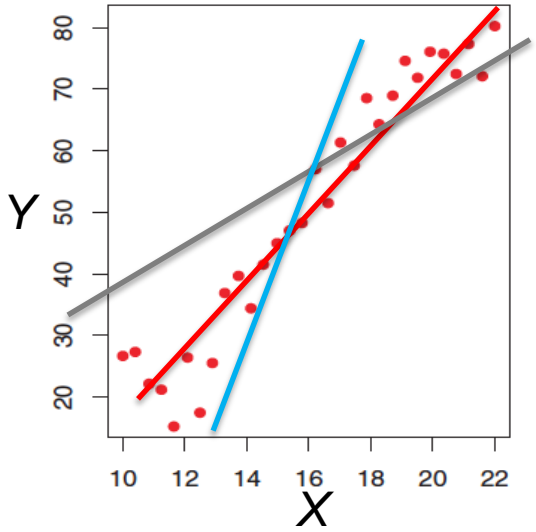
  – 缺点：需要大量的训练样本

举例，线性分类器与k近邻

线性回归（Linear regression）

$$Y = \theta_0 + \theta_1 X$$

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$$



对于自变量观测值x，模型预测因变量取值 $\hat{y}$

$\hat{\theta}_0, \hat{\theta}_1$ 是我们通过训练数据估计的模型参数

参数学习

- Residual sum of squares（*RSS*）

$$e = y - \hat{y} = y - \hat{\theta}_0 - \hat{\theta}_1 x$$

$$RSS = e^{(1)^2} + e^{(2)^2} + \cdots + e^{(n)^2}$$

- 学习目标：使得*RSS*最小化

可解得：$\hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x}$　　$\hat{\theta}_1 = \dfrac{\sum_{i=1}^{n} (\overline{y} - y^{(i)})(\overline{x} - x^{(i)})}{\sum_{i=1}^{n} (\overline{x} - x^{(i)})^2}$

其中：$\overline{x} = \sum_{i=1}^{n} x^{(i)}$　　$\overline{y} = \sum_{i=1}^{n} y^{(i)}$

如何衡量回归效果？

- 决定系数R^2 : The proportion of variance explained

$$R^2 = \frac{\sum_{i=1}^{n}(\bar{y} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\bar{y} - y^{(i)})^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

其中： $TSS = \sum_{i=1}^{n}(\bar{y} - y^{(i)})^2$

R^2与相关系数Corr(X,Y)之间的关系？

多元线性回归

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p$$

定义函数： $f_\Theta(X) = \theta_0 X_0 + \theta_1 X_1 + \cdots + \theta_p X_p = \Theta^{\mathrm{T}}\mathbf{X}$

其中X_0=1

定义损失函数（Loss function）：

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - f_\Theta(\mathbf{x}^{(i)}))^2$$

最小平方误差准侧等价于 $\min_\Theta J(\Theta)$ 如何求解？

矩阵形式求解析解

$$\boldsymbol{X}=\begin{bmatrix} x_{10} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{np} \end{bmatrix}=\begin{bmatrix} \boldsymbol{x}^{(1)T} \\ \vdots \\ \boldsymbol{x}^{(n)T} \end{bmatrix} \qquad \mathbf{Y}=\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$\hat{\mathbf{Y}}=\mathbf{X}\Theta$$

$$J(\Theta)=\frac{1}{2}(X\Theta-Y)^T(X\Theta-Y)=\frac{1}{2}(\Theta^TX^TX\Theta-2\Theta^TX^TY-Y^TY)$$

$$\frac{\partial J(\Theta)}{\partial\Theta}=0\Rightarrow 2X^TX\Theta-2X^TY=0$$

$$\frac{\partial x^TAx}{\partial x}=Ax+A^Tx$$

当$\mathbf{X}^{\mathrm{T}}\mathbf{X}$满秩的时候有：$\hat{\Theta}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$

向量求导参考：https://en.wikipedia.org/wiki/Matrix_calculus

梯度下降法（数值解法）

$$\frac{\partial J(\Theta)}{\theta_j} = \sum_{i=1}^{n} (f_\Theta(\mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

$$\theta_j(t) = \theta_j(t-1) - \alpha \sum_{i=1}^{n} (f_\Theta(\mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

其中α为学习率

（当$\mathbf{X}^T\mathbf{X}$不满秩时也可以用）

从概率的角度理解线性回归

$$y^{(i)} = \Theta^{T}\mathbf{x}^{(i)} + \varepsilon^{(i)}$$

$\varepsilon^{(i)}$为满足高斯分布的独立同分布随机变量, 即

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2})$$

$$p(y^{(i)} \mid \mathbf{x}^{(i)};\Theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y^{(i)} - \Theta^{T}\mathbf{x}^{(i)})^2}{2\sigma^2})$$

理解为在参数Θ下，给定$\mathbf{x}^{(i)}$，$y=y^{(i)}$的概率

最大似然准则

课程第四章将详细介绍最大似然估计，书p45页

定义似然函数（likelihood function）：

$$L(\Theta)=\prod_{i=1}^{n} p(y^{(i)} \mid \mathbf{x}^{(i)};\Theta)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\Theta^{T}\mathbf{x}^{(i)})^{2}}{2\sigma^{2}}\right)$$

最大似然准则： $\max_{\Theta}(L(\Theta))$

Log likelihood：$l(\Theta)=\log(L(\Theta))=n\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^{2}}\sum_{i=1}^{n}(y^{(i)}-\Theta^{T}\mathbf{x}^{(i)})^{2}$

最大化$l(\Theta)$等价于最小化 $\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\Theta^{T}\mathbf{x}^{(i)})^{2}$

与前面定义的准则函数等价 $\min_{\Theta} J(\Theta)$

定性变量的回归

- 例如某一维特征的取值为｛男，女｝
- 引入虚拟变量(dummy variable)

$$x_1 = \begin{cases} 1 & \text{该样本是男生} \\ 0 & \text{该样本是女生} \end{cases}$$

- 这时的回归模型为：

$$y^{(i)} = \theta_0 + \theta_1 {x_1}^{(i)} + \varepsilon^{(i)} = \begin{cases} \theta_0 + \theta_1 + \varepsilon^{(i)} & \text{男生} \\ \theta_0 + \varepsilon^{(i)} & \text{女生} \end{cases}$$

线性模型的扩展

- 变量之间的交互项(interaction term)

- 输入变量与输出变量之间不是简单线性关系
  例如$y \sim x^2$

关于模型参数的显著性检验*

空假设：$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

备择假设：$H_a :$ at least one β_j is non-zero

F统计量：

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

线性分类方法

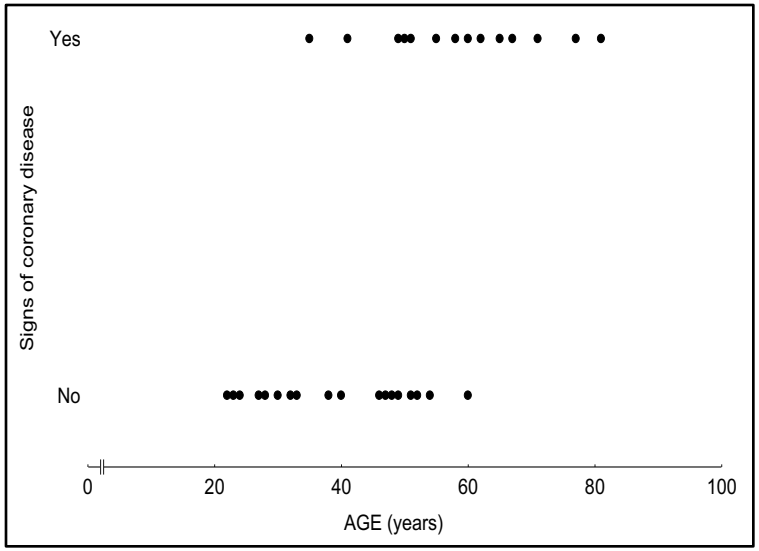
Logistic regression

- 如果Y的取值是｛0,1｝

- 我们希望 $x \to \infty$ 时 $y \to 1$

  $x \to -\infty$ $y \to 0$

心脏病与年龄的关系



一种取值方式：

$$f_\Theta(\mathbf{x}) = \frac{1}{1+e^{-\Theta^T\mathbf{x}}}$$

Logistic function, or sigmoid function

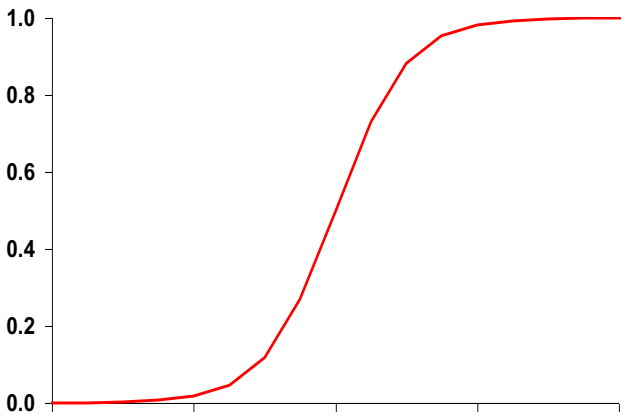
Logit 函数特点

- 该函数形式便于求导

对于 $g(z)=\frac{1}{1+e^{-z}}$ $g(z)'=g(z)(1-g(z))$

$$P(y=1\mid x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\log\frac{P(y=1\mid x)}{1-P(y=1\mid x)}=\alpha+\beta x$$



最大似然法求解

求解概率问题：

$$P(y=1 \mid \mathbf{x};\Theta) = f_\Theta(\mathbf{x})$$

$$P(y=0 \mid \mathbf{x};\Theta) = 1 - f_\Theta(\mathbf{x})$$

等价于：$P(y \mid \mathbf{x};\Theta) = f_\Theta(\mathbf{x})^y (1 - f_\Theta(\mathbf{x}))^{1-y}$

对所有训练样本的似然函数：

$$L(\Theta) = \prod_{i=1}^{n} f_\Theta(\mathbf{x}^{(i)})^{y^{(i)}} (1 - f_\Theta(\mathbf{x}^{(i)}))^{1-y^{(i)}}$$

梯度下降法求解：

$$\theta_j(t) = \theta_j(t-1) - \alpha \sum_{i=1}^{n} (f_\Theta(\mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

广义线性 logistic regression

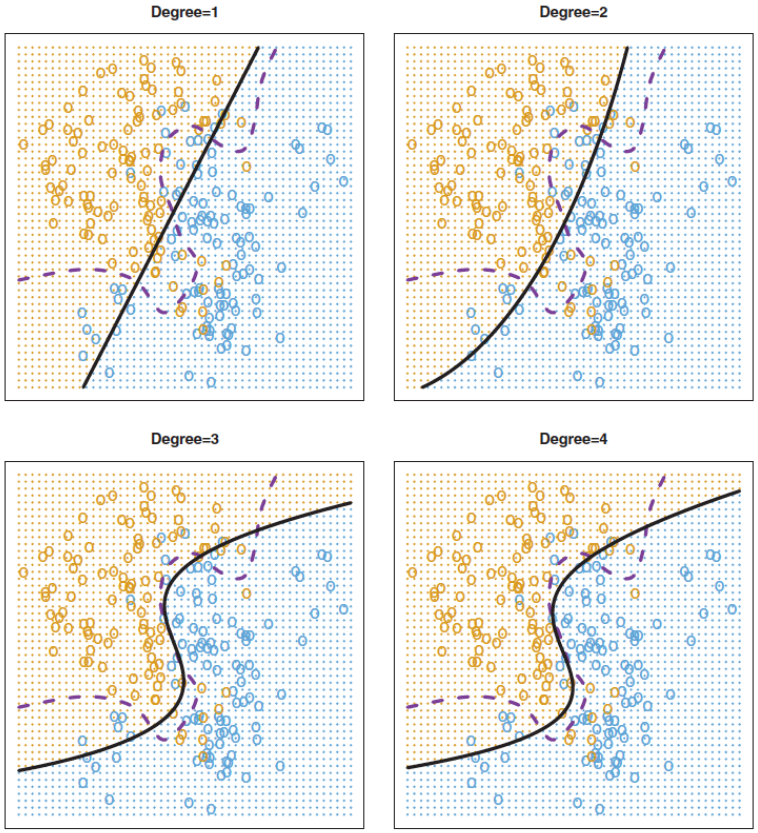


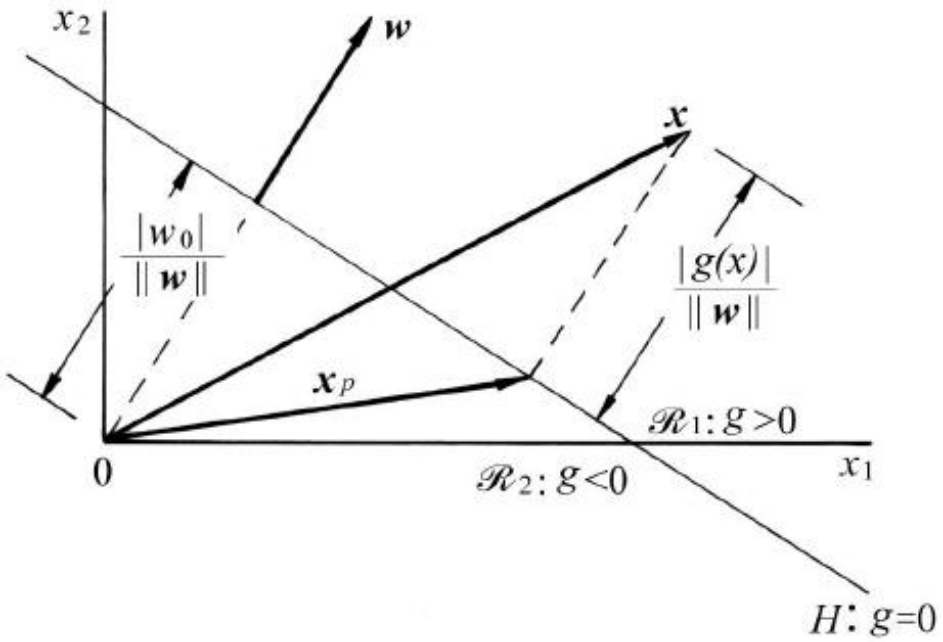
FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively* 0.201, 0.197, 0.160, *and* 0.162, *while the Bayes error rate is* 0.133.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

Gareth James et al., *An Introduction to Statistical Learning with Applications in R*, Springer, 2015

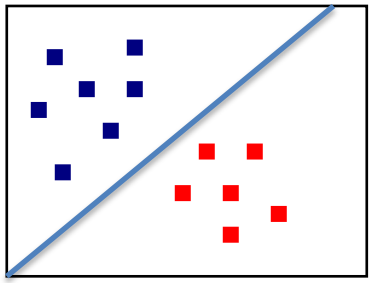
线性判别函数

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$



感知器（Perceptron）

如何直接求分类面：$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$



写成增广向量的形式：$z = g(X) = \Theta^{\mathrm{T}}\mathbf{X}$

决策函数：
$$f(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{other wise} \end{cases}$$

求解权向量 Θ ：
$$\text{if } y_i \in \omega_1, \text{then } \Theta^{\mathrm{T}}\mathbf{X}_i > 0$$
$$\text{if } y_i \in \omega_2, \text{then } \Theta^{\mathrm{T}}\mathbf{X}_i < 0$$

感知器（Perceptron）

- 令第一类样本$y_i=1$, 第二类样本$y_i=0$

损失函数： $J(\Theta)=\sum_{i=1}^{n}(y^{(i)}-f_{\Theta}(z^{(i)})(-\Theta^{\mathrm{T}}\mathbf{X}^{(i)})$

目标： $\min_{\Theta} J(\Theta)$

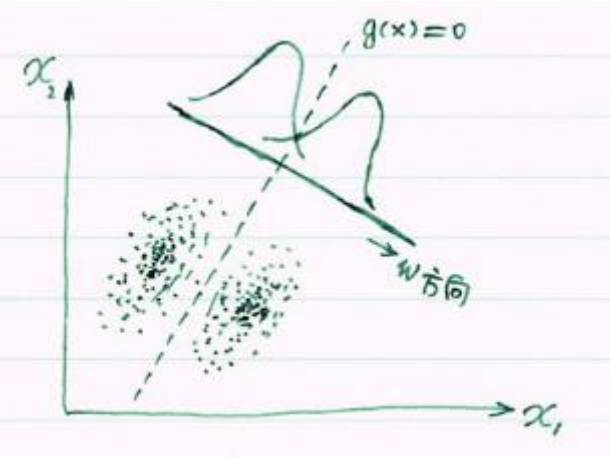
通过梯度下降法求解：

$$\theta_j(t)=\theta_j(t-1)+\alpha\sum_{i=1}^{n}(y^{(i)}-f_{\Theta}(\Theta^{\mathrm{T}}\mathbf{X}^{(i)})x_j^{(i)}$$

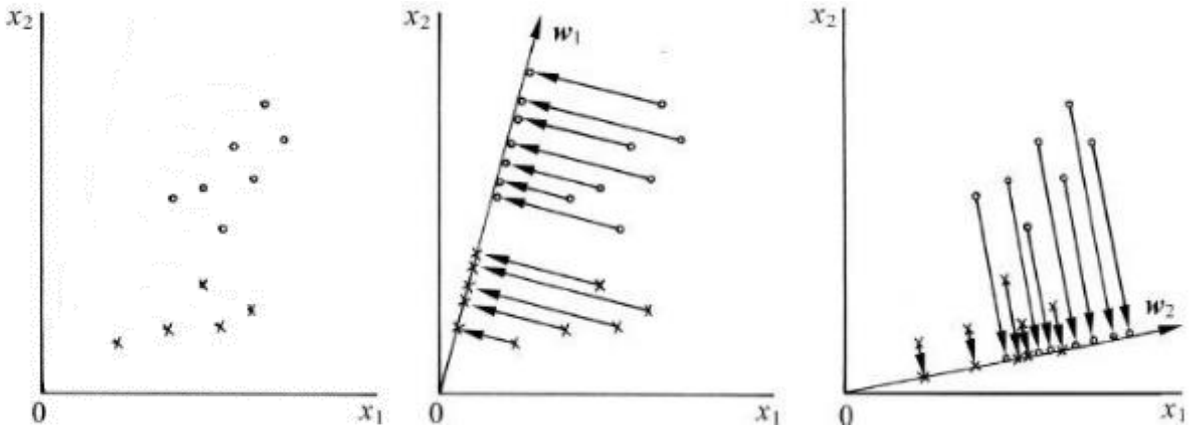
Fisher线性判别

出发点：把所有样本都投影到一维，
使在投影线上最易于分类。

———　寻找投影方向



投影　$y_i = \mathbf{w}^T \mathbf{x}_i$



使两类之间尽可能分开，各类内部尽可能聚集

Fisher准则函数(Fisher's Criterion)

在Y空间（一维投影）：

$$\max J_F(w) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

类均值：
$$\tilde{\mu}_i = \frac{1}{m_i} \sum_{x_j \in \mathcal{Y}_i} y_j, \ i = 1,2$$

类内散度：
$$\tilde{S}_i^2 = \sum_{x_j \in \mathcal{Y}_i} (y_j - \tilde{\mu}_i)^2, \ i = 1,2$$

代入 $y=\mathbf{w}^T\mathbf{x}$，可得：

$$J_F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

求解投影方向

- 目标：求解 $\mathbf{w}^* = \max \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$

  分子分母都是w的二次函数，取值与w长度无关，只与方向有关。

  不妨设 $\mathbf{w}^T S_w \mathbf{w} = c \neq 0$ ，最大化分子

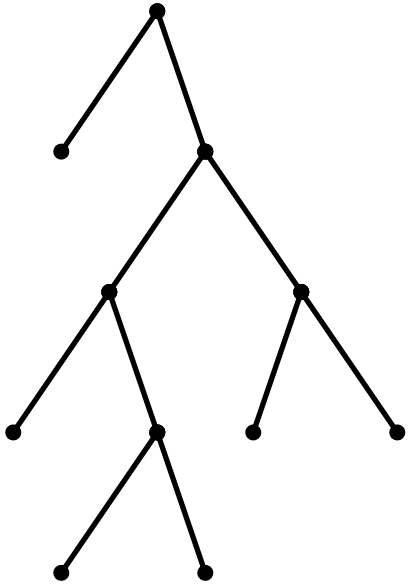
- Lagrange乘子法求解，得到

$$\mathbf{w}^* = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

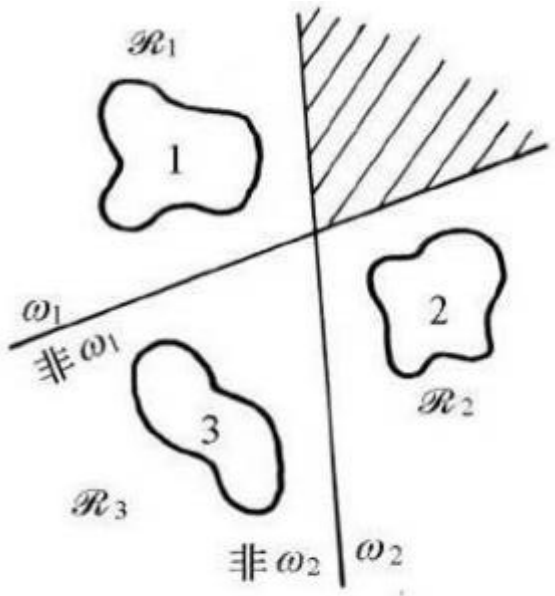
只给出投影方向，没有给出阈值。如何选阈值？

关于多类划分问题的线性判别

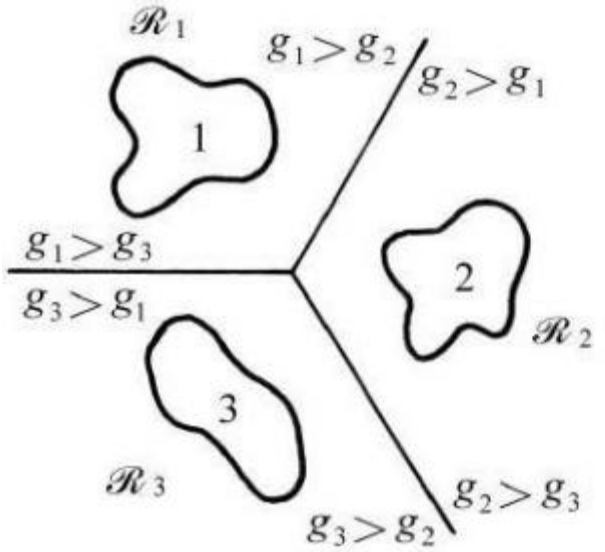
思路一：转化为多个两类问题
思路二：多个判别函数



用二叉树实现多类分类



(a) ω_i/非ω_i



(a) 三类

线性模型回答的问题

- 自变量X和因变量Y之间是否有联系(relationship)
- 如果有联系，如何进行量化
- X的变化能在多大程度上解释Y的变化
- X的哪些分量对预测Y有作用
- X和Y之间的关系是否是线性的
- 自变量之间是否有协同作用(synergy)
- ……

本章小结

- 线性回归模型及其求解方法
- 广义线性回归模型
- 线性分类器
  - Logistic回归
  - 线性判别函数
  - 感知机器
  - Fisher线性判别