数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

上节回顾

- 特征
  - 图像的形状特征
  - 图像的纹理特征
- 线性分类器
  - 线性分类器基本概念

本节提要

- 线性分类器
  - 垂直平分分类器
  - Fisher投影准则
  - 感知准则

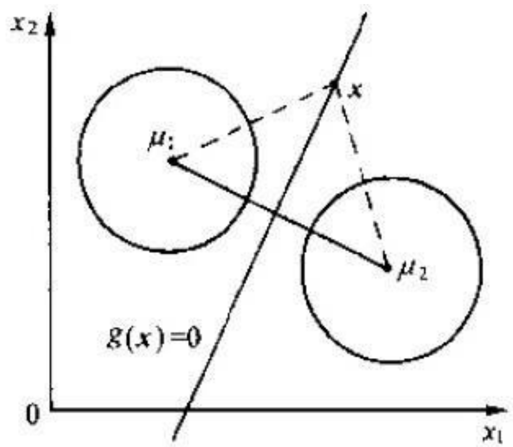
2　垂直平分分类器

- 2.1　问题与思路
- 2.2　垂直平分形式
- 2.3　最小距离形式
- 2.4　实例
- 2.5　特点

2.1 问题与思路

- **垂直平分分类器又称为最小距离分类器。**

- **设计思路**
  - 基于两类样本均值点作垂直平分线



2.1 问题与思路

- **已知**
  - 给定类别已知的训练样本集Z有N个样本，
    - 其中ω_1类样本有N_1个，样本集用Z_1表示；
    - ω_2类样本有N_2个，样本集用Z_2表示；
  - 显然
    - $N_1 + N_2 = N$
    - $Z_1 + Z_2 = Z$
- **试求垂直平分分类器**

2.2 垂直平分形式

- **判别函数与决策面方程**
  - 对于两类二维问题
  - C = 2，D =2
  - 垂直平分线性判别函数
    - $g(x) = w^T x + w_0$
  - 垂直平分直线方程
    - $g(x) = 0$ 即 $w^T x + w_0 = 0$

2.2 垂直平分形式

- **求解权向量与阈值权**
  - 先求均值向量
    - m_1 和 m_2
  - 利用垂直几何关系，设权向量
    - $w = (m_1 - m_2)$
  - 则直线方程为
    - $(m_1 - m_2)^T x + w_0 = 0$

（注意正侧在m_1这边）

2.2 垂直平分形式

- **求解权向量与阈值权**
  - 再利用平分几何关系，中点x_0在直线上
    - $x_0 = (m_1 + m_2) / 2$
  - 代入方程求得
    - $w_0 = -(m_1 - m_2)^T (m_1 + m_2) / 2$

2.2 垂直平分形式

- **最终结果**
  - 线性判别函数
    - $g(x) = (m_1 - m_2)^T x - (m_1 - m_2)^T (m_1 + m_2) / 2$
    - $\quad = (m_1 - m_2)^T (x - (m_1 + m_2) / 2)$
  - **决策面方程**
    - $(m_1 - m_2)^T (x - (m_1 + m_2) / 2) = 0$

2.2 垂直平分形式

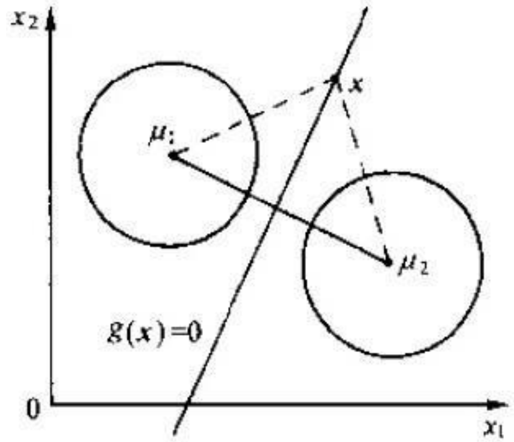
- **决策规则**
  - 已知垂直平分判别函数
    - $g(x) = (m_1 - m_2)^T (x - (m_1 + m_2) / 2)$
  - 垂直平分决策规则为
    - 对于未知样本x，若$g(x) > 0$，则x决策为ω_1类
    - 若$g(x) < 0$，则x决策为ω_2类

2.2 垂直平分形式

- **判别函数与决策面方程**
  - 很容易推广到两类多维问题
  - C = 2, D任意
  - 垂直平分线性判别函数
    - $g(x) = w^T x + w_0$
  - 垂直平分决策面方程
    - $g(x) = 0$ 即 $w^T x + w_0 = 0$

2.3 垂直平分分类器的最小距离形式

- **最小距离等价形式的由来**
  - 定义欧式距离（非线性）为判别函数
    - $G_1(x) = d_1(x) = \|x - m_1\|$
    - $G_1(x) = d_2(x) = \|x - m_2\|$



2.3　最小距离形式

- **决策规则**
  - 等价的最小距离决策规则为
    - 对于未知样本x，若$d_1(x) < d_2(x)$，则x决策为ω_1类
    - 若$d_1(x) > d_2(x)$，则x决策为ω_2类

2.4 实例

- **已知**
  - 甲类：$[0\ \ 3]^T$、$[2\ \ 4]^T$、$[1\ \ 3]^T$、$[2\ \ 3]^T$、$[0\ \ 2]^T$
  - 乙类：$[4\ \ 1]^T$、$[3\ \ 2]^T$、$[2\ \ 1]^T$、$[3\ \ 0]^T$、$[3\ \ 1]^T$
- **试问**
  - 待分类样本为$x = [5\ \ 0]^T$，问x应决策为哪一类？

2.5 特点

- **最小距离分类器的主要特点**
  - 解决两类分类问题的线性分类器
  - 原则上对样本集无特殊要求
  - 未采用准则函数求极值解（非最佳决策）
  - 算法最简单，分类器设计最容易

3　Fisher投影准则

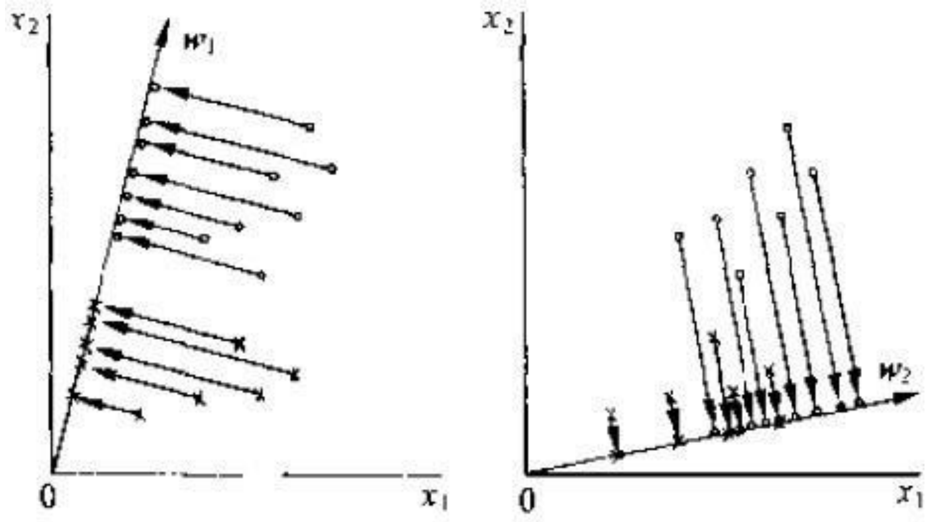
- 3.1　问题与思路
- 3.2　Fisher准则函数
- 3.3　准则函数化简
- 3.4　求极值解
- 3.5　特点
- 3.6　后续研究

3.1 问题和思路

- **原因**
  - 高维问题——特征个数太多
  - （经典理论）分类器设计困难
  - 分类困难

3.1 问题和思路

- **设计思路**
  - 通过投影对高维分类问题降维
  - Fisher将高维特征空间的样本投影到一维直线上



3.1 问题和思路

- **问题**
  - 已知C = 2，D维分类问题的样本集
  - 设投影向量为p
  - 则一维投影方程为$y = p^T x$
  - 求最佳投影向量p（的方向）

3.2 Fisher准则函数

- **Fisher定义的准则函数**
  - 定义各类均值m_1和m_2
  - 定义各类离散度S_1和S_2
  - 定义总离散度$S_W = S_1 + S_2$
  - 定义类间离散度S_B

1. 在 d 维 X 空间

(1) 各类样本均值向量 $\boldsymbol{m}_i$

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{x \in \mathscr{X}_i} \boldsymbol{x}$$

(2) 样本类内离散度矩阵 S_i 和总类内离散度矩阵 S_w

$$S_i = \sum_{x \in \mathscr{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

$$S_w = S_1 + S_2$$

(3) 样本类间离散度矩阵 S_b[①]

$$S_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

2. 在一维 Y 空间

(1) 各类样本均值 $\tilde{m}_i$

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y \in \mathscr{Y}_i} y$$

(2) 样本类内离散度 $\tilde{S}_i^2$ 和总类内离散度 $\tilde{S}_w$

$$\tilde{S}_i^2 = \sum_{y \in \mathscr{Y}_i} (y - \tilde{m}_i)^2$$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2$$

(3) 样本的类间离散度：

$$(\tilde{m}_1 - \tilde{m}_2)^2$$

3.2　Fisher准则函数

- **Fisher定义的准则函数**
  - 定义Fisher投影准则

$$J_F(p)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2}$$

- **Fisher投影准则的物理含义**
  - 投影后异类样本尽量远离
  - 投影后同类样本尽量靠近

3.3 准则函数化简

- **化简Fisher准则函数**
  - 分子的化简

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in \mathscr{Y}_i} y = \frac{1}{N_i}\sum_{x\in \mathscr{X}_i} w^T x$$

$$= w^T\left(\frac{1}{N_i}\sum_{x\in \mathscr{X}_i} x\right) = w^T m_i$$

分子便成为

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (w^T m_1 - w^T m_2)^2$$

$$= w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_b w$$

3.3 准则函数化简

- **化简Fisher准则函数**
  - 分母的化简

$$\tilde{S}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 = \sum_{x \in \mathcal{X}_i} (\boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{m}_i)^2$$

$$= \boldsymbol{w}^T \left[\sum_{x \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T \right] \boldsymbol{w} = \boldsymbol{w}^T S_i \boldsymbol{w}$$

$$\tilde{S}_1^2 + \tilde{S}_2^2 = \boldsymbol{w}^T (S_1 + S_2) \boldsymbol{w} = \boldsymbol{w}^T S_w \boldsymbol{w}$$

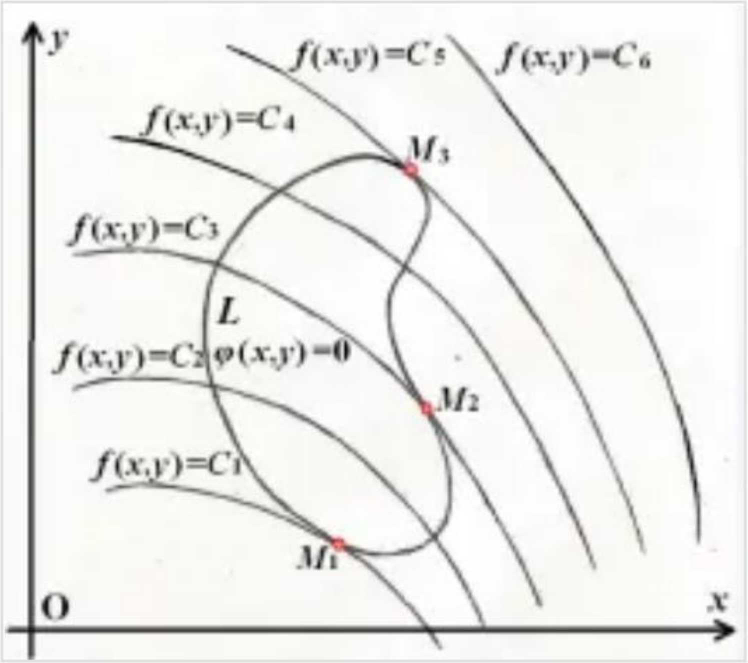
3.3 准则函数化简

- **Fisher准则函数**
  - 化简的结果

$$J_F(p) = \frac{p^T S_b p}{p^T S_W p}$$

3.4 求极值解

曲线 L 为约束条件 $\varphi(x,y)=0$ ，$f(x,y)=C$ 为目标函数的等值线



- 求Fisher函数的极值解
  - 采用Lagrange乘子法求极值
    - 等式约束条件：令分母为常数
    - 目标函数：分子

$$w^T S_w w = c \neq 0$$

定义 Lagrange 函数为

$$L(w,\lambda) = w^T S_B w - \lambda(w^T S_w w - c)$$

式中 λ 为 Lagrange 乘子。将式(4-28)对 w 求偏导数，得

$$\frac{\partial L(w,\lambda)}{\partial w} = S_b w - \lambda S_w w$$

令偏导数为零，得

$$S_b w^* - \lambda S_w w^* = 0$$

即

$$S_b w^* = \lambda S_w w^*$$

拉格朗日函数

$$F(x,y,\lambda) = f(x,y) + \lambda\varphi(x,y)$$

3.4 求极值解

- **求Fisher函数的极值解**

其中 w^* 就是 $J_F(w)$ 的极值解。因为 S_w 非奇异，式(4-29)两边左乘 S_w^{-1}，可得

$$S_w^{-1}S_b w^* = \lambda w^* \tag{4-30}$$

解式(4-30)为求一般矩阵 $S_w^{-1}S_b$ 的本征值问题。但在我们这个特殊情况下，利用式(4-19)S_b 的定义，式(4-30)左边的 $S_b w^*$ 可以写成

$$S_b w^* = (m_1 - m_2)(m_1 - m_2)^T w^* = (m_1 - m_2)R$$

式中 $$R = (m_1 - m_2)^T w^*$$

为一标量，所以 $S_b w^*$ 总是在向量 $(m_1 - m_2)$ 的方向上。由于我们的目的是寻找最好的投影方向，w^* 的比例因子对此并无影响，因此，从式(4-30)可得

$$\lambda w^* = S_w^{-1}(S_b w^*) = S_w^{-1}(m_1 - m_2)R$$

从而可得

$$w^* = \frac{R}{\lambda} S_w^{-1}(m_1 - m_2) \tag{4-31}$$

忽略比例因子 R/λ，得

$$w^* = S_w^{-1}(m_1 - m_2) \tag{4-32}$$

3.4 求极值解

- **求Fisher函数的极值解**
  - 极值解（极大值）

$$p^* = S_W^{-1}(m_1 - m_2)$$

3.5 特点

- **Fisher投影的特点**
  - 解决两类问题的线性投影
  - 原则上对样本集无特殊要求（S_w矩阵可逆）
  - 采用Fisher投影准则函数求极值解（最佳决策）
  - 分类器设计较容易

3.6 后续研究

- 1936 年，Fisher发表经典论文，提出投影准则。Wilks和Duda分别提出判别向量集概念，由判别向量集构成子空间，对原始样本在子空间中的投影向量进行分类判别。
- 1970年，Sammon提出基于Fisher准则的最佳判别平面。Foley和Sammon提出采用正交条件下的最佳判别向量集进行特征提取的方法。
- 1988年，Duchene等给出多类情况最佳判别向量集的计算公式。
- ……
- Linear Discriminent Analysis (LDA)

实验2：垂直平分分类器

- 给定训练数据，学习得到一个垂直平分分类器
- 对测试样本进行分类

- Python编程实现

4　感知准则

- 4.1　样本集线性可分
- 4.2　解向量和解区
- 4.3　感知准则函数
- 4.4　求极值解
- 4.5　特点
- 4.6　后续研究

4.1 样本集线性可分

- **样本集的线性可分性**
  - [线性可分] 若训练样本集可以被某个线性分类器完全正确分类，则该样本集是线性可分的。
  - 样本集是线性可分的——至少存在一个权向量，能将该样本集中的每个样本都正确分类；
  - 否则就是线性不可分的（异或问题）。

4.1 样本集线性可分

- **问题**
  - 已知C = 2，D维分类问题的样本集（其它略）
  - 设该样本集是线性可分的
  - 提出感知准则（因此称为感知器）
  - 求能够对样本集正确分类的解（某个线性分类器）

  - 感知器用来解决线性可分样本集分类问题

4.1 样本集线性可分

- **线性可分性样本集的规范化**
  - 感知准则采用增广向量形式

    判别函数 $g(x) = a^T y$

    对于未知样本x，若 $g(x) > 0$，则x决策为 ω_1 类

    若 $g(x) < 0$，则x决策为 ω_2 类

  - 规范化
    - 对 ω_2 类样本的增广向量全部乘以-1
  - 规范化之后的分类结果
    - $a^T y_i > 0$——正确分类
    - $a^T y_i < 0$——错误分类

4.2 解向量和解区

- **概念**
  - 解向量——能将线性可分样本集中的每个样本都正确分类的权向量。
  - 解区——解向量往往不是一个，而是由无穷多个解向量组成的（角度）区域，称为解区。

4.3 感知准则函数

- **Rosenblatt定义感知准则函数**
  - 对于规范化的增广样本集
    - $a^T y_i < 0$——错误分类
  - 定义感知准则函数，作为优化准则函数

$$\min J_p(a) = \sum_{y \in Z_E} (-a^T y)$$

  - 求解向量（或解区）

4.3 感知准则函数

- 图示法求解区
- 解区可以直接画图求出（二维条件时）

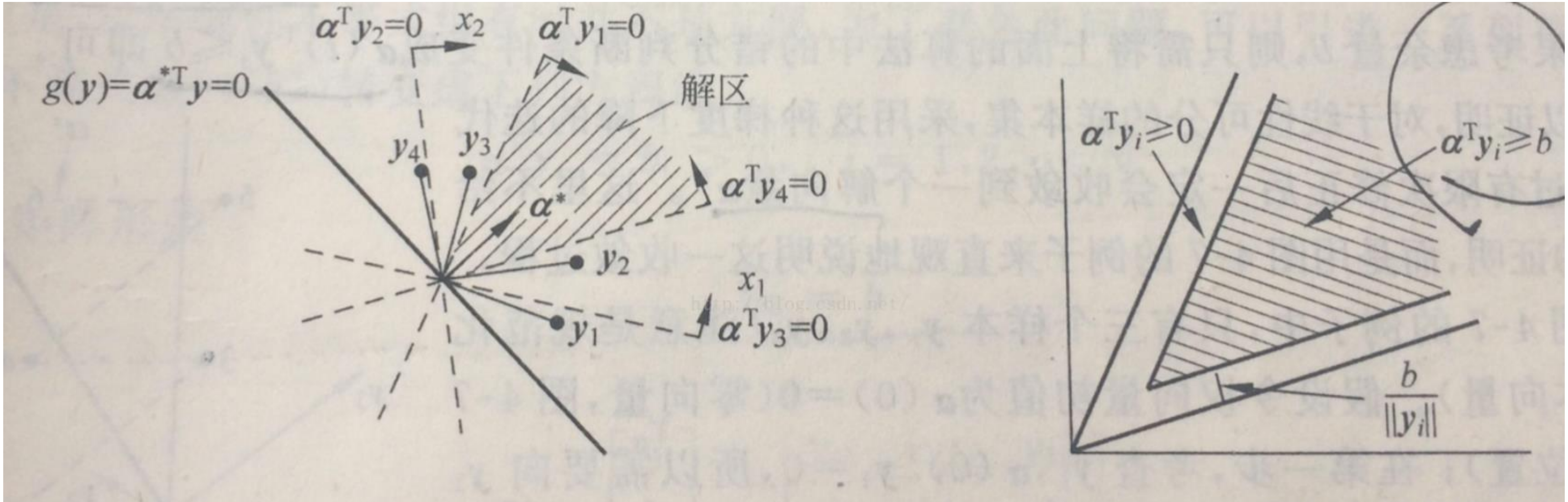


图 4-5 解向量和解区

图 4-6 带有余量的解区

4.4　求极值解

- **求解感知器**
  - 采用梯度下降法求优化准则函数极值（极小值）
    - 先求梯度方向
    - 计算参数改变量
    - 得到迭代公式

$$\nabla J_P(\boldsymbol{a}) = \frac{\partial J_P(\boldsymbol{a})}{\partial \boldsymbol{a}} = \sum_{y\in\mathcal{Y}^k}(-\boldsymbol{y})$$

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \rho_k \nabla J$$

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) + \rho_k \sum_{y\in\mathcal{Y}^k} \boldsymbol{y}$$

4.4 求极值解

- **求解感知器**
  - 梯度下降法求极值的问题
    - 收敛性
    - 步长的选择

4.5 特点

- **感知准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集必须是线性可分的
  - 采用感知准则函数求极值解（最优决策）
  - 分类器设计过程复杂

5　最小错分样本数准则

- 5.1　问题与思路
- 5.2　最小错分样本数准则一
- 5.3　最小错分样本数准则二
- 5.4　特点

5.1 问题与思路

- **问题的提出**
  - 感知准则只适用线性可分样本集——无错分
  - 实际情况未必线性可分——有错分
  - 另外线性可分的判断也很困难

  - 既然存在错分样本——求错分样本数最少

5.1 问题与思路

- **数学描述**
  - 仿照线性可分样本集的规范化（ω_2类样本的增广向量乘以-1）
    - $a^T y_i > 0$——正确分类
    - $a^T y_i < 0$——错误分类
  - 设样本数为N，N个不等式联立
    - $a^T y_i > 0 \qquad (i = 1,\ldots,N)$
  - 求满足不等式最多的解（权向量）

5.1 问题与思路

- **数学描述**
  - 写成矩阵形式

用矩阵形式重写式(4-44)所表示的不等式组，

$$Ya > 0$$

$$Y = \begin{bmatrix} \boldsymbol{y}_1^T \\ \boldsymbol{y}_2^T \\ \vdots \\ \boldsymbol{y}_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix}$$

为使解更可靠，引入余量 $\boldsymbol{b}>0$

$$Ya \geqslant \boldsymbol{b} > 0$$

5.2　最小错分样本数准则一

- **最小错分样本数准则一**
  - 准则函数

$$\min J_{q1}(\boldsymbol{a}) = \| (Y\boldsymbol{a} - \boldsymbol{b}) - |Y\boldsymbol{a} - \boldsymbol{b}| \|^2$$

  - 求极值解
    - 共轭梯度下降法

共轭梯度下降法

在数值线性代数中，共轭梯度法是一种求解对称正定线性方程组Ax=b的迭代方法。

事实上，求解Ax=b等价于求解：$min||Ax-b||_2^2$,将其展开后可以得到：$min\quad x^TA^TAx-b^TAx+b^Tb$,也就是等价于求解 $min\quad \frac{1}{2}x^TA^TAx-b^TAx$ 。于是解方程问题就转化为了求解二次规划问题(QP)。

共轭梯度法是介于梯度下降法与牛顿法之间的一个方法，是一个**一阶方法**。它克服了梯度下降法收敛慢的缺点，又避免了存储和计算牛顿法所需要的二阶导数信息。

在n维的优化问题中，共轭梯度法最多n次迭代就能找到最优解（是找到，不是接近），但是只针对二次规划问题。

共轭梯度法的思想就是找到n个两两共轭的共轭方向，每次沿着一个方向优化得到该方向上的极小值，后面再沿其它方向求极小值的时候，不会影响前面已经得到的沿哪些方向上的极小值，所以理论上对n个方向都求出极小值就得到了n维问题的极小值。

5.3　最小错分样本数准则二

- **最小错分样本数准则二**
  - 准则函数

$$\max J_{q2}(a) = \sum_{i=1}^{N} \frac{1+\mathrm{sgn}(y_i a)}{2}$$

$$\mathrm{sgn}(y_i a) = \begin{cases} +1, \text{对于 } y_i a \geq 0 \text{①} \\ -1, \text{对于 } y_i a < 0 \end{cases}$$

  - 求极值解
    - 搜索算法

5.4　特点

- **最小错分样本数准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集不限，可以是线性不可分的
  - 求满足不等式个数最多的权向量（最优）
  - 分类器设计过程复杂

6　最小平方误差准则

- 6.1　问题与思路
- 6.2　最小平方误差准则
- 6.3　余量的选择
- 6.4　特点

6.1 问题与思路

- **问题的提出**
  - 对于线性不可分问题
  - 最小错分样本数准则——求错分样本数最少
  - 工程上往往是求误差平方和最小

6.1 问题与思路

- **数学描述**
  - 引入余量b_i，将不等式组改造为等式组
    - $a^T y_i = b_i > 0 \quad (i = 1,\ldots,N)$
  - 求满足等式组的最小平方误差解（权向量）

6.1 问题与思路

- **数学描述**
  - 写成矩阵形式

$$Ya = b$$

$$Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{Nd} \end{bmatrix}$$

$$b = [b_1, b_2, \cdots, b_N]^T$$

6.2 最小平方误差准则

- **最小平方误差准则——工程上常用准则**
  - 定义优化准则函数

$$e = Ya - b$$

$$J_s(a) = \| e \|^2 = \| Ya - b \|^2 = \sum_{i=1}^{N} (a^t y_i - b_i)^2$$

6.2 最小平方误差准则

- **最小平方误差准则优化结果**
  - 直接求极值解

首先对式(4-63)中的 $J_s(\boldsymbol{a})$ 求梯度，

$$\nabla J_s(\boldsymbol{a}) = \sum_{n=1}^{N} 2(\boldsymbol{a}^T \boldsymbol{y}_n - b_n)\boldsymbol{y}_n = 2Y^T(Y\boldsymbol{a} - \boldsymbol{b})$$

令 $\nabla J_s(\boldsymbol{a})=0$，得

$$Y^TY\boldsymbol{a}^* = Y^T\boldsymbol{b} \tag{4-65}$$

这样，求解 $Y\boldsymbol{a}=\boldsymbol{b}$ 的问题转化为求解 $Y^TY\boldsymbol{a}^*=Y^T\boldsymbol{b}$ 的问题了。这一方程的最大优点是，矩阵 Y^TY 是 $\hat{d}\times\hat{d}$ 方阵，而且一般是非奇异的，因此可唯一地解得

$$\boldsymbol{a}^* = (Y^TY)^{-1}Y^T\boldsymbol{b} = Y^+\boldsymbol{b} \tag{4-66}$$

式中 $(\hat{d}\times N)$ 矩阵

$$Y^+ = (Y^TY)^{-1}Y^T \tag{4-67}$$

是 Y 的左逆矩阵，$\boldsymbol{a}^*$ 就是式(4-62)的 MSE 解。

6.3 余量的选择

- **选择不同余量的结果**
  - $b_i > 0$ $(i = 1,\ldots,N)$

    – 选项一情况

$$\boldsymbol{b} = \begin{bmatrix} N/N_1 \\ \vdots \\ N/N_1 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} N_1\text{个} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} N_2\text{个} \end{matrix}$$

    – 等价于Fisher解

6.3　余量的选择

- **选择不同余量的结果**
  - $b_i > 0\ (i = 1,\ldots,N)$

– 选项二情况：全部为1，即$b_i = 1$

– 当N趋于无穷时，逼近Bayes解（最优分类器）

6.4 特点

- **最小平方误差准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集不限，可以是线性不可分的
  - 求最小平方误差的权向量（最优）
  - 分类器设计过程相对简单

Bayes分类器

- 4.1 基本概念
- 4.2 最小错误率Bayes决策
- 4.3 最小风险Bayes决策
- 4.4 最小最大Bayes决策
- 4.5 Bayes分类器设计

4.1 基本概念

- [错误率] 几乎所有的分类器在识别时都有可能出现错误分类（简称错分 / 误判）的情况，这种错误分类的可能性称为分类器识别结果的错误概率，简称错误率 / 误判率。

- [正确率] （通常意义的）正确率 = 1 – 错误率

4.1 基本概念

- **线性分类器**
  - 垂直平分分类器
    - 未经优化，错误率通常较大
  - 感知器
    - 优化（求线性可分样本集的解），最终错误率未知
  - 最小平方误差
    - 优化（样本集MSE的解），最终错误率未知

4.1 基本概念

- **Bayes分类器设计思路**
  - 寻求概率意义上的最小错误率的分类器
  - 即具有最小错分概率的分类器——分类器设计的最优解

4.1 基本概念

- **数学基础回顾**
  - 概率论与数理统计
    - 随机事件
    - 概率
    - 条件概率
    - Bayes公式
    - 随机变量
    - 概率密度函数

4.1 基本概念

- **数学基础回顾**
  - Bayes分类相关
    - 随机事件——样本的状态/ 类别
    - 概率——状态/ 类别的概率
    - 随机变量——随机向量
    - 概率密度函数

贝叶斯公式

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

$$P\left(B_i|A\right) = \frac{P\left(B_i\right)P\left(A|B_i\right)}{\sum_{j=1}^{n} P\left(B_j\right)P\left(A|B_j\right)}$$

- 先验和后验：**P(θ)和P(θ|x)**

机器学习两大流派——贝叶斯派和频率派

- **频率派旨在求最大似然估计**
  - 认为待求参数 θ 是唯一存在的
  - θ 可以是模型参数，也可以是分类标签或预测结果
  - 利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MLE}} &= \arg\max P(X;\theta) \\
&= \arg\max P(x_1;\theta)P(x_2;\theta)\cdots\cdot P(x_n;\theta) \\
&= \arg\max \log \prod_{i=1}^{n} P(x_i;\theta) \\
&= \arg\max \sum_{i=1}^{n} \log P(x_i;\theta) \\
&\boxed{= \arg\min -\sum_{i=1}^{n} \log P(x_i;\theta)}
\end{aligned}
$$

– 负对数似然函数

贝叶斯派和频率派

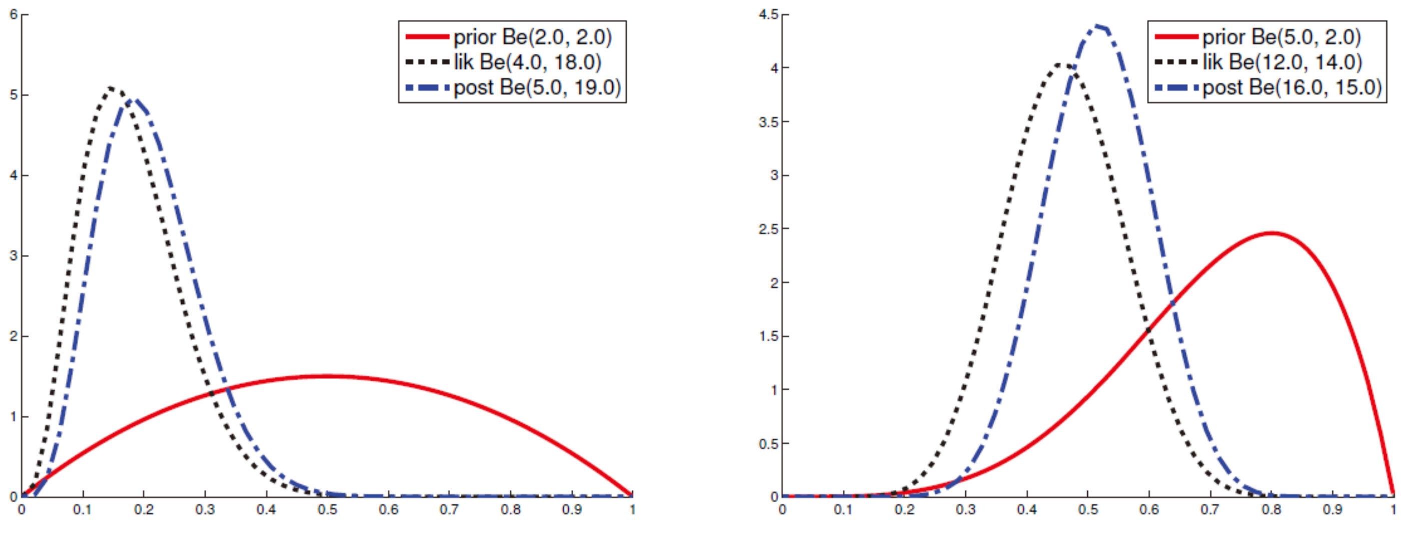
- 贝叶斯派旨在求最大后验估计
  - 认为待求参数 θ 是一个随机变量，符合一定的概率分布
  - 预设一个参数 θ 的概率分布，再用已有样本去修正这个预设（先验概率），得到最有利于样本出现的分布参数（后验概率）

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &= \arg\max P(\theta|X) \\
&= \arg\min -\log P(\theta|X) \\
&= \arg\min -\log P(X|\theta) - \log P(\theta) + \log P(X) \\
&= \arg\min \boxed{-\log P(X|\theta)} - \log P(\theta)
\end{aligned}
$$

*分号表示确定性，竖条表示不确定性

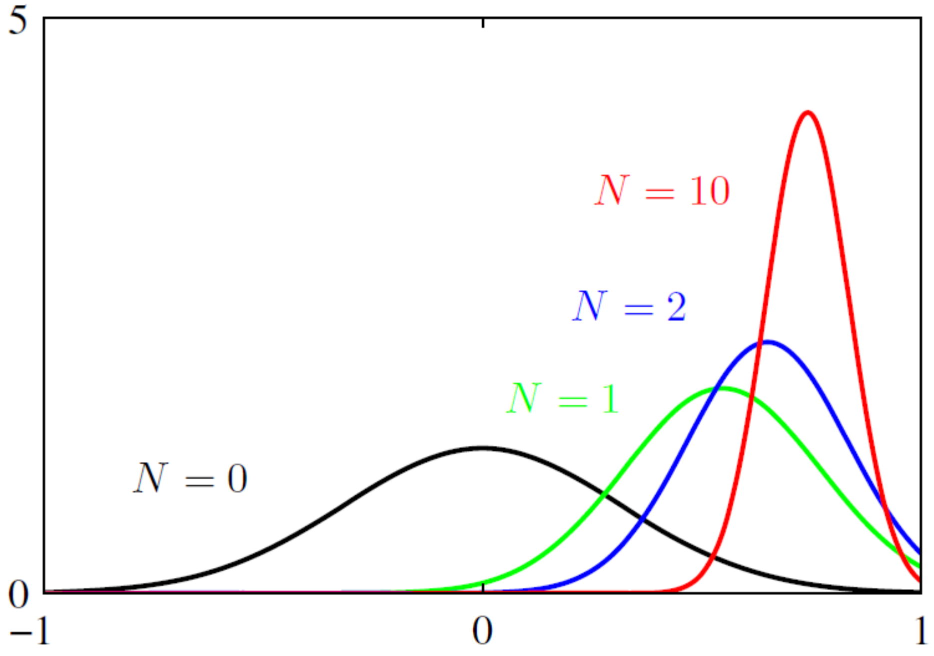
贝叶斯派和频率派

- **频率派的优势**
  - 样本足够大的情况下较容易得到接近无偏的估计
  - 样本少的情况下，偏差较大（例：投5次硬币）
- **贝叶斯派的优势**
  - 实际上是基于先验的校正，由于先验的存在，样本少时效果也不会太差
  - 先验非常重要



(a) (b)

- 不同先验的影响



- 训练样本的影响–稀释先验

- 频率派和贝叶斯派的等价关系

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

THE END !