线性分类模型(1)

引言

- 目标：将输入矢量 x 赋给 K 个离散类别 $c_k, k=1,\ldots,K$ 之一
  - 绝大多数情形下，类别之间不相交，称为互斥的。
- 术语
  - 决策区域(decision regions)：类别在输入空间所占据的区域
  - 决策边界(decision boundaries)或决策表面(decision surfaces)：决策区域的边缘
- 线性分类模型：解决分类问题的一类模型
  - 决策表面：输入矢量 x 的线性函数，D 维输入空间中的 D-1 维超平面
- 线性可分(linearly separable)
  - 数据集可以被线性决策表面完全地分开

引言

- 目标变量(target variable)
  - 两类问题
    - 单个目标变量 $t \in \{0,1\}$，$t=1$ 表示类别 c_1，$t=0$ 表示类别 c_2。
    - 单个目标变量 $t \in [0,1]$，表示属于类别 c_1 的概率
  - 多类问题
    - t 是一个长度为 K 的矢量，如果是类别 c_j，则除了 $t_j=1$ 之外全为零
    - 同样，每个分量的取值也可以理解为属于对应类别的概率

决策方法

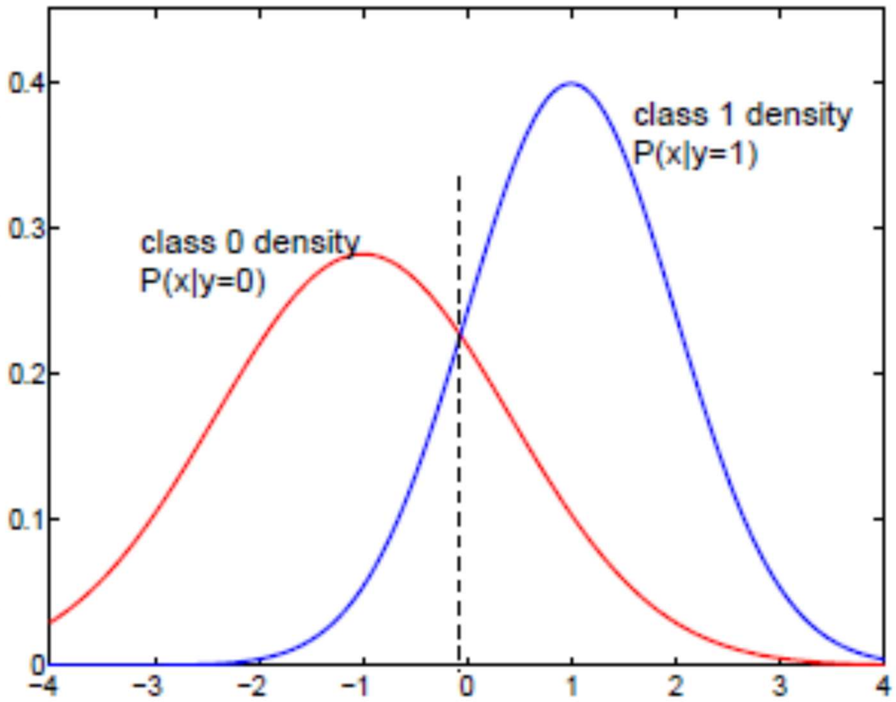
- 判别式(discriminant)方法
  - 最简单的方法
  - 直接将每个矢量 x 赋给特定类别
- 直接对条件概率 $p(c_k \mid x)$ 建模
  - 如将条件概率表示为参数模型，使用训练样本集合优化参数。
- 产生式(generative)方法
  - 对类别条件密度 $p(x \mid c_k)$ 建模
  - 对类别先验概率 $p(c_k)$ 建模
  - 使用 Bayes 定理计算后验概率

$$p(c_k|x) = \frac{p(x|c_k)\,p(c_k)}{p(x)}$$

决策理论

- 假如已知每个类别样本的分布，即 $p(x \mid y=0)$ 和 $p(x \mid y=1)$。如何给出新样本 x' 的最优决策？
- 最优决策
  - 最优：最小错误分类
  - 方法：基于对数似然比

$$y=\begin{cases}1 & \text{if } \log\dfrac{p(x' \mid y=1)}{p(x' \mid y=0)}>0 \\ 0 & \text{otherwise}\end{cases}$$



决策理论

- 当某类样本数目多于另一个类别时，需要修改决策规则

$$y=\begin{cases}1 & \text{if } \log\dfrac{p(x'\mid y=1)P(y=1)}{p(x'\mid y=0)P(y=0)}>0\\ 0 & \text{otherwise}\end{cases}$$

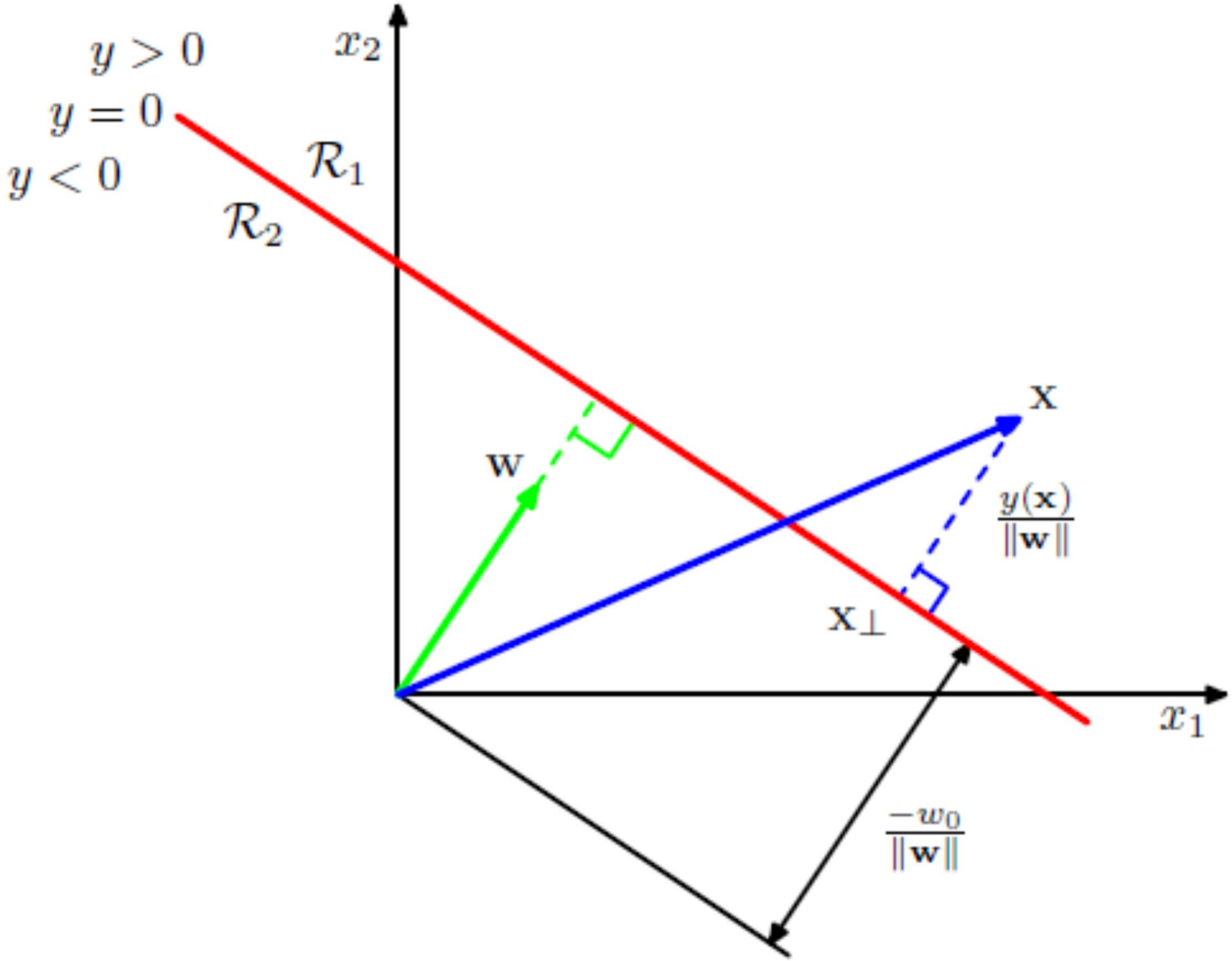
- Bayes 最优决策规则

$$\begin{aligned}y' &= \arg\max_{y\in\{0,1\}}\left\{p(x'\mid y)P(y)\right\}\\ &= \arg\max_{y\in\{0,1\}}\{P(y\mid x')\}\end{aligned}$$

  - 只有当拥有正确的密度函数和先验频度时，才是最优的。

线性判别式

❖ 二维空间线性判别式函数的几何解释



线性判别式

- 如果 x 是决策面上的一个点，则 $y(\mathbf{x})=0$，从原点到决策面的法线距离为

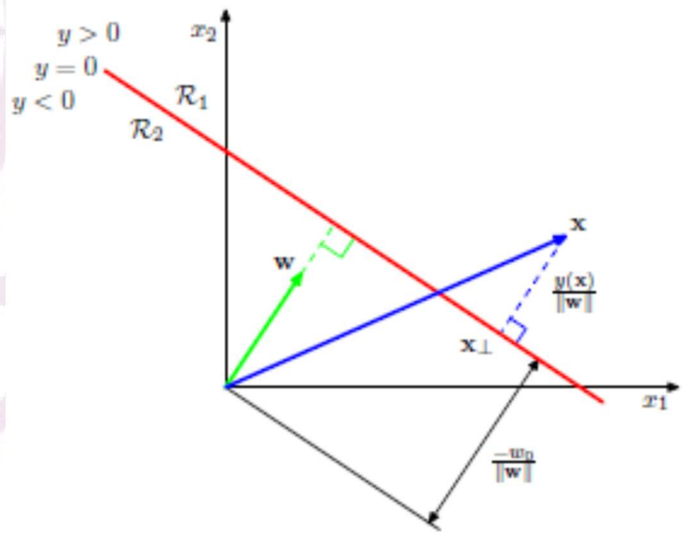
$$\frac{\mathbf{w}^{\mathrm{T}}\mathbf{x}}{\|\mathbf{w}\|}=-\frac{w_0}{\|\mathbf{w}\|}$$

  - 偏置参数 w_0 决定了决策面的位置
- 点 x 到决策面的垂直距离 r
  - 任一点 x 和它在决策面的正交投影$\mathbf{x}_\perp$，有

$$\mathbf{x}=\mathbf{x}_\perp+r\frac{\mathbf{w}}{\|\mathbf{w}\|}\Rightarrow$$

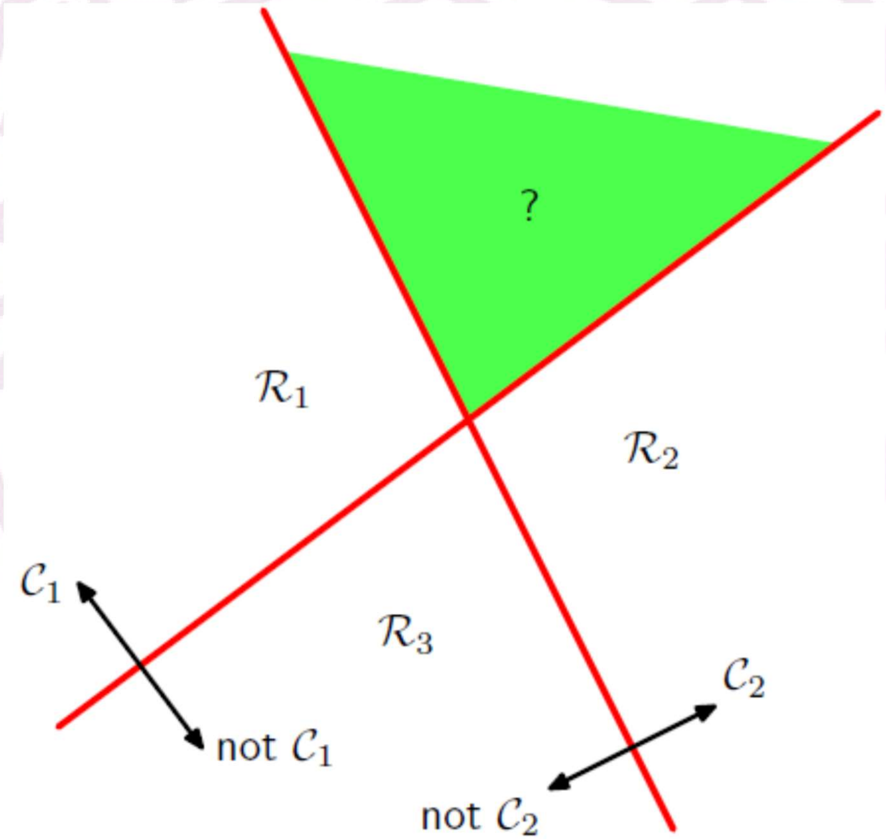
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0=\mathbf{w}^{\mathrm{T}}\mathbf{x}_\perp+r\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|}+w_0\Rightarrow$$

$$y(\mathbf{x})=y(\mathbf{x}_\perp)+r\|\mathbf{w}\|\Rightarrow r=\frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$



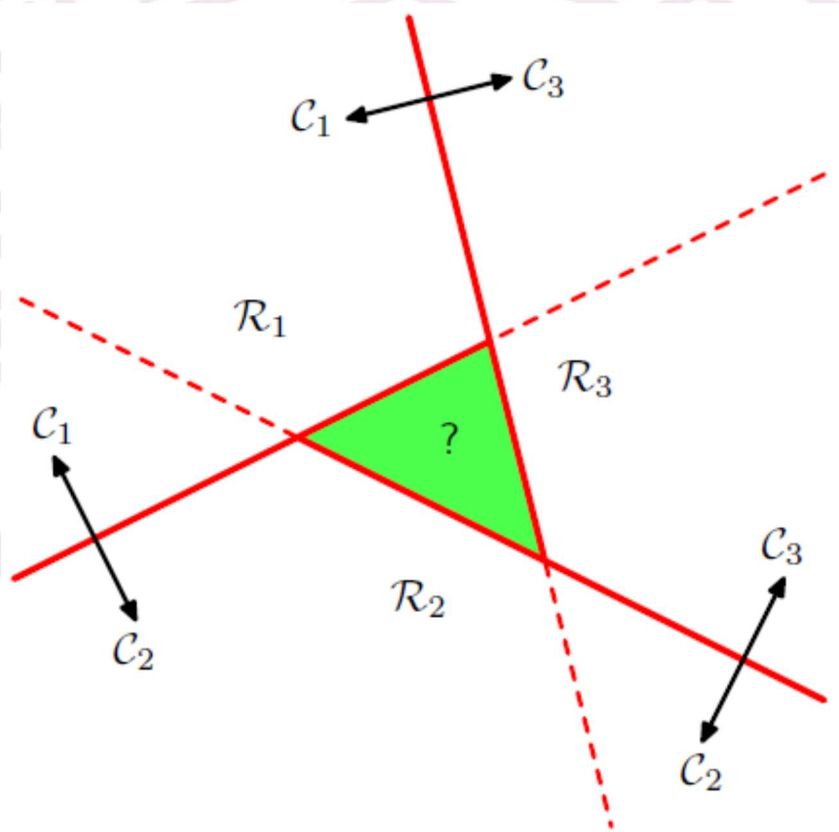
多类问题

- “一对其它(one-versus-the-rest)” 分类器
  - 使用 K – 1 个两类分类器
  - 每个分类器将属于特定类别 c_k 的样本与不属于该类的样本分开
  - 问题：导致“绿色”的歧义区域



多类问题

- “一对一(one-versus-one)”分类器
  - 使用 K(K – 1)/2 个两类分类器
  - 每个分类器解决每个可能的一对类别，根据多数票原则进行分类决策
  - 问题：仍然存在歧义区域



多类问题

- K 类判别式
  - 由 K 个线性函数构成：

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

  - 决策规则：

$$\mathbf{x} \in c_k, \text{ if } y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for all } j \neq k$$

  - 类别c_k和c_j 之间的决策边界是 (D – 1) 维超平面

$$\left(\mathbf{w}_k - \mathbf{w}_j\right)^{\mathrm{T}}\mathbf{x} + \left(w_{k0} - w_{j0}\right) = 0$$

多类问题

- 决策区域是单连通和凸的
  - 考虑决策区域$\mathcal{R}_k$ 内的两个点 $\mathbf{x}_A, \mathbf{x}_B$，二者连线上的任一点 $\hat{\mathbf{x}}$ 表示为

$$\hat{\mathbf{x}} = \lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B$$

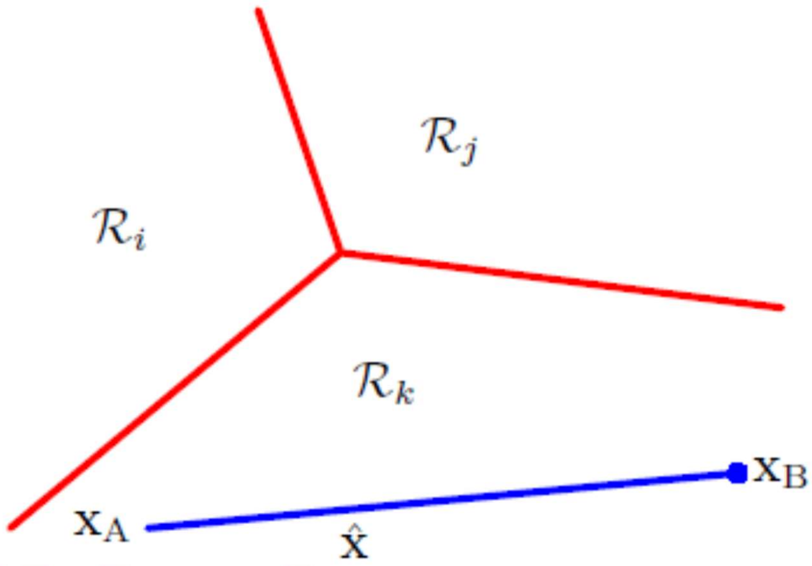
    其中：$0 \le \lambda \le 1$
  - 利用判别式函数的线性关系，有

$$y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_A) + (1-\lambda) y_k(\mathbf{x}_B)$$

  - 因为，存在

$$\left.\begin{array}{l} y_k(\mathbf{x}_A) > y_j(\mathbf{x}_A) \\ y_k(\mathbf{x}_B) > y_j(\mathbf{x}_B) \end{array}\right. \text{for all } j \neq k$$

  - 故，$y_k(\hat{\mathbf{x}}) > y_j(\hat{\mathbf{x}})$
  - 所以，$\hat{\mathbf{x}}$ 也在 $\mathcal{R}_k$ 内，因此$\mathcal{R}_k$是单连通和凸的。



最小二乘法

- 问题
  - K 类分类问题
  - 目标变量 t 使用 1-of-K 二值编码
- 每类 c_k 使用线性模型表示

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0},\ \ k = 1,\ldots,K$$

矩阵形式

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{x}}$$

其中，矩阵 $\tilde{\mathbf{W}}$ 的第 k 列是 D+1 维矢量 $\tilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^{\mathrm{T}})^{\mathrm{T}}$， $\tilde{\mathbf{x}}$ 是对应的增广输入矢量 $(1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$。

- 未知输入 $\mathbf{x}$ 的类别是

$$c_k = \arg\max_k y_k = \arg\max_k \tilde{\mathbf{w}}_k^{\mathrm{T}}\tilde{\mathbf{x}}$$

最小二乘法

- 最小化平方和误差函数
  - 训练数据集：$\{\mathbf{x}_n, \mathbf{t}_n\}, n=1,\ldots N$
  - 定义：矩阵 T 的第 n 行是矢量 $\mathbf{t}_n^{\mathrm{T}}$，矩阵 $\tilde{\mathbf{X}}$ 的第 n 行是 $\tilde{\mathbf{x}}_n^{\mathrm{T}}$。
  - 平方和误差函数

$$\mathbb{E}_{\mathcal{D}}(\tilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{\left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}\right)^{\mathrm{T}}\left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}\right)\right\}$$

  - 令对 $\tilde{\mathbf{W}}$ 的导数为零，获得解

$$\tilde{\mathbf{W}} = \left(\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{X}\right)^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \tilde{\mathbf{X}}^{?}\mathbf{T}$$

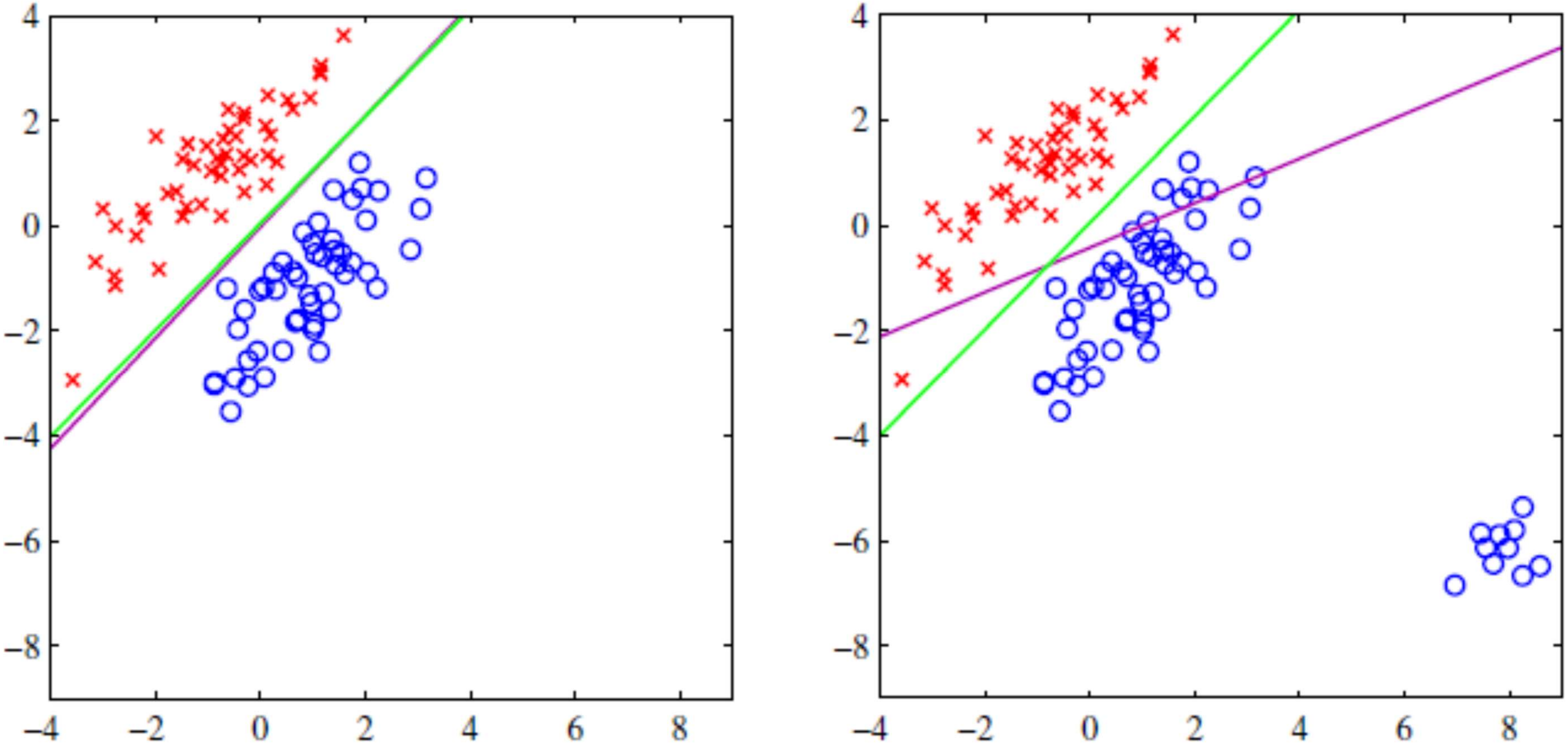
其中： $\tilde{\mathbf{X}}^{\dagger}$是 $\tilde{\mathbf{X}}$的伪逆矩阵

  - 判别式函数

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\tilde{\mathbf{X}}^{?}\right)^{\mathrm{T}}\tilde{\mathbf{x}}$$

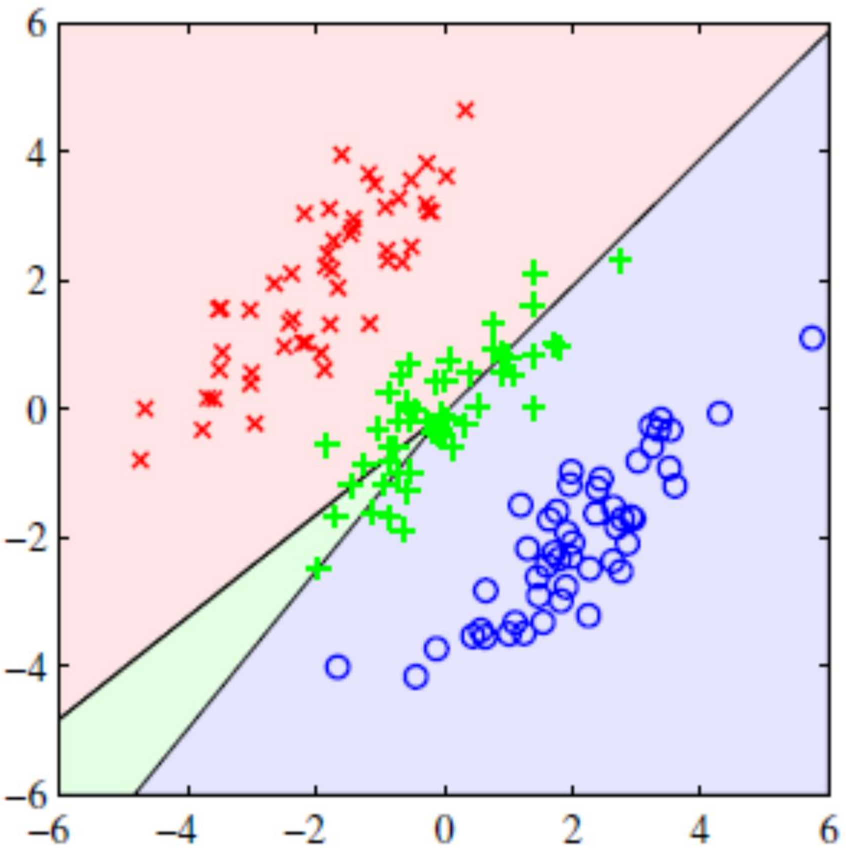
最小二乘法

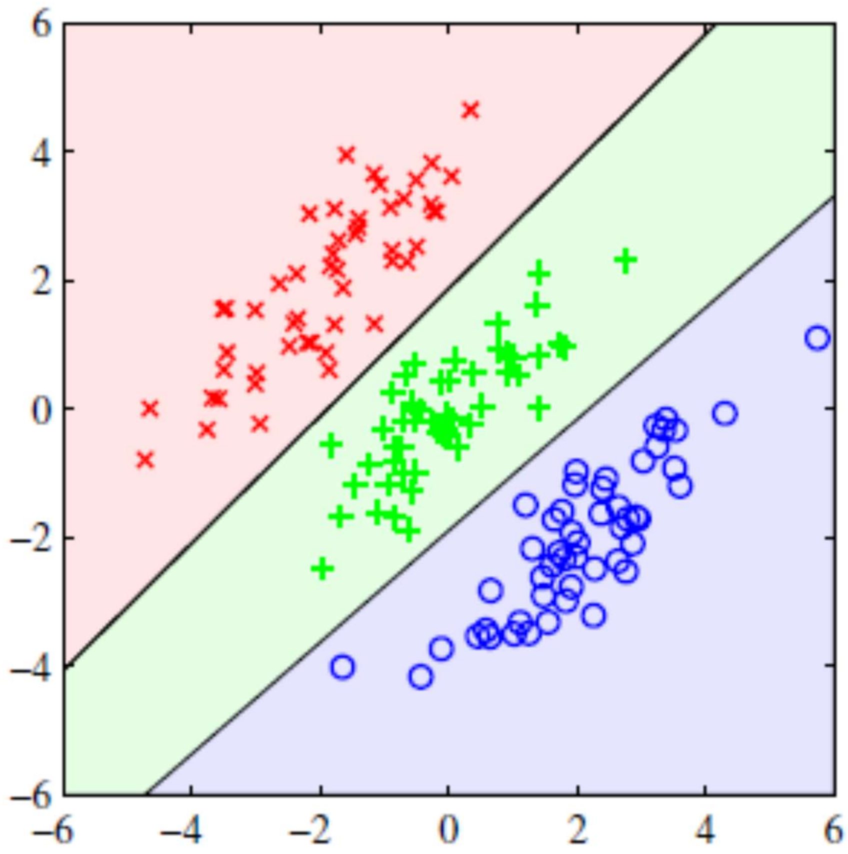
- **缺点**
  - 对离群点缺乏鲁棒性



最小二乘法

- 因为，最小二乘法对应正态条件分布假设下的最大似然，但二值目标矢量的分布离正态分布太远，故，造成问题。
  - 左边：最小二乘法
  - 右边：逻辑回归





Fisher线性判别式

- 两类问题
  - D 维输入矢量 x 投影到一维空间，公式为

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

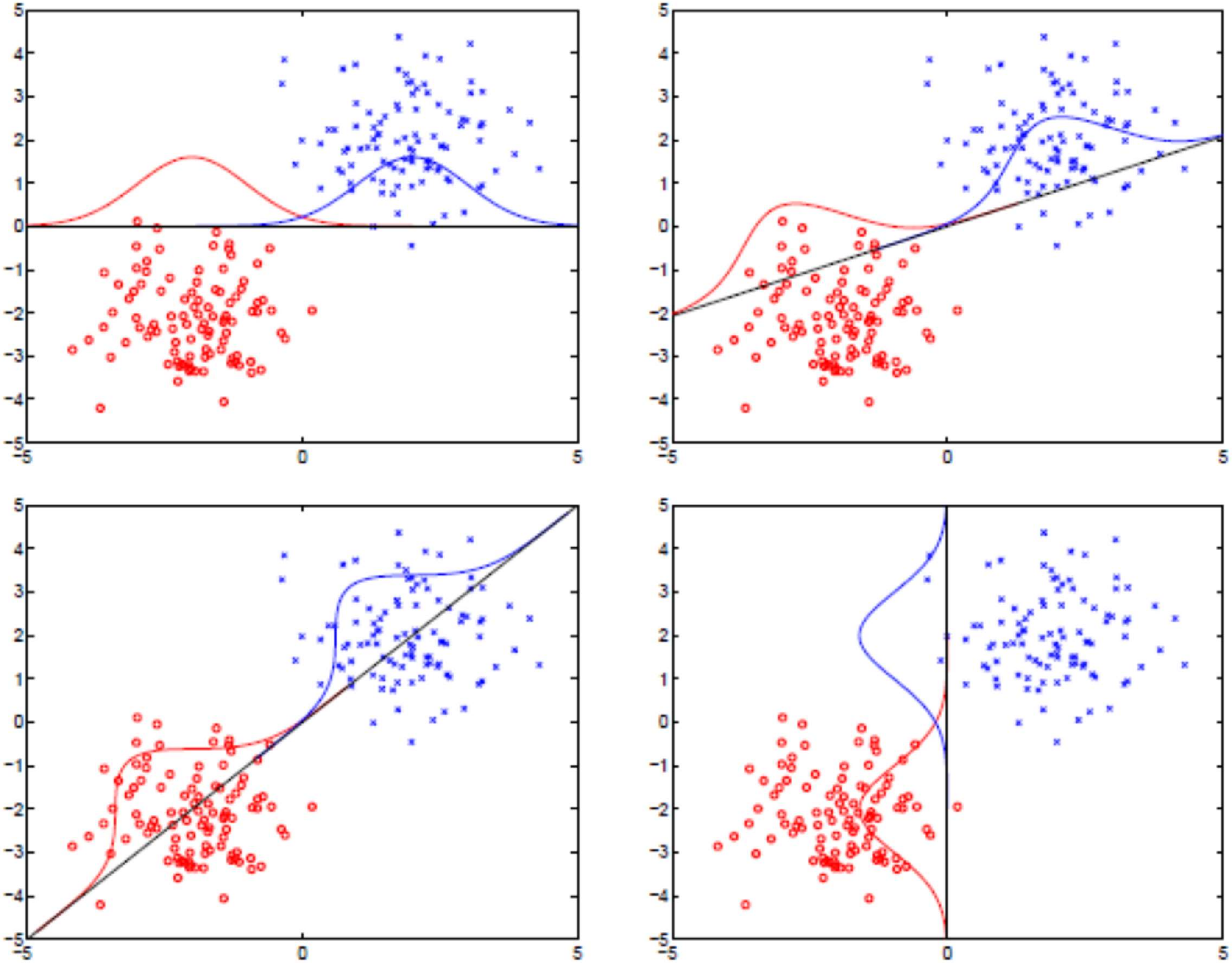
  - 构造线性判别式

$$\mathbf{x} \in \begin{cases} c_1 & \text{if } y \geq -w_0 \\ c_2 & \text{otherwise} \end{cases}$$

- 问题
  - 投影到一维会造成信息损失
  - 在 D 维空间中可分离类别可能在一维空间发生重叠

Fisher线性判别式

- 通过改变 **w**，可以使类别之间分开的程度不同。



Fisher线性判别式

- 假设
  - 类别 c_1 有 N_1 个点，类别 c_2 有 N_2 个点
  - 均值矢量

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in c_1}\mathbf{x}_n \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in c_2}\mathbf{x}_n$$

- 最简单度量：投影到 w 上，类别均值投影之间距离最大化

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$$

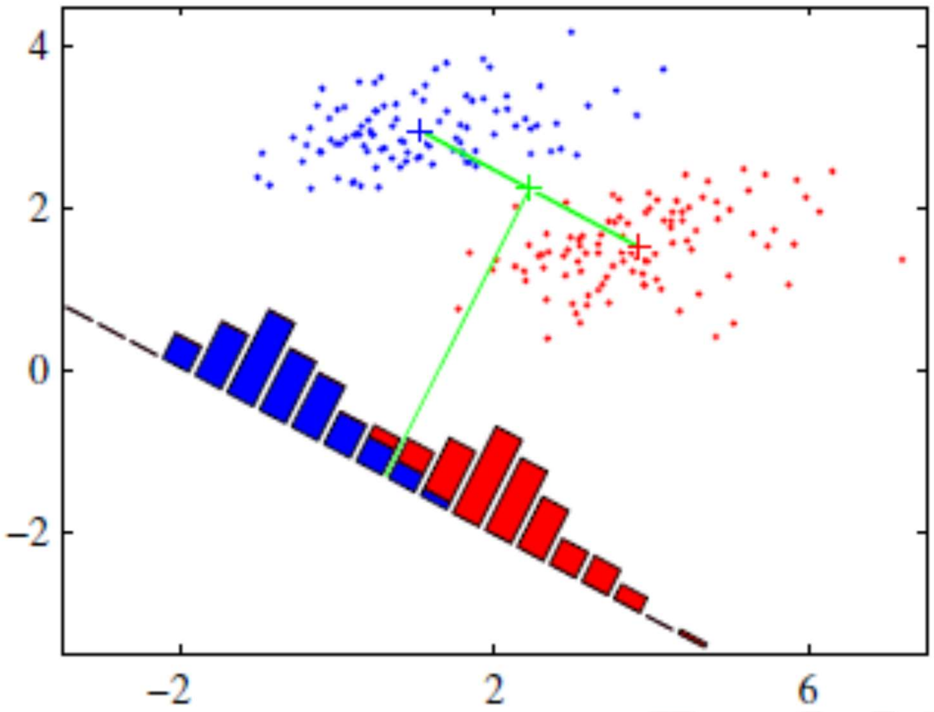
但投影距离与 w 的幅值有关。

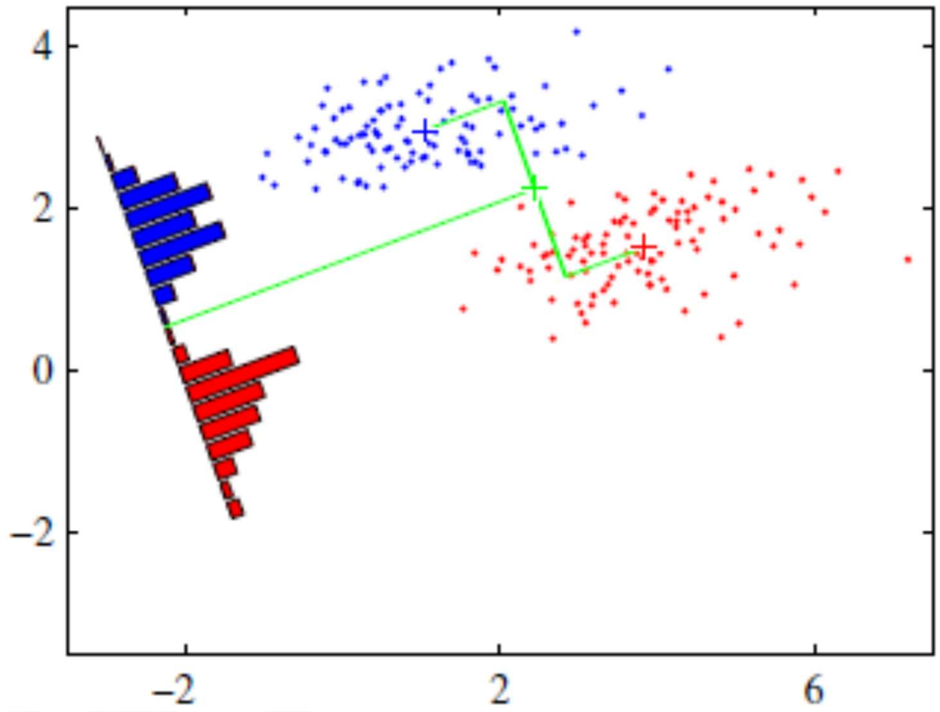
- 拉格朗日乘子法
  - 约束 w 是单位矢量，排除其对最大化操作的影响
  - 得到解

$$\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$$

Fisher线性判别式

- 投影直方图
  - 左：投影到类别均值连线
  - 右：Fisher 线性判别式

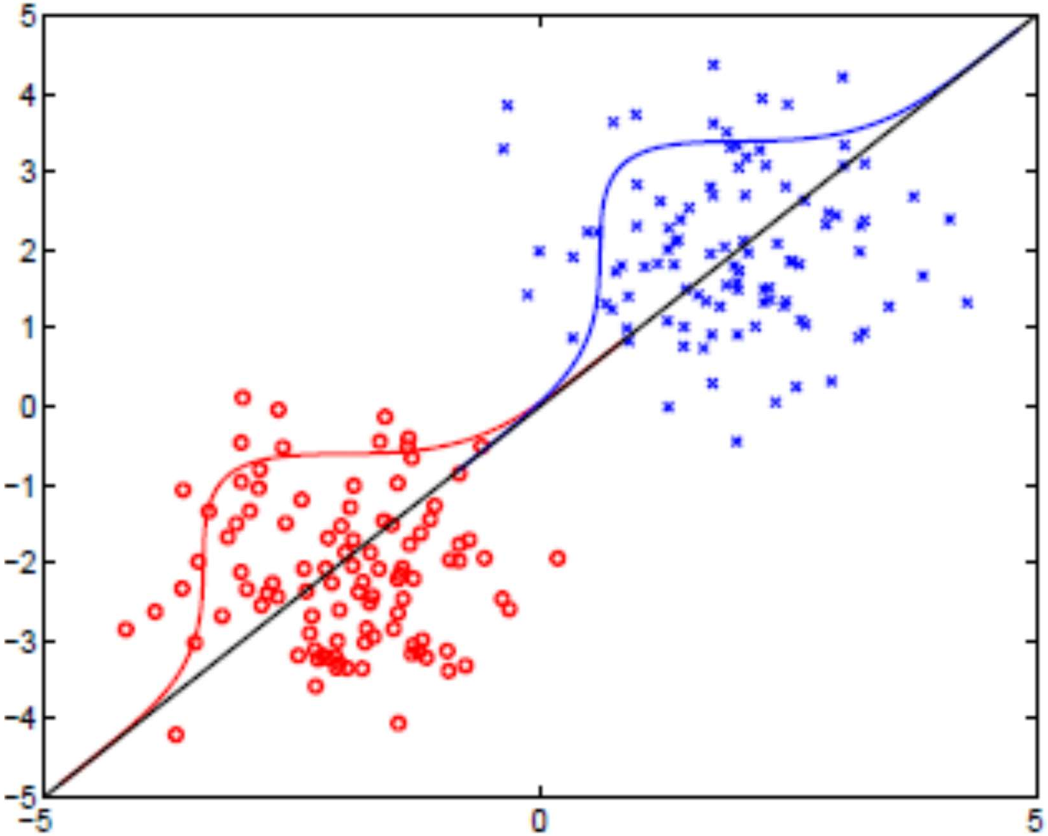




  - 投影重叠的原因：类别点分布的协方差矩阵非对角化

Fisher线性判别式

- 优化目标：在输入空间寻找一个方向 w，使得投影点变得“很好分开”
  - 类均值投影之间具有较大分离间隔
  - 每个类别内部较小方差



Fisher线性判别式

- 定义
  - c_k 类点投影的类内方差

$$s_k^2 = \sum_{n \in c_k} (y_n - m_k)^2$$

  - 所有数据点的总类内方差 $s_1^2 + s_2^2$

- Fisher 准则

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{B}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{W}} \mathbf{w}}$$

  - 类间协方差矩阵(between-class covariance matrix)

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

  - 类内协方差矩阵(within-class covariance matrix)

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in c_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in c_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$$

Fisher线性判别式

- $J(\mathbf{w})$ 对 w 求导，得到其最大化的条件是

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_w\mathbf{w} = (\mathbf{w}^T\mathbf{S}_w\mathbf{w})\mathbf{S}_B\mathbf{w}$$

  - 利用 $\mathbf{S}_B\mathbf{w}$ 总是在 $(\mathbf{m}_2 - \mathbf{m}_1)$ 的方向，且不关心 w 的幅值，化简得到 Fisher 线性判别式（投影方向）

$$\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

  - 两类样本是协方差矩阵相等的正态分布时，解是Bayes最优的

- 点评
  - 虽然不是传统意义的判别式形式，但经过投影后，非常容易得到线性判别式函数。

最小二乘法 vs. Fisher线性判别式

- 思路
  - 最小二乘法：模型预测尽可能地接近目标值集合
  - Fisher 线性判别式：在输出空间样本类别具有最大可分离性
- 二者关系
  - 令：类别c_1 的目标值为 N/N_1，近似先验概率的倒数；类别 c_2的目标值为 $-N/N_2$，其中N_1, N_2, N 分别为类别 c_1、类别c_2 和总体的样本数。
  - 平方和误差函数

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)^2$$

  - 令 E 对权值的导数为零，得到

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right) = 0$$

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)\mathbf{x}_n = 0$$

最小二乘法 vs. Fisher线性判别式

- 对第一个方程，利用

$$\sum_{n=1}^{N} t_n = N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2} = 0$$

得到

$$w_0 = -\mathbf{w}^{\mathrm{T}} \mathbf{m}$$

其中

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \left(N_1 \mathbf{m}_1 + N_2 \mathbf{m}_2 \right)$$

- 对第二个方程，同样利用 t_n，得到

$$\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1 N_2}{N} \mathbf{S}_{\mathrm{B}} \right) \mathbf{w} = N \left(\mathbf{m}_1 - \mathbf{m}_2 \right)$$

利用各项的定义，并且忽略不相关的比例因子，推导可得到

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1} (\mathbf{m}_2 - \mathbf{m}_1)$$

最小二乘法 vs. Fisher线性判别式

❖ 因为，存在

$$w_0 = -\mathbf{w}^{\mathrm{T}}\mathbf{m}$$

所以，对未知矢量 x

$$\mathbf{x} \in \begin{cases} c_1 & \text{if } y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}(\mathbf{x}-\mathbf{m}) > 0 \\ c_2 & \text{otherwise} \end{cases}$$

❖ 结论：对于两类问题，Fisher 准则是最小二乘法的一个特例。

多类 Fisher 判别式

- 假设
  - 输入空间维数 D > 类别数目 K
  - $D' > 1$ 个线性特征 $y_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{x}$，其中 $k = 1, \ldots, D'$， 形成特征 y
  - 矩阵 W 的列矢量为 $\mathbf{w}_k$

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$$

- 在输入空间中，泛化各个量
  - K 类的类内协方差矩阵

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k$$

其中

$$\mathbf{S}_k = \sum_{n \in c_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}$$

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in c_k} \mathbf{x}_n$$

多类 Fisher 判别式

- 总协方差矩阵

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}}$$

其中

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N}\sum_{k=1}^{K} N_k \mathbf{m}_k$$

- 总协方差矩阵分解为

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}}$$

将 $\mathbf{S}_{\mathrm{B}}$认为是类间协方差矩阵的度量

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

多类 Fisher 判别式

- 在 D' 维 y-空间定义相似矩阵

$$\mathbf{s}_{\mathrm{W}} = \sum_{k=1}^{K}\sum_{n\in c_k}\left(\mathbf{y}_n-\mu_k\right)\left(\mathbf{y}_n-\mu_k\right)^{\mathrm{T}}$$

$$\mathbf{s}_{\mathrm{B}} = \sum_{k=1}^{K} N_k\left(\mu_k-\mu\right)\left(\mu_k-\mu\right)^{\mathrm{T}}$$

其中

$$\mu_k = \frac{1}{N_k}\sum_{n\in c_k}\mathbf{y}_n,\quad \mu = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k$$

- Fisher 准则

$$J(\mathbf{W}) = \mathrm{tr}\{\mathbf{s}_{\mathrm{W}}^{-1}\mathbf{s}_{\mathrm{B}}\} \Rightarrow J(\mathbf{w}) = \mathrm{tr}\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\}$$

  - 投影由 $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ 的 D' 个最大特征值对应的特征矢量决定。

多类 Fisher 判别式

- 强调
  - 因为

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

    为 K 个矢量外积之和（秩为 1）
  - 因为约束

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N}\sum_{k=1}^{K} N_k \mathbf{m}_k$$

    K 个矩阵（矢量外积）只有 K – 1个是独立的。
  - 由于S_{B}的秩最大值是 K – 1，所以最多有 K – 1个非零特征值
  - 结论：不可能发现超过 K – 1个线性特征

概率生成式模型

- 生成式(generative)方法
  - 建模类条件密度 $p(\mathbf{x}\,|\,c_k)$ 和类先验概率 $p(c_k)$
  - 根据 Bayes 定理计算后验概率 $p(c_k\,|\,\mathbf{x})$
- 两类问题
  - 类 c_1 的后验概率

$$p(c_1|\mathbf{x}) = \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_1)p(c_1)+p(\mathbf{x}|c_2)p(c_2)}$$

$$= \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha)$$

定义

$$\alpha = \ln\frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_2)p(c_2)}$$

概率生成式模型

❖ 定义：logistic sigmoid函数

$$\sigma(\alpha)=\frac{1}{1+\exp(-\alpha)}$$

- 也称为“挤压函数”
- 对称性

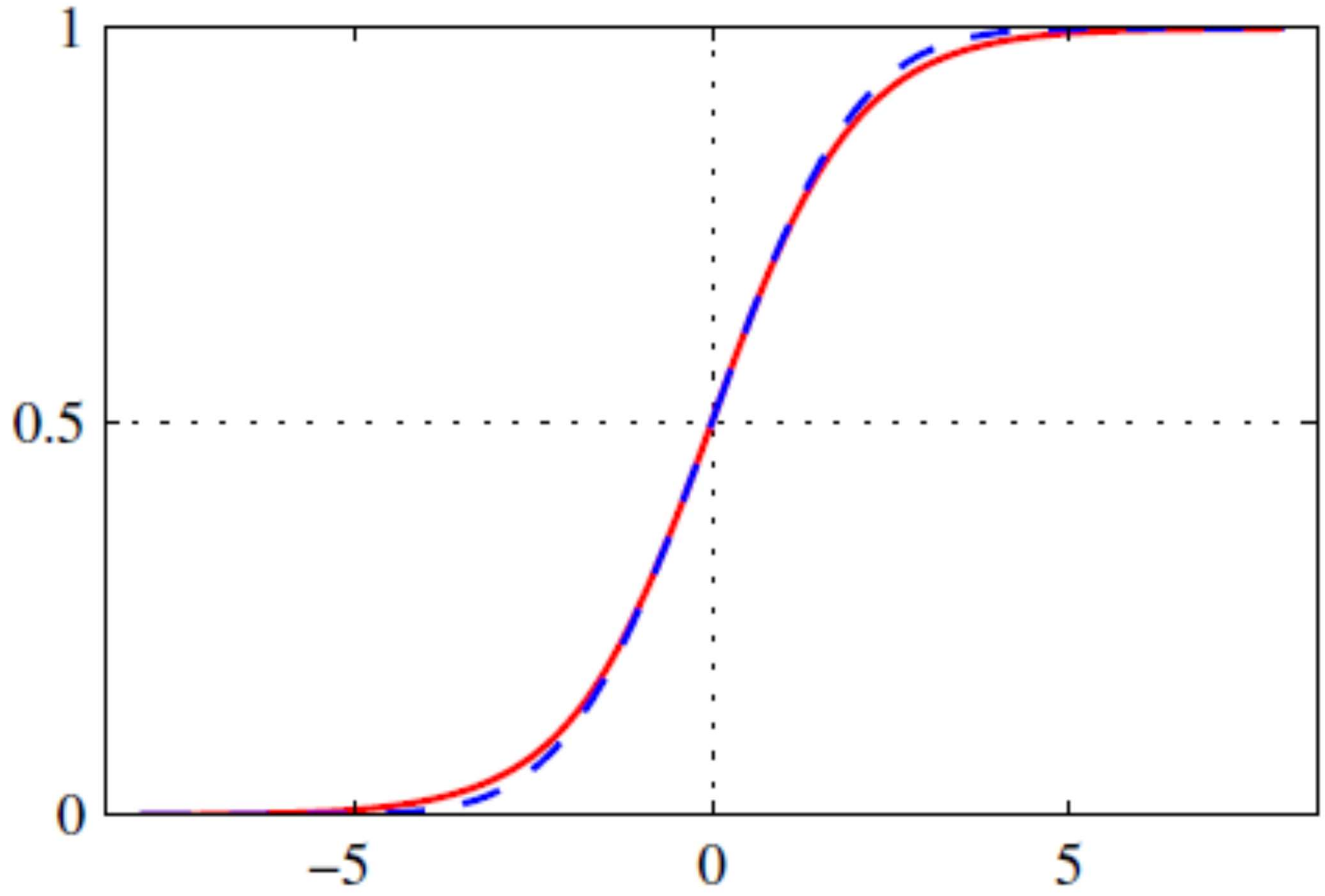
$$\sigma(-\alpha)=1-\sigma(\alpha)$$

- 逆函数(logit function)

$$\alpha=\ln\left(\frac{\sigma}{1-\sigma}\right)$$

概率生成式模型

- Logistic sigmoid函数（红色），标度probit函数（蓝色）



概率生成式模型

- 多类问题 K > 2
  - 后验概率

$$p(c_k|\mathbf{x}) = \frac{p(\mathbf{x}|c_k)p(c_k)}{\sum_j p(\mathbf{x}|c_j)p(c_j)}$$

$$= \frac{\exp(\alpha_k)}{\sum_j \exp(\alpha_j)}$$

就是归一化指数，也称为 Softmax 函数（平滑版的max函数）

  - 定义：

$$\alpha_k = \ln p(\mathbf{x}|c_k)p(c_k)$$

连续的输入

- 假设类条件密度是高斯分布
- 所有类别具有相同的协方差矩阵，类 c_k 的密度函数为

$$p(\mathbf{x}|c_k)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu_k)\right\}$$

- 两类问题

$$p(c_1|\mathbf{x})=\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0)$$

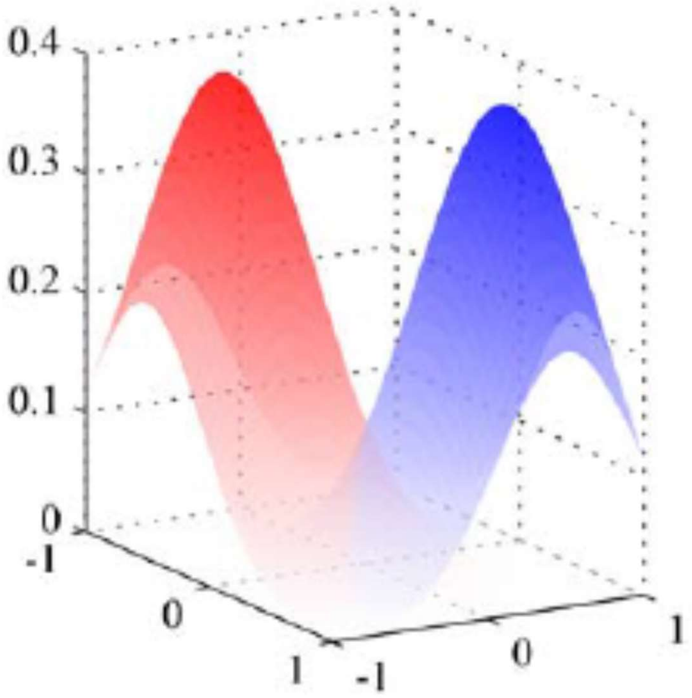
其中：

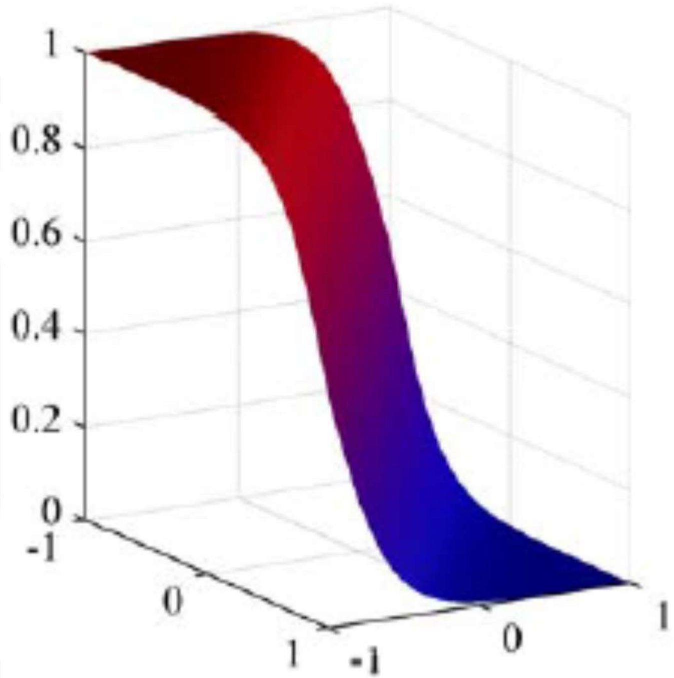
$$\mathbf{w}=\Sigma^{-1}(\mu_1-\mu_2)$$

$$w_0=-\frac{1}{2}\mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1+\frac{1}{2}\mu_2^{\mathrm{T}}\Sigma^{-1}\mu_2+\ln\frac{p(c_1)}{p(c_2)}$$

连续的输入

- 左：两类的类条件密度；右：对应的后验概率（x 线性函数的 logistic sigmoid）





- 决策边界（后验概率为常数）在输入空间是线性的
- 先验概率变化对决策边界的平行移动有影响

连续的输入

- K 类问题

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

其中：

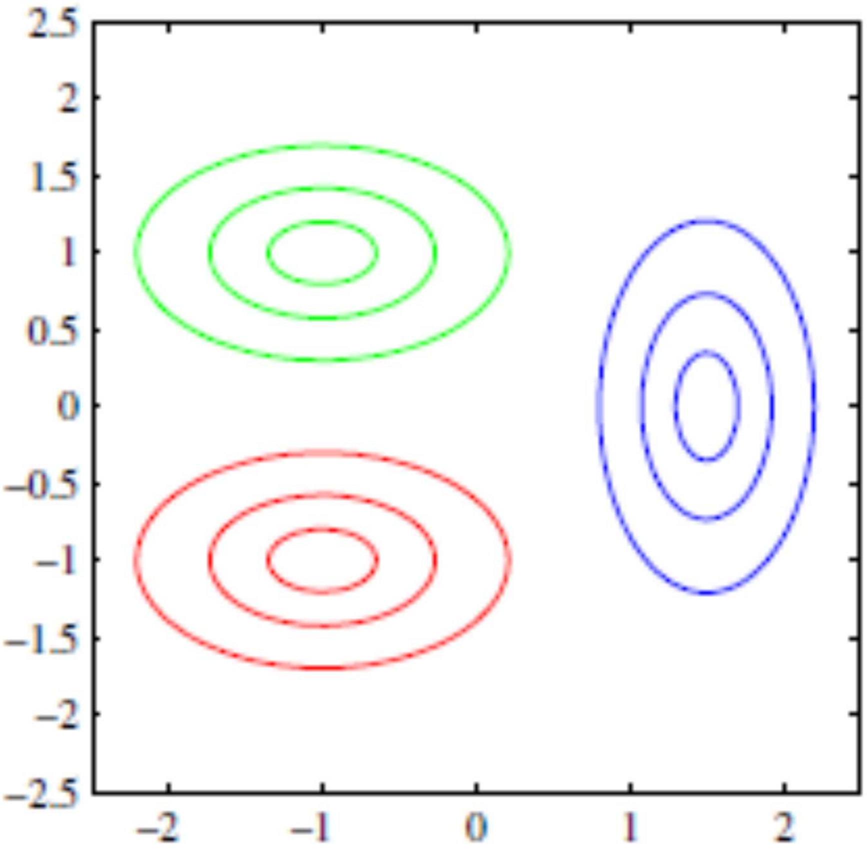
$$\mathbf{w}_k = \Sigma^{-1}\mu_k$$

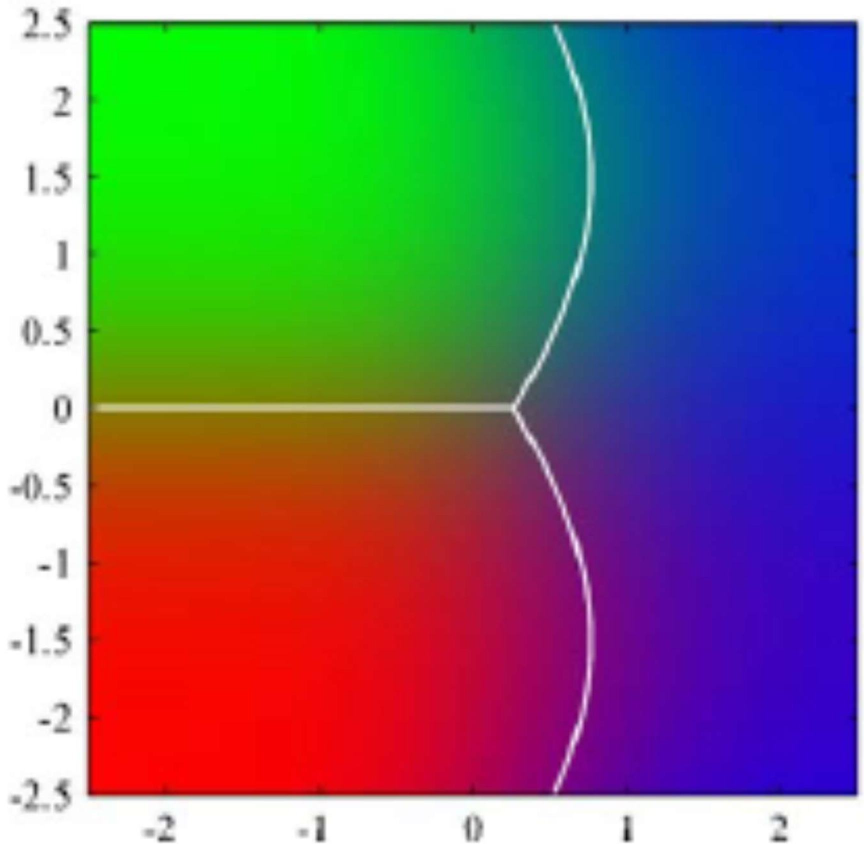
$$w_{k0} = -\frac{1}{2}\mu_k^{\mathrm{T}}\Sigma^{-1}\mu_k + \ln p(c_k)$$

- 如果放松共享协方差矩阵的假设，那么将获得 x 的二次函数，上升到二次判别式。

连续的输入

❖ 三类问题示意图





最大似然解

- 两类问题
  - 每类满足共享协方差矩阵的正态类条件分布
  - 数据集 $\{\mathbf{x}_n, t_n\}$，$t_n = 1$ 表示类别 c_1，$t_n = 0$ 表示类别 c_2
  - 类别先验概率 $p(c_1) = \pi, p(c_2) = 1 - \pi$
  - 来自类别 c_1 的数据点 $\mathbf{x}_n$，有

$$p(\mathbf{x}_n, c_1) = p(c_1) p(\mathbf{x}_n | c_1) = \pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)$$

  - 相似地，来自类别 c_2 的数据点 $\mathbf{x}_n$，有

$$p(\mathbf{x}_n, c_2) = p(c_2) p(\mathbf{x}_n | c_2) = \pi \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)$$

  - 似然函数

$$p(\mathbf{t} | \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[\pi \mathcal{N}\left(\mathbf{x}_n | \mu_1, \Sigma\right) \right]^{t_n} \left[(1 - \pi) \mathcal{N}\left(\mathbf{x}_n | \mu_2, \Sigma\right) \right]^{1 - t_n}$$

其中 $\mathbf{t} = \left(t_1, \ldots, t_N\right)^{\mathrm{T}}$

最大似然解

- 在对数似然函数中，与 π 有关的项为

$$\sum_{n=1}^{N}\left\{t_n \ln \pi + \left(1-t_n\right)\ln\left(1-\pi\right)\right\}$$

令对 π 的导数为零，有

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

对于多类情形，可以得到相似结果。

- 在对数似然函数中，与 μ_1 有关的项为

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}\left(\mathbf{x}_n \middle| \mu_1, \Sigma\right) = -\frac{1}{2}\sum_{n=1}^{N} t_n \left(\mathbf{x}_n - \mu_1\right)^{\mathrm{T}} \Sigma^{-1}\left(\mathbf{x}_n - \mu_1\right) + \text{const}$$

令对 μ_1 的导数为零，有

$$\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n$$

最大似然解

相似地，对 μ_2，有

$$\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathbf{x}_n$$

- 在对数似然函数中，与共享协方差矩阵 Σ 有关的项，有

$$-\frac{1}{2}\sum_{n=1}^{N}t_n\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}t_n(\mathbf{x}_n-\mu_1)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n-\mu_1)$$

$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n-\mu_2)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n-\mu_2)$$

$$= -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\mathrm{tr}\{\Sigma^{-1}\mathbf{S}\}$$

最大似然解

其中

$$\mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in c_1}(\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^{\mathrm{T}}$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in c_2}(\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^{\mathrm{T}}$$

$\Sigma = \mathbf{S}$ 表示两个类别各自协方差矩阵的加权平均

- 结果容易扩展到 K 类问题
- 注意：类别数据拟合正态分布对离群点缺少鲁棒性，因为正态分布的最大似然估计缺少鲁棒性

离散特征

- 考虑二值特征矢量 $x_i \in \{0,1\}$
- 如果 D 个输入，则一般分布对应着每类 2^D 个数，包含 $2^D - 1$ 个独立变量
- 朴素 Bayes 假设：对类别 c_k，特征值是条件独立的，有

$$p(\mathbf{x}|c_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{1-x_i}$$

因为

$$\alpha_k = \ln p(\mathbf{x}|c_k) p(c_k)$$

有

$$\alpha_k(\mathbf{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki})\} + \ln p(c_k)$$

仍然是输入变量的线性函数

离散特征

- 类后验概率
  - K = 2：由带有 Logistic sigmoid 的泛化线性模型给出
  - K ≥ 2：由带有 Softmax 的泛化线性模型给出

诚信　创新　实践