数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

上节回顾

- 线性分类器
  - 垂直平分分类器
  - Fisher投影
- 特征
  - Hough变换特征
  - 邻接和距离

实验课安排

- 下周四（4.8）上午1、2节，二主楼B403

实验2：提取图像的纹理特征

- 给定4张图像，利用灰度共生矩阵特征或局部二值模式特征对这4张图像进行特征提取
- 图像是W*H*3的矩阵
- 将最终提取到的特征通过plot的形式展示，直观对比不同纹理提取到的特征

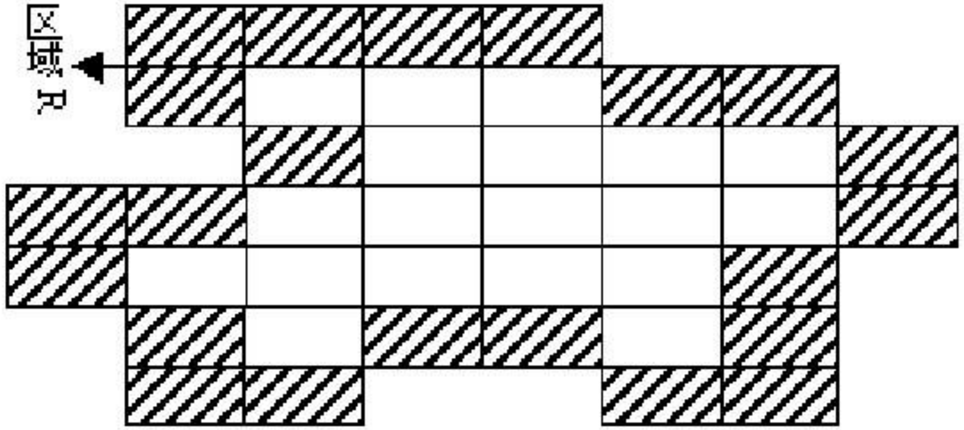
- Python编程实现

续上节：图像的形状特征

4.区域的测量

✓ 区域的大小及形状表示方法主要包括以下几种：

（1）面积S：图像中的区域面积S可以用同一标记的区域内像素的个数总和来表示。



区域的面积和周长

按上述表示法区域R的面积S=41。区域面积可以通过扫描图像，累加同一标记像素得到，或者是直接在加标记处理时计数得到。

（2）周长L：区域周长L是用区域中相邻边缘点间距离之和来表示。采用不同的距离公式，关于周长L的计算有很多方法。常用的有两种：

一种计算方法是采用欧式距离，在区域的边界像素中，设某像素与其水平或垂直方向上相邻边缘像素间的距离为1，与倾斜方向上相邻边缘像素间的距离为$\sqrt{2}$ 。周长就是这些像素间距离的总和。这种方法计算的周长与实际周长相符，因而计算精度比较高。

另一种计算方法是采用8邻域距离，将边界的像素个数总和作为周长。也就是说，只要累加边缘点数即可得到周长，比较方便，但是，它与实际周长间有差异。

（3）圆形度R_0：圆形度R_0用来描述景物形状接近圆形的程度，它是测量区域形状常用的量。其计算公式为：

$$R_0 = 4\pi S/L^2$$

式中为S区域面积；L为区域周长R_0值的范围为$0 < R_0 \le 1$，R_0值的大小反映了被测量边界的复杂程度，越复杂的形状取值越小。R_0值越大，则区域越接近圆形。

（4）形状复杂性e：形状复杂性常用离散指数表示，其计算公式为：

$$e = L^2/S$$

该式描述了区域单位面积的周长大小，e值越大，表明单位面积的周长大，即区域离散，则为复杂形状；反之，则为简单形状。e值最小的区域为圆形。

典型连续区域的计算结果为：圆形e=12.6；正方形e=16.0；正三角形e=20.8。

此外，常用的特征量还有区域的幅宽、占有率和直径等。

8.4 图像的纹理分析技术

8.4.1 纹理分析概念

- 指的是图像像素灰度级或颜色的某种变化，主要研究如何获得图像纹理特征和结构的定量描述和解释，以便于图像分析、分割和理解。
- 一般来说，可以认为纹理由许多相互接近、相互编织的元素构成，并常富有周期性。
- 纹理的定义大体可以从三个方面来描述：
  - 具有某种局部的序列性，并在该序列更大的区域内不断重复；
  - 序列由基本部分非随机排列组成；
  - 各个部分大致都是均匀的统一体。

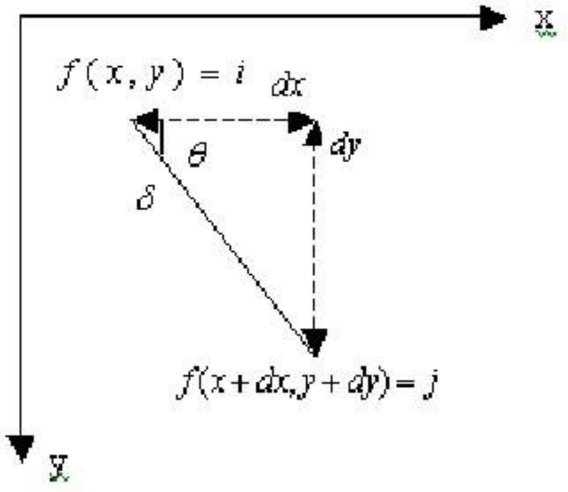
图像的纹理特征

- 纹理分析是指通过一定的图像处理技术抽取出纹理特征，从而获得纹理的定量或定性描述的处理过程。
- 纹理特征是从图像中计算出来的一个值，它对区域内部灰度级变化的特征进行量化。
- 纹理分析基本过程是从像素出发，在纹理图像中提取出一些辨识力比较强的特征，作为检测出的纹理基元，并找出纹理基元排列的信息，建立纹理基元模型，然后再利用此纹理基元模型对纹理图像进一步分割、分类或是辨识等处理。

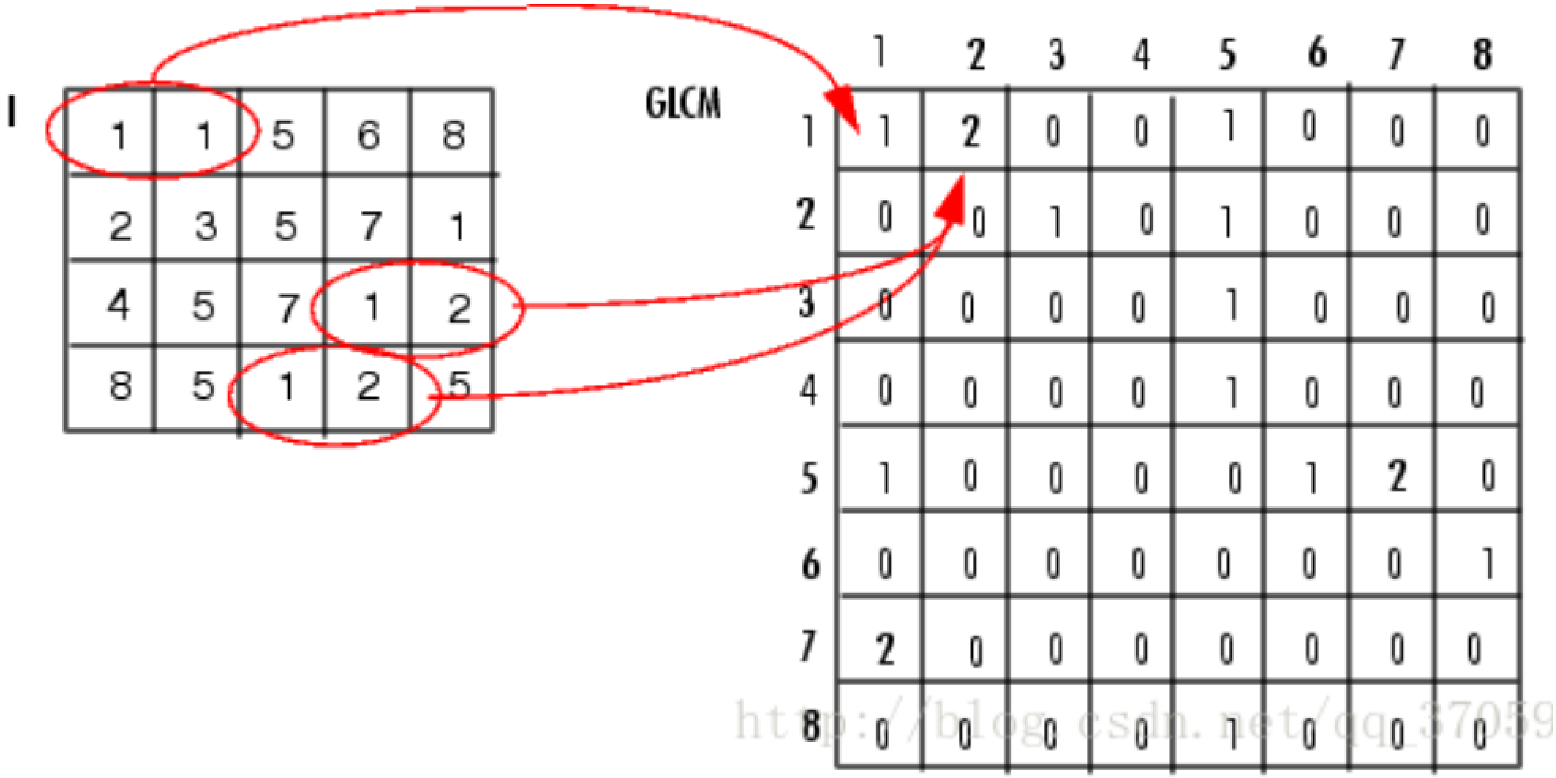
8.4.2 空间灰度共生矩阵

✓ 灰度共生矩阵就是从$N \times N$的图像f (x, y) 的灰度为i的像素出发，统计与距离为$\delta = (dx^2 + dy^2)^{1/2}$，灰度为 j的像素同时出现的概率$P(i, j, \delta, \theta)$。用数学表达式则为：

$$P(i, j, \delta, \theta) = \{[(x, y), (x + dx, y + dy)] \mid f(x, y) = i, f(x + dx, y + dy) = j\}$$



灰度共生矩阵的像素对



- 分方向
- 统计个数
- 构造矩阵
- 形成特征向量

1. 0°方向灰度共生矩阵

当 $\theta = 0°$， $dx = 1, dy = 0$ 由于所给图像中只有4个灰度级，因此所求得的灰度共生矩阵的大小为 4×4 。

可设置有向或无向

0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3

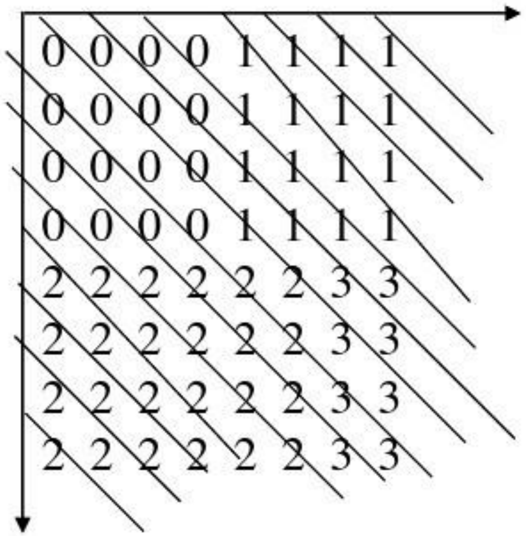
一幅数字灰度图像

0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3

0°方向灰度共生矩阵计算示意图

2. 45°方向灰度共生矩阵

当 $\theta = 45°$ 时，$dx = 1, dy = -1$。



45°方向灰度共生矩阵计算示意图

3. 90°方向灰度共生矩阵

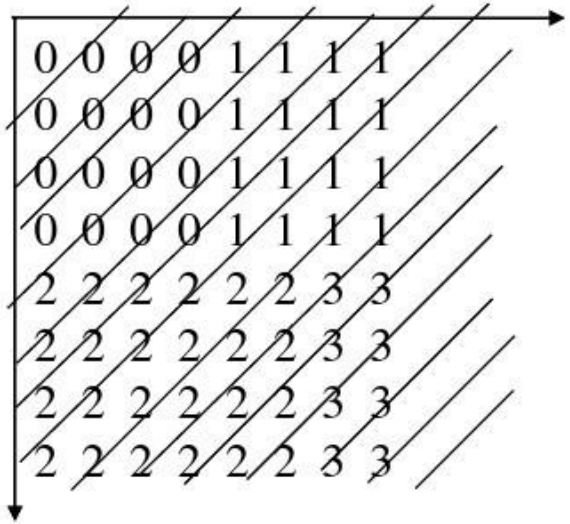
当 $\theta = 90^\circ$ 时，$dx = 0, dy = -1$。

0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3

90°方向灰度共生矩阵计算示意图

4. 135°方向灰度共生矩阵

当 $\theta = 135^{\circ}$ 时，$dx = -1, dy = -1$ 。



135°方向灰度共生矩阵计算示意图

灰度共生矩阵计算结果（双向）

$$P(0°)=\begin{bmatrix}12 & 4 & 0 & 0\\ 4 & 12 & 0 & 0\\ 0 & 0 & 20 & 4\\ 0 & 0 & 4 & 4\end{bmatrix}$$

(a)

$$P(45°)=\begin{bmatrix}9 & 3 & 4 & 0\\ 3 & 9 & 1 & 2\\ 4 & 1 & 15 & 3\\ 0 & 2 & 3 & 3\end{bmatrix}$$

(b)

$$P(90°)=\begin{bmatrix}12 & 0 & 4 & 0\\ 0 & 12 & 2 & 2\\ 4 & 2 & 18 & 0\\ 0 & 2 & 0 & 6\end{bmatrix}$$

(c)

$$P(135°)=\begin{bmatrix}9 & 3 & 3 & 0\\ 3 & 9 & 3 & 1\\ 3 & 3 & 15 & 3\\ 0 & 1 & 3 & 3\end{bmatrix}$$

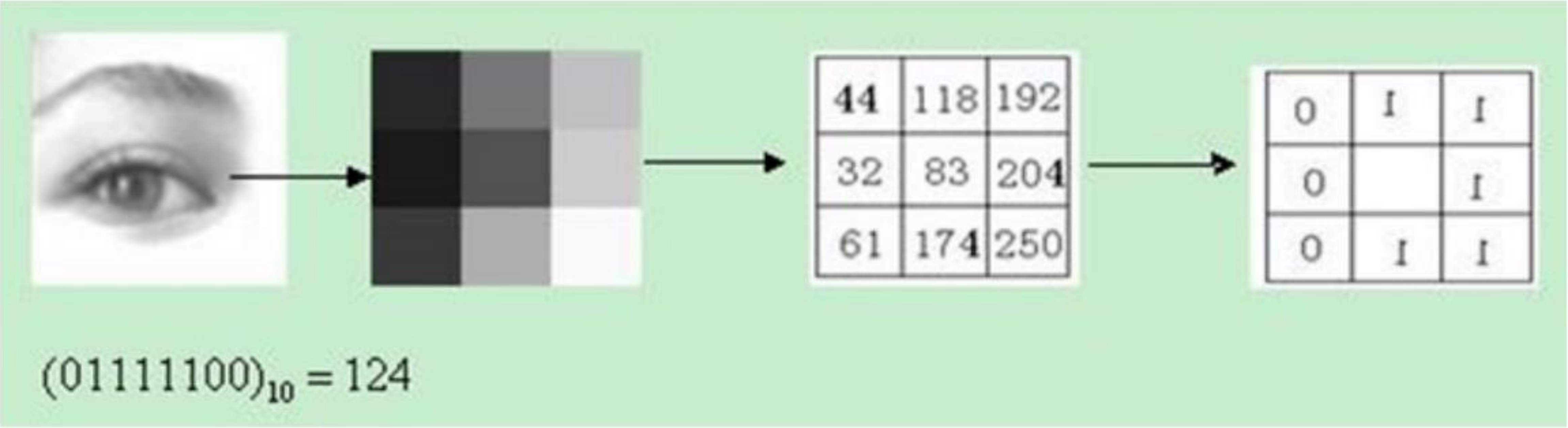
(d)

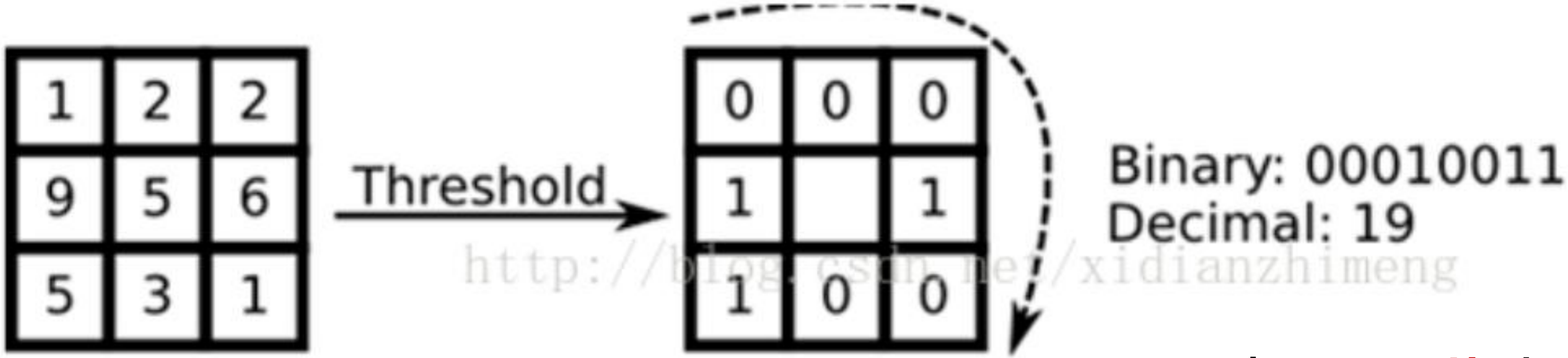
注意事项：

- 特征提取算法无需严格遵从某规则，只要规则统一即可
- 各方向可串联或相加，为了便于计算，本实验仅需取0°和90°
- 图像灰阶通常为256，构造矩阵后特征维数过高，通常需采取某些操作使维数降低
  - 将灰阶从256压缩至16
  - 对矩阵行求和或列求和作为特征
  - 一般特征维数在“百”量级

局部二值模式特征（Local Binary Pattern, LBP）

- LBP算子定义为在3*3的窗口内，以窗口中心像素为阈值，将相邻的8个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素点的位置被标记为1，否则为0。这样，3*3邻域内的8个点经比较可产生8位二进制数（通常转换为十进制数即LBP码，共256种），即得到该窗口中心像素点的LBP值，并用这个值来反映该区域的纹理信息（例如亮点和暗点）
- 最终得到LBP图，然后计算直方图


44 118 192
32 83 204
61 174 250
0 1 1
0 1
0 1 1
(01111100)10 = 124


1 2 2
9 5 6
5 3 1
Threshold
0 0 0
1 1
1 0 0
Binary: 00010011
Decimal: 19
http://blog.csdn.net/xidianzhimeng

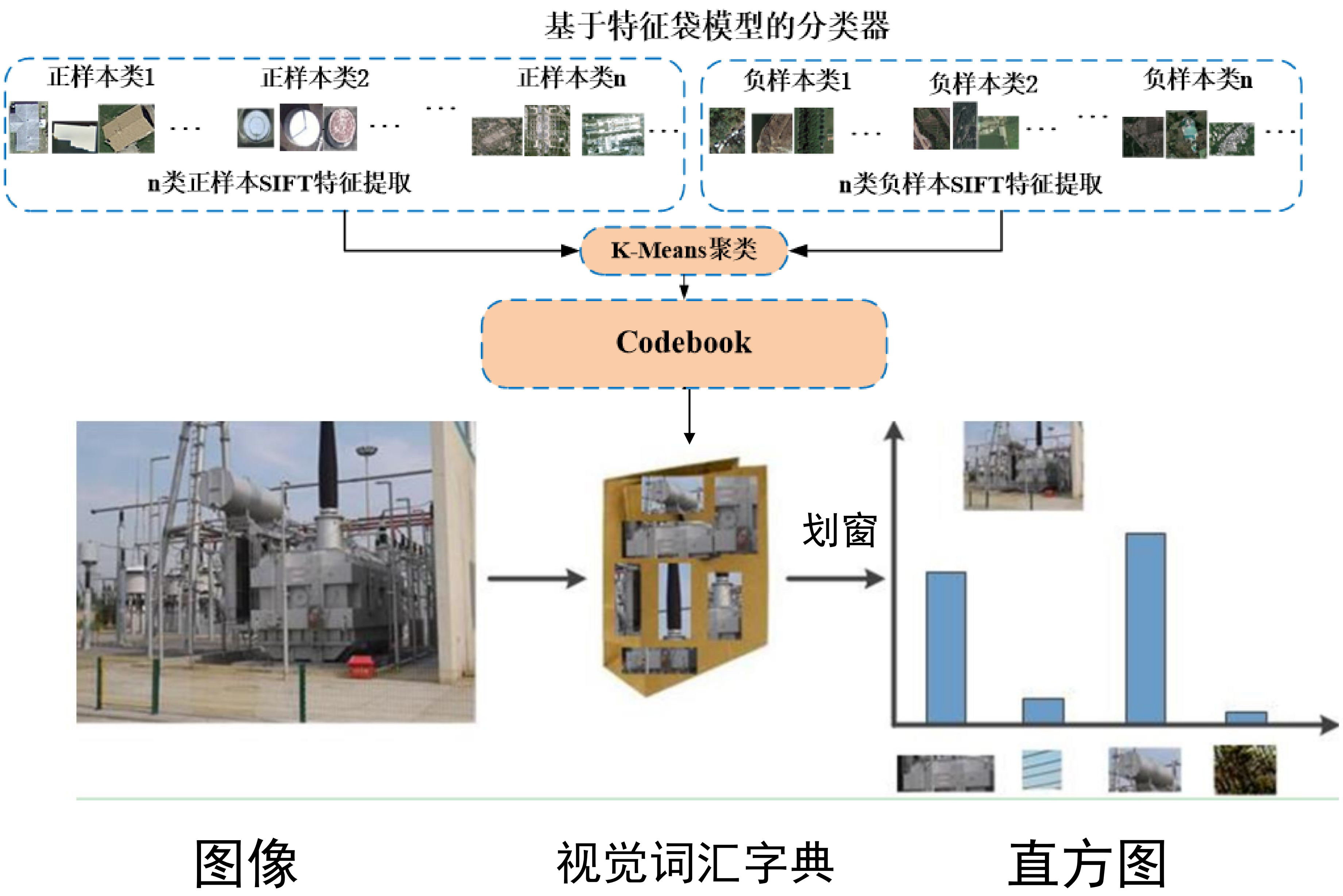
■思考：边缘如何处理？

视觉词袋（Bag-of-words，BOW）

词袋是一种统计关键词出现频率的文档表示方法，2004年被引入计算机视觉领域。词袋模型将图像局部特征空间进行量化，得到一组离散的具有代表性的特征向量，成为视觉单词，视觉单词的集合称为视觉词典，通常采用的视觉词典的构建方法是利用K-means算法对局部特征进行聚类，每个聚类中心对应于一个视觉词汇，视觉词袋方法将图像中的局部特征匹配到距离最近的一个视觉词汇，用视觉词汇出现频率的直方图表示图像。

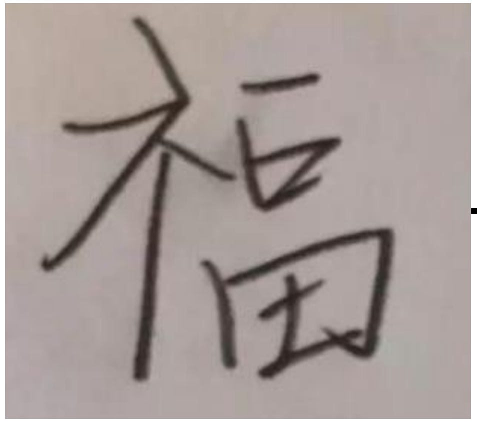
特征袋（Bag-of-Features，BoF）：

该模型将多幅图像作为文档集合，分别提取出文档集合中单个文档的局部特征，然后采用聚类算法生成码书(codebook)，每个聚类的中心即代表一个视觉词汇，码书则表示为视觉词汇的集合，通过量化提取的输入图像特征来计算属于码书中视觉词汇的统计直方图，即为图像的特征袋模型。


基于特征袋模型的分类器
正样本类1
正样本类2
正样本类n
n类正样本SIFT特征提取
负样本类1
负样本类2
负样本类n
n类负样本SIFT特征提取
K-Means聚类
Codebook
划窗
图像
视觉词汇字典
直方图

大作业1：福字识别

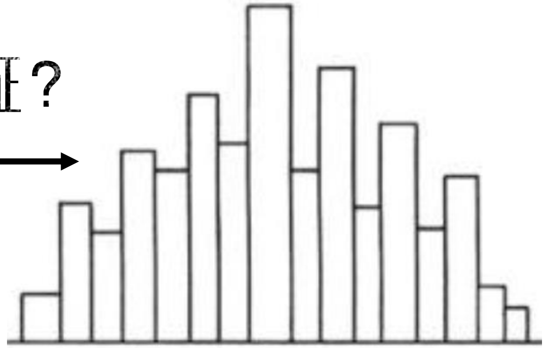
训练?



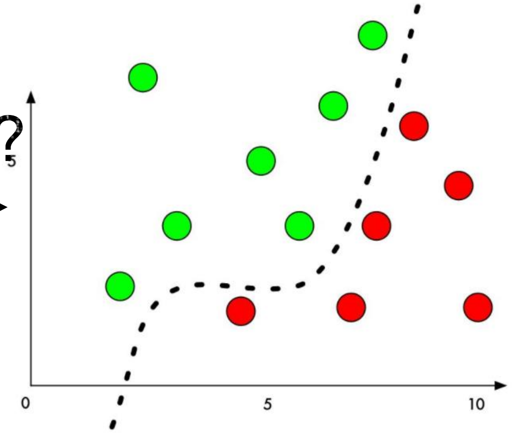
提供矩阵形式

预处理?

特征?



分类器?



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）

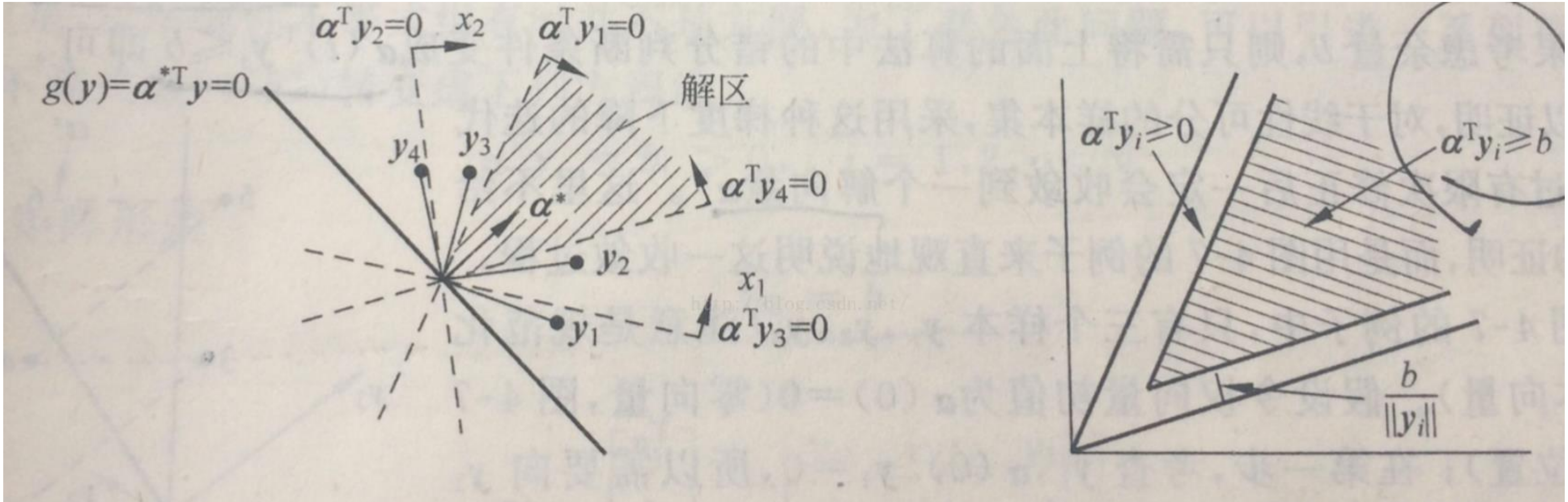


图 4-5 解向量和解区

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

$$\nabla J_P(\mathbf{a}) = \frac{\partial J_P(\mathbf{a})}{\partial \mathbf{a}} = \sum_{\mathbf{y} \in \Psi^k} (-\mathbf{y})$$

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

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \rho_k \sum_{\mathbf{y} \in \Psi^k} \mathbf{y}$$

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
    - $a^T y_i > 0 \qquad (i = 1,\dots,N)$
  - 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，

$$Ya > 0$$

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

为使解更可靠，引入余量 $b>0$

$$Ya \geqslant b > 0$$

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

$$\max J_{q2}(\boldsymbol{a}) = \sum_{i=1}^{N} \frac{1+\operatorname{sgn}(\boldsymbol{y}_i\boldsymbol{a})}{2}$$

$$\operatorname{sgn}(\boldsymbol{y}_i\boldsymbol{a}) = \begin{cases} +1, 对于\ \boldsymbol{y}_i\boldsymbol{a} \geq 0 ① \\ -1, 对于\ \boldsymbol{y}_i\boldsymbol{a} < 0 \end{cases}$$

  - 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

$$e = Ya - b$$

$$J_s(a) = \| e \|^2 = \| Ya - b \|^2 = \sum_{n=1}^{N} (a^t y_n - b_n)^2$$

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**
  - $b_i > 0\ (i = 1,\dots,N)$

– 选项一情况

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

– 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,…,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

4 Bayes分类

- 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

4.2 最小错误率Bayes决策

- **4.2.1 问题**
- **4.2.2 后验概率**
- **4.2.3 决策规则**
- **4.2.4 实例**
- **4.2.5 其它问题**
- **4.2.6 特点**

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)**

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

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

– 负对数似然函数

贝叶斯派和频率派

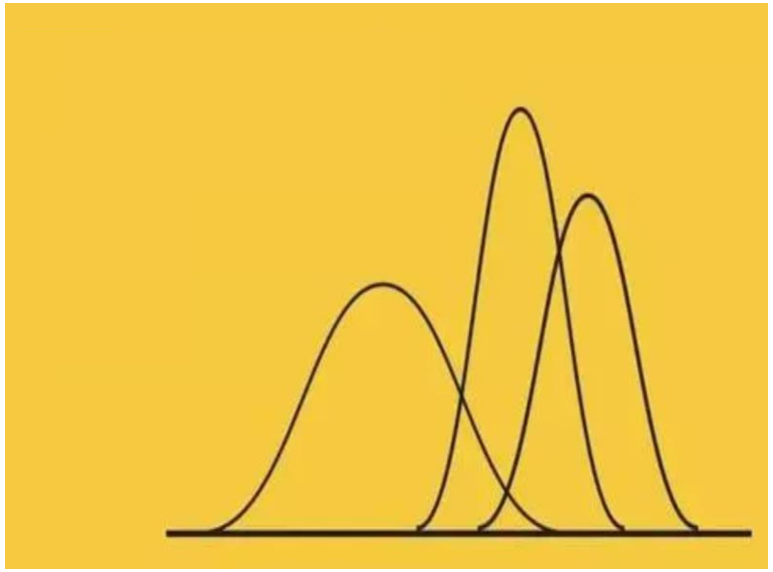
- 贝叶斯派旨在求最大后验估计
  - 认为待求参数 θ 是一个随机变量，符合一定的概率分布
  - 预设一个参数 θ 的概率分布，再用已有样本去修正这个预设（先验概率），得到最有利于样本出现的分布参数（后验概率）

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

*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



4.2.1 问题（P代表概率，p代表概率密度函数）

- **Bayes决策已知条件和问题（先考虑C = 2分类问题）**
  - 两类问题：ω_1和ω_2
  - 先验概率：$P(\omega_1)$ 和$P(\omega_2)$
  - 类条件概率密度函数：$p(x|\omega_1)$ 和$p(x|\omega_2)$
  - 发生了一个随机事件，其观察值为：特征向量x

  - **求最小错误率分类器**

4.2.2 后验概率

- **Bayes（条件概率）公式**
  - $p(x|\omega_1)\,P(\omega_1) = p(x)\,P(\omega_1|x)$
  - $p(x|\omega_2)\,P(\omega_2) = p(x)\,P(\omega_2|x)$

- 后验概率
  - $P(\omega_1|x) = p(x|\omega_1)\,P(\omega_1)\,/\,p(x)$
  - $P(\omega_2|x) = p(x|\omega_2)\,P(\omega_2)\,/\,p(x)$
  - $p(X) = p(x|\omega_1)\,P(\omega_1) + p(x|\omega_2)\,P(\omega_2)$

4.2.3 决策规则

- **决策规则**
  - 比较后验概率，取最大值进行类别判断
  - 对于未知样本x，若$P(\omega_1|x) > P(\omega_2|x)$，则$x \in \omega_1$
  - 若$P(\omega_1|x) < P(\omega_2|x)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则一 后验概率分子**
  - 比较分子$p(x|\omega_1)P(\omega_1)$ 和 $p(x|\omega_2)P(\omega_2)$，取最大
  - 对于未知样本x，若$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$，则$x \in \omega_1$
  - 若$p(x|\omega_1)P(\omega_1) < p(x|\omega_2)P(\omega_2)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则二 似然比**
  - 定义似然比函数$l(x) = p(x|\omega_1) / p(x|\omega_2)$
  - 对于未知样本x，若$l(x) > P(\omega_2) / P(\omega_1)$，则$x \in \omega_1$
  - 若$l(x) < P(\omega_2) / P(\omega_1)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则三 负对数似然比**
  - 定义负对数似然比函数$h(x) = -\ln l(x) = -\ln p(x|\omega_1) + \ln p(x|\omega_2)$
  - 对于未知样本x，若$h(x) < \ln [P(\omega_1) / P(\omega_2)]$，则$x \in \omega_1$
  - 若$h(x) > \ln [P(\omega_1) / P(\omega_2)]$，则$x \in \omega_2$

4.2.4 实例

- 已知
  - 癌细胞图像识别：正常和异常两类（即C = 2）
  - 已知未知样本特征观察值：$x = 0.5$
  - 又已知$P(\omega_1) = 0.9$和$P(\omega_2) = 0.1$
  - 查函数曲线得$p(0.5|\omega_1) = 0.2$和$p(0.5|\omega_2) = 0.4$

  - 试对未知样本$x = 0.5$进行分类

4.2.4 实例

- 已知
  - 癌细胞图像识别：正常和异常两类（即C = 2）
  - 已知未知样本特征观察值：x = 0.5
  - 又已知$P(\omega_1) = 0.9$和$P(\omega_2) = 0.1$
  - 查函数曲线得$p(0.5|\omega_1) = 0.2$和$p(0.5|\omega_2) = 0.4$
  - 试对未知样本x = 0.5进行分类

解：利用贝叶斯公式，分别计算出 ω_1 及 ω_2 的后验概率。

$$P(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{\sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)} = \frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1} = 0.818$$

$$P(\omega_2|x) = 1 - p(\omega_1|x) = 0.182$$

根据贝叶斯决策规则式(2-2)，有

$$P(\omega_1|x) = 0.818 > P(\omega_2|x) = 0.182$$

所以合理的决策是把 x 归类于正常状态。

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 错误率P(e)的定义

首先应指出所谓错误率是指平均错误率，以 $P(e)$ 来表示，其定义为

$$P(e)=\int_{-\infty}^{\infty}P(e,\mathbf{x})\mathrm{d}\mathbf{x}=\int_{-\infty}^{\infty}P(e|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2-6}$$

其中 $\int_{-\infty}^{\infty}(\quad)\mathrm{d}\mathbf{x}$ 表示在整个 d 维特征空间上的积分。

对两类别问题，从式(2-2)的决策规则可知，如果 $P(\omega_2|\mathbf{x})>P(\omega_1|\mathbf{x})$，则决策应为 ω_2。显然在作出决策 ω_2 时，$\mathbf{x}$ 的条件错误概率为 $P(\omega_1|\mathbf{x})$；反之，则应为 $P(\omega_2|\mathbf{x})$。可表示为

$$P(e|\mathbf{x})=\begin{cases}P(\omega_1|\mathbf{x}),\text{当 } P(\omega_2|\mathbf{x})>P(\omega_1|\mathbf{x})\\ P(\omega_2|\mathbf{x}),\text{当 } P(\omega_1|\mathbf{x})>P(\omega_2|\mathbf{x})\end{cases} \tag{2-7}$$

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 错误率P(e)的推导
    - 做ω_1判别时的错误概率

$$P(e_{12})=\int_{\Re_1} P(\omega_2 \mid x)p(x)dx=\int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2)dx$$

    - 做ω_2判别时的错误概率

$$P(e_{21})=\int_{\Re_2} P(\omega_1 \mid x)p(x)dx=\int_{\Re_2} P(\omega_1)p(x \mid \omega_1)dx$$

    - 总错误概率

$$P(e)=\int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2)dx+\int_{\Re_2} P(\omega_1)p(x \mid \omega_1)dx$$

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 再假设t为唯一分界点（**C=2，D=1**）
    - 做ω_1判别时的错误概率

$$P(e_{12}) = \int_{-\infty}^{t} P(\omega_2)\, p(x \mid \omega_2) dx$$

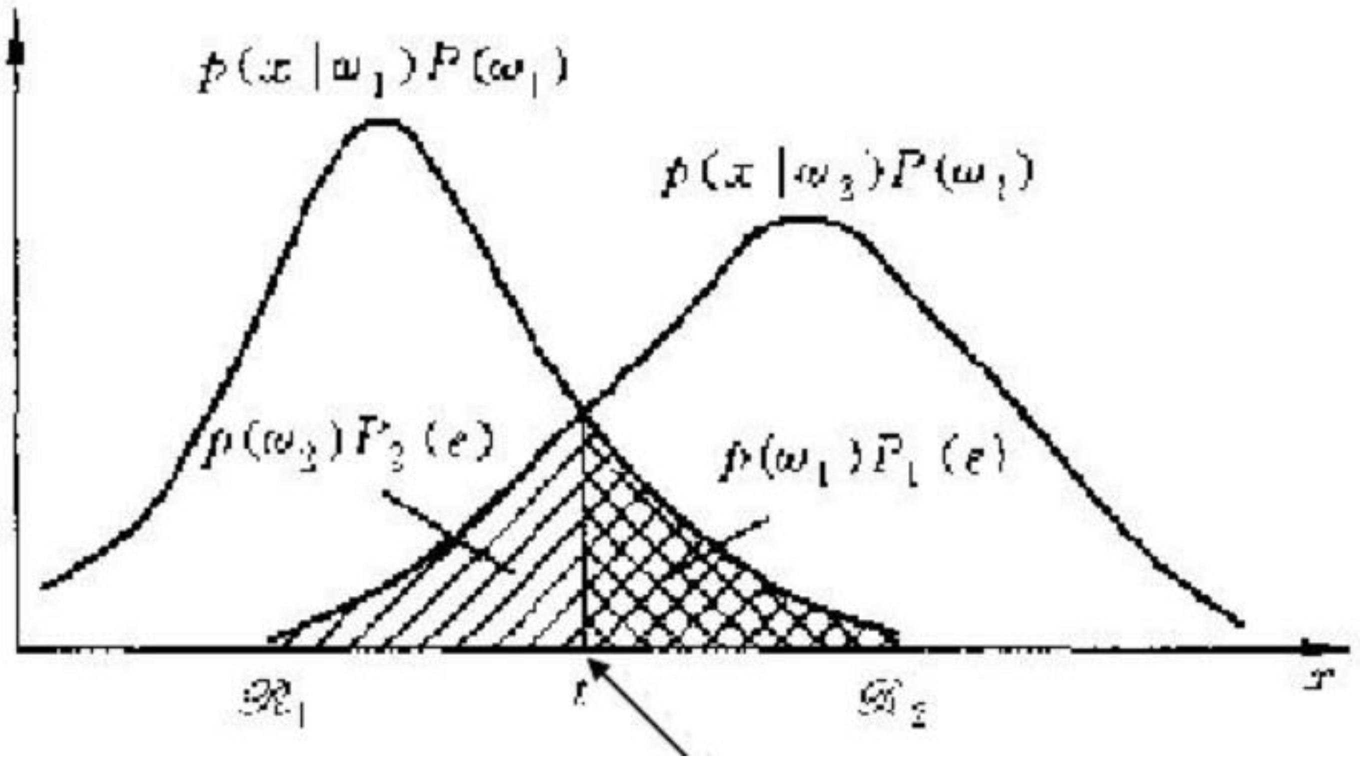
    - 做ω_2判别时的错误概率

$$P(e_{21}) = \int_{t}^{\infty} P(\omega_1)\, p(x \mid \omega_1) dx$$

    - 总错误概率

$$P(e) = \int_{-\infty}^{t} P(\omega_2)\, p(x \mid \omega_2) dx + \int_{t}^{\infty} P(\omega_1)\, p(x \mid \omega_1) dx$$

4.2.5 最小错误率的说明



$$P_1(e) = \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$$

$$P_2(e) = \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x}$$

4.2.5 最小错误率的说明

- **推广到多类（任意C类）**
  - 已知条件和问题

  - C类D维问题：ω_1， ω_2， …， ω_C
  - 先验概率：$P(\omega_1)$， $P(\omega_2)$， …， $P(\omega_C)$
  - 类条件概率密度函数：$p(x \mid \omega_1)$， $p(x \mid \omega_2)$， …， $p(x \mid \omega_C)$
  - 发生了一个随机事件，其观察值为：特征向量x

  - **求最小错误率分类器**

4.2.5 最小错误率的说明

- **推广到多类（任意C类）**
  - 判别函数
    - $P(\omega_i|x) = p(x|\omega_i)\, P(\omega_i) / p(x)$　　　$i = 1, \ldots, C$
  - 决策规则
    - 对于未知样本x，若$P(\omega_j|x) = \max P(\omega_i|x)$，则$x \in \omega_j$

4.2.6 特点

- **最小错误率Bayes决策的特点**
  - 已知条件多——各类概率分布
  - 最小错误率——概率意义上最优
  - 非线性分类器
  - 设计过程复杂

4. 3 最小风险Bayes决策

- 4.3.1 问题的提出
- 4.3.2 决策规则
- 4.3.3 其它说明
- 4.3.4 特点

4.3.1 问题的提出

- **最小错误率Bayes决策**
  - 最小错误率——概率意义上最优
  - 工程上是否最优?

- 错误分类的结果、代价或风险会是怎样的？
  - 考虑癌细胞图像识别的例子

- 出错的可能情况
  - 正常细胞ω_1错分为异常ω_2
  - 异常细胞ω_2错分为正常ω_1

4.3.1 问题的提出

- **区别状态和决策概念**
  - 状态：识别的目的是分类，把样本归类于其可能的自然状态（即类别）之一，将这种自然状态简称为状态，记为ω
  - 状态空间：所有可能的状态的集合构成状态空间，记为$\boldsymbol{\Omega}$

4.3.1 问题的提出

- **区别状态和决策概念**
  - 决策：把样本归类于某个状态，或不能进行归类，都是决策，记为α
  - 决策空间：所有可能的决策（包括拒绝决策）的集合构成决策空间，记为A

4.3.1 问题的提出

- **已知条件（类别C = 2，决策A = 2）**
  - ω_1、$P(\omega_1)$、 $p(x \mid \omega_1)$和ω_2 、$P(\omega_2)$、 $p(x \mid \omega_2)$
  - $\Omega = \{\omega_1，\omega_2\}$
  - $A = \{\alpha_1，\alpha_2\}$
  - 定义损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小风险分类器**

4.3.2 决策规则

- **判别函数**
  - $P(\omega_j | x) = p(x | \omega_j) P(\omega_j) / p(x)$ $\quad j = 1, 2$
  - $R(\alpha_i | x) = E[\lambda(\alpha_i, \ \omega_j) | x] = \lambda_{i1} P(\omega_1 | x) + \lambda_{i2} P(\omega_2 | x)$ $\quad i = 1, 2$

- **决策规则**
  - 对于未知样本x，若$R(\alpha_k | x) = \min R(\alpha_i | x)$，则$x \in \omega_k$，**即决策**$\alpha_k$

4.3.2 决策规则

- **推广到任意情况（C个类别，A个决策）**
  - ω_j、$P(\omega_j)$、 $p(x|\omega_j)$　　　　　$j = 1, \ldots, C$
  - $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$
  - $A = \{\alpha_1, \alpha_2, \ldots, \alpha_A\}$
  - 定义损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小风险分类器**

4.3.2 决策规则

- **判别函数**
  - $P(\omega_j \mid x) = p(x \mid \omega_j)\, P(\omega_j) / p(x)$　　　j = 1, 2, …, C
  - $R(\alpha_i, x) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1} \lambda(\alpha_i, \omega_j) P(\omega_j \mid x)$　　i = 1, 2, …, A
- **决策规则**
  - 对于未知样本x，若$R(\alpha_k \mid x) = \min R(\alpha_i \mid x)$，则$x \in \omega_k$，**即决策**$\alpha_k$

4.3.3 其它说明

- **最小风险与最小错误率的关系**
  - 0-1损失函数

$\lambda(\alpha_i, \omega_j) = 0 \qquad i = j$

$\lambda(\alpha_i, \omega_j) = 1 \qquad i \neq j$

$$R(\alpha_i, x) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1} \lambda(\alpha_i, \omega_j) P(\omega_j \mid x)$$

$$R(\alpha_i, x) = \sum_{\substack{j=1 \\ j \neq i}} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

4.3.3 其它说明

- **最小风险与最小错误率的关系**

$$R(\alpha_i, x) = \sum_{\substack{j=1 \\ j \neq i}}^{L} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

  - 对于未知样本x，若$R(\alpha_k \mid x) = \min R(\alpha_i \mid x)$，则$x \in \omega_k$
  - 对于未知样本x，若$P(\omega_k|x) = \max P(\omega_i|x)$，则$x \in \omega_k$

  - **结论：最小错误率Bayes决策，等价于0-1损失函数的最小风险Bayes决策**

4.3.4 特点

- **最小风险Bayes决策的特点**
  - 已知条件多——各类概率分布及风险系数
  - 最小错误风险——概率意义上最优
  - 非线性分类器
  - 设计过程复杂

例题

- **1、已知**
  - 甲类：$P(\omega_1) = 0.7$和类条件概率密度函数$p(x|\omega_1)$
  - 乙类：$P(\omega_2) = 0.3$和类条件概率密度函数$p(x|\omega_2)$
  - 今有待分类样本特征观察值$x = 10$，且由函数曲线查得$p(10|\omega_1) = 0.2$，$p(10|\omega_2) = 0.5$
  - (1)试用最小错误率Bayes决策对样本$x = 10$进行分类
  - (2)试用最小风险Bayes决策对该样本进行分类，设$\lambda_{11}=\lambda_{22}=0$，$\lambda_{12}=2$，$\lambda_{21}=1$

4.4 最小最大决策

- 4.4.1 问题的提出
- 4.4.2 期望损失
- 4.4.3 最小最大风险
- 4.4.4 特点

4.4.1 问题的提出

- 已知条件和问题（C = 2情况）
  - 先验概率：考虑$P(\omega_1)$ 和$P(\omega_2)$未知或不确定的情况
  - 此时绝对意义的最小风险不存在
  - 如何求Bayes分类器

  - 思路
    - 假设$P(\omega_1)$ 和$P(\omega_2)$确定
    - 设计系列最小风险Bayes分类器
    - 取其中最大风险为最小的一个来用
    - 目的是控制最大风险

4.4.1 问题的提出

- **问题（类别C = 2，决策A = 2）**
  - ω_1、$p(x \mid \omega_1)$和ω_2、$p(x \mid \omega_2)$
  - $\Omega = \{\omega_1，\omega_2\}$
  - $A = \{\alpha_1，\alpha_2\}$
  - 损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小最大风险分类器**

4.4.2 期望损失

- 期望损失的推导（回顾最小错误率证明）

$$P(e)=\int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2)dx+\int_{\Re_2} P(\omega_1)\, p(x \mid \omega_1)dx$$

$$P(e_{12})=\int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2)dx$$

$$P(e_{21})=\int_{\Re_2} P(\omega_1)\, p(x \mid \omega_1)dx$$

4.4.2 期望损失

- 期望损失的推导

$$R=\int R(\alpha(\boldsymbol{x})|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_{\mathcal{R}_1}R(\alpha_1|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\int_{\mathcal{R}_2}R(\alpha_2|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$=\int_{\mathcal{R}_1}[\lambda_{11}P(\omega_1)p(\boldsymbol{x}|\omega_1)+\lambda_{12}P(\omega_2)p(\boldsymbol{x}|\omega_2)]\mathrm{d}\boldsymbol{x}+\int_{\mathcal{R}_2}[\lambda_{21}P(\omega_1)p(\boldsymbol{x}|\omega_1)$$

$$+\lambda_{22}P(\omega_2)p(\boldsymbol{x}|\omega_2)]\mathrm{d}\boldsymbol{x}$$

$$P(\omega_2)=1-P(\omega_1)$$

$$\int_{\Re_2}p(x|\omega_1)dx=1-\int_{\Re_1}p(x|\omega_1)dx$$

4.4.2 期望损失

- **期望损失的推导**

$$R=\lambda_{22}+(\lambda_{12}-\lambda_{22})\int_{\mathcal{R}_1}p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}+P(\omega_1)[(\lambda_{11}-\lambda_{22})$$
$$+(\lambda_{21}-\lambda_{11})\int_{\mathcal{R}_2}p(\boldsymbol{x}|\omega_1)\mathrm{d}\boldsymbol{x}-(\lambda_{12}-\lambda_{22})\int_{\mathcal{R}_1}p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}]$$

$$a=\lambda_{22}+(\lambda_{12}-\lambda_{22})\int_{\mathcal{R}_1}p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}$$

$$b=(\lambda_{11}-\lambda_{22})+(\lambda_{21}-\lambda_{11})\int_{\mathcal{R}_2}p(\boldsymbol{x}|\omega_1)\mathrm{d}\boldsymbol{x}-(\lambda_{12}-\lambda_{22})\int_{\mathcal{R}_1}p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}$$

推导思路：替换后改成只有P(w2)有关的形式

4.4.3 最小最大风险

- **最小最大风险的解**
  - 任给$P(\omega_1)^*$的值，$P(\omega_2)^* = 1 - P(\omega_1)^*$
  - ↓
  - 求对应的最小风险Bayes分类器，设决策域为R_1和R_2
  - ↓
  - 计算对应的a^*和b^*，得到$R^* = a^* – b^* P(\omega_1)^*$
  - ↓
  - 重复上述计算步骤，得到一系列最小风险Bayes分类器，并 可得$P(\omega_1)^*$——R^* 关系曲线
  - ↓
  - 比较所有最大风险，取其中最小的一个，作为最终的分类器

4.4.4 最小最大决策

- **最小最大决策的特点**
  - 已知条件多——各类概率分布及风险系数
  - 最小最大风险——概率意义上最优
  - 非线性分类器
  - 设计过程很复杂

THE END!