数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

实验课安排

- 本周四上午1、2节，二主楼B403

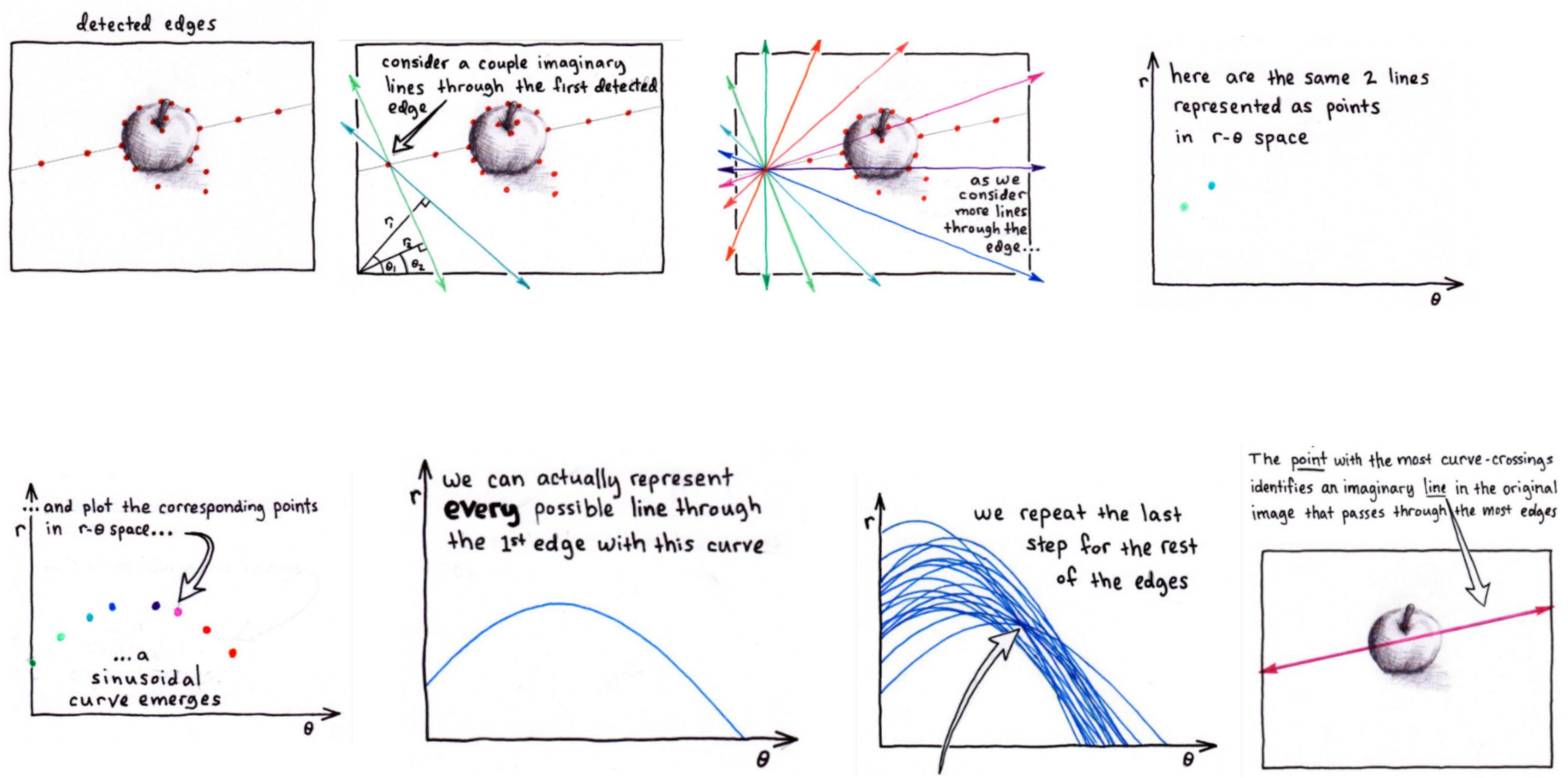
上节回顾

- 不同模型的性能比较
  - 假设检验方法
  - 偏差-方差分解
- 分类任务
  - 特征
  - 分类模型与学习方法

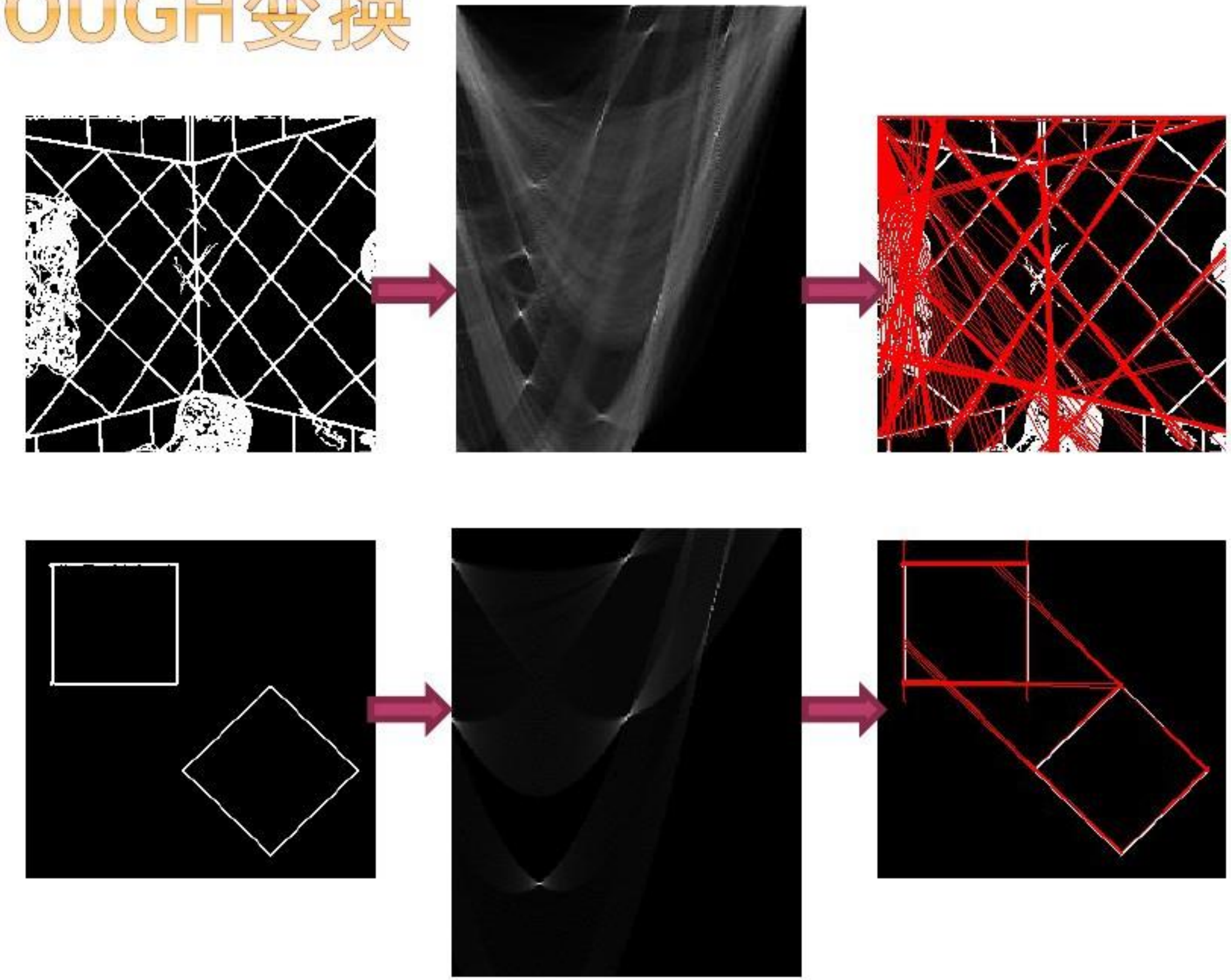
HOUGH变换

- 经典的Hough变换主要涉及图像中的直线检测，但是后来Hough变换得到了扩展，被用于任意形状位置的检测，其中最常用的是圆形或者椭圆。
  - Hough变换最简单的实例就是用于直线检测的线性变换。
- 使用参数空间(r,θ)，直线公式可以重写为
  - $y = \left(-\frac{\cos\theta}{\sin\theta}\right)x + \left(\frac{r}{\sin\theta}\right)$，
  - 这也就等价于公式
    - $r = x\cos\theta + y\sin\theta$。

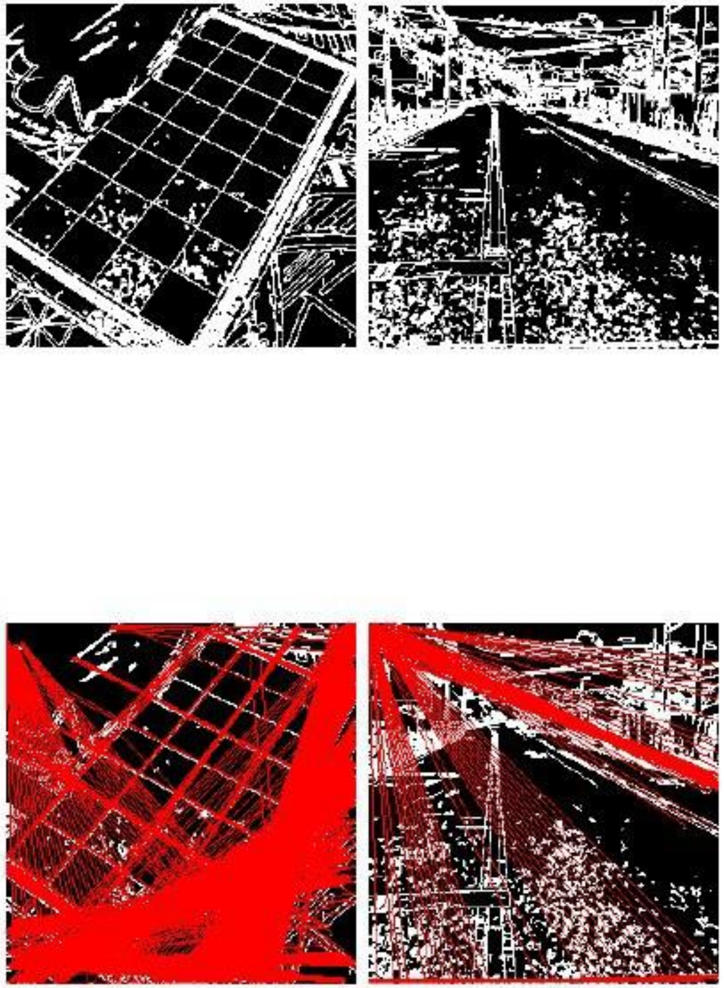
直角坐标系中的一条直线对应于极坐标系下的一个点.

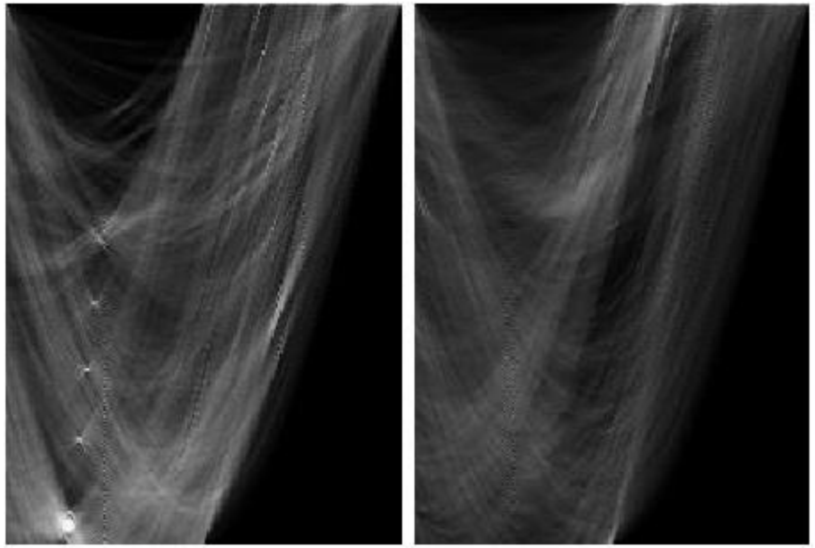

detected edges
consider a couple imaginary lines through the first detected edge
r_1
r_2
θ_1
θ_2
as we consider more lines through the edge...
here are the same 2 lines represented as points in r-θ space
r
θ
...and plot the corresponding points in r-θ space...
...a sinusoidal curve emerges
r
θ
we can actually represent **every** possible line through the 1st edge with this curve
r
θ
we repeat the last step for the rest of the edges
r
θ
The point with the most curve-crossings identifies an imaginary line in the original image that passes through the most edges

HOUGH变换



HOUGH变换





RADON变换

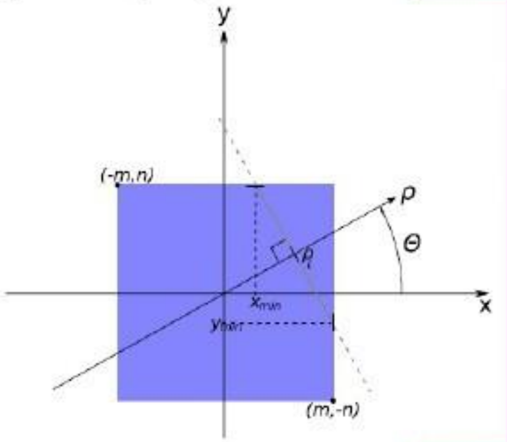
- 二维Radon变换属于一种积分变换，是由众多直线函数上的积分组成。
- Radon变换被广泛的应用于断层摄影术。
  - 如果f表示一个未知的密度函数，那么Radon变换表示一种发散数据，可用于断层摄影术的输出。
  - 因此，Radon逆变换可以用于从发散数据中重构原始密度函数，从而形成了断层摄影术重构技术的数学基础。
- Radon变换数据通常被称为正弦图，因为Dirac三角函数的Radon变换结果看上去是呈一个正弦波的分布。
  - 许多小物体的Radon变换在显示上都会呈现出模糊正弦波的样子，而这些正弦波具有不同的振幅和相位。

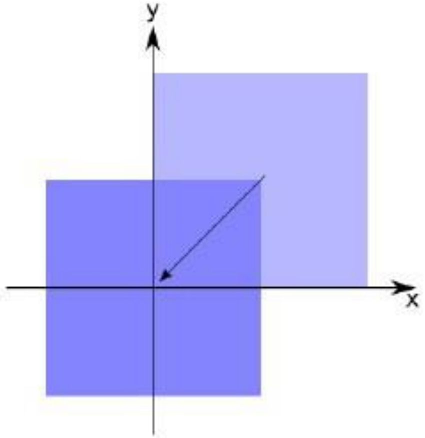
RADON变换

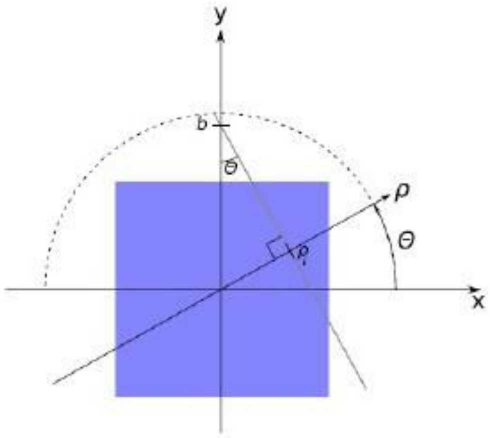
- Radon变换的数学公式可以描述为
  - $R(\rho,\theta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\delta(\rho-x\cos\theta-y\sin\theta)dxdy$,
  - 其中$\rho=x\cos\theta+y\sin\theta$表示一条直线，$\delta(.)$为Dirac三角函数。
- Dirac三角函数大致可以认为是一种在原点无穷大而其余地方为零的函数，即
  - $\delta(x)=\begin{cases}+\infty & x=0\\ 0 & x\neq 0\end{cases}$,
  - 并且满足如下条件
    - $\int_{-\infty}^{\infty}\delta(x)dx=1$ 。
- Radon变换实际上是从笛卡尔直角坐标(x,y)到极坐标(ρ,θ)的映射

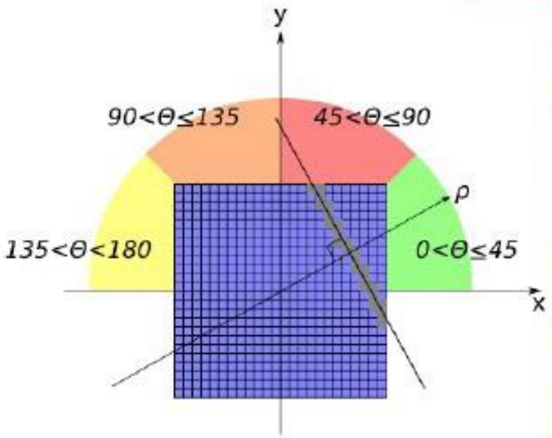
RADON变换

- Radon变换的具体实现，大致可以分为四个步骤
  - 将坐标原点移至图像中心
  - 直线参数计算
  - 直线坐标的最大最小值计算
  - 根据角度进行分区讨论

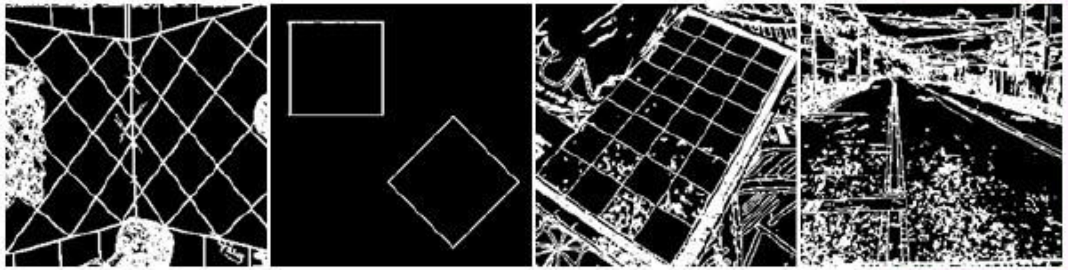


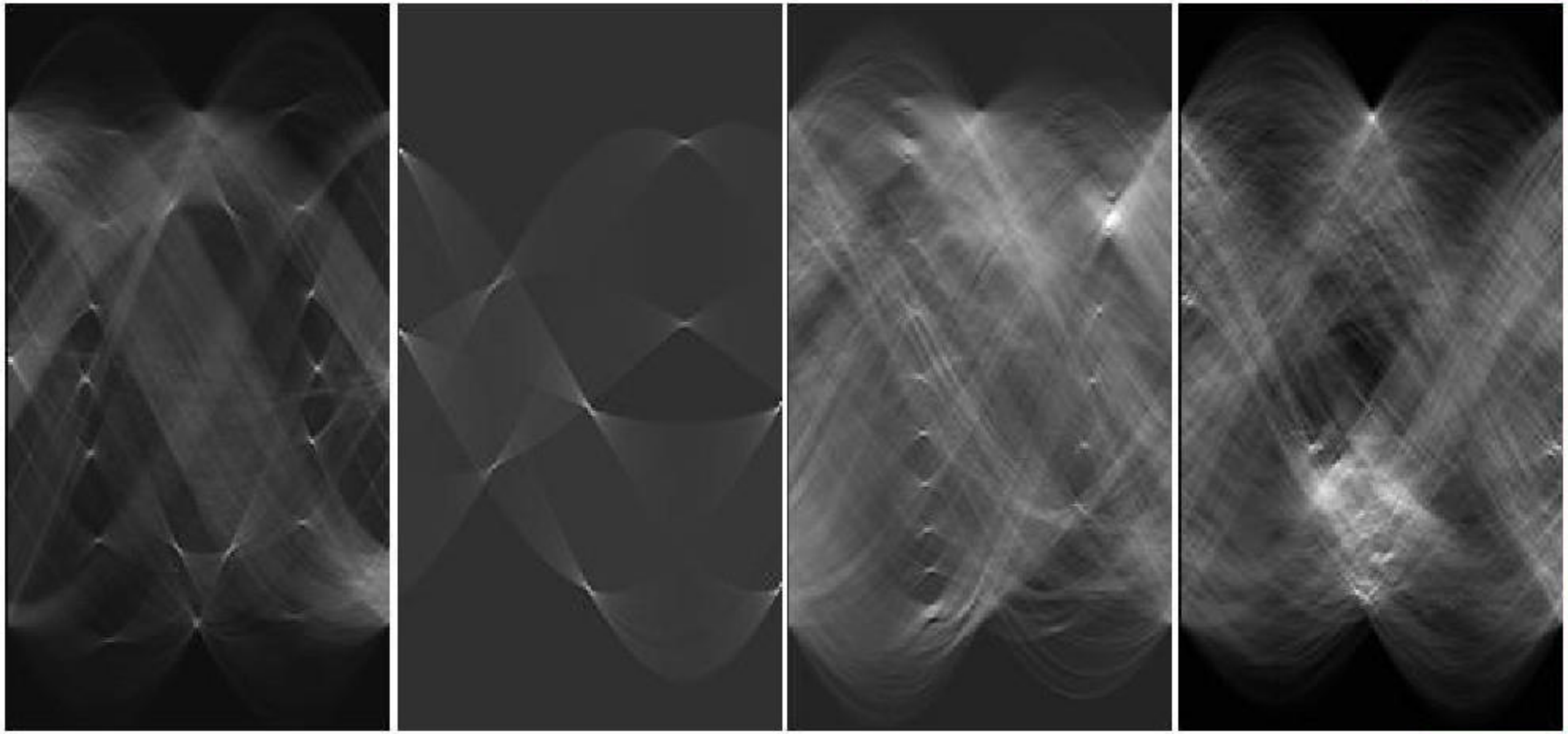






RADON变换



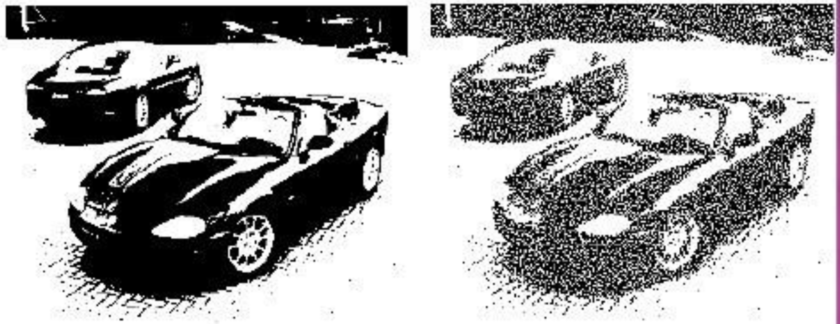


傅里叶描述子

- 傅里叶描述子，就是对封闭曲线进行傅里叶变换，将变换后得到的傅里叶系数作为表示图像形状的特征。
- 可以将形状表示为具有N个顶点的集合$\{z(i): i, \ldots, N\}$，那么傅里叶描述子$\{c(k): k = -{}^{N}/_{2} + 1, \ldots, {}^{N}/_{2}\}$就是集合$z$的傅里叶变换系数集合，即
  - $z_{\mathrm{i}} = \sum_{k=-{}^{N}/_{2}+1}^{{}^{N}/_{2}} c_k e^{2\pi j \frac{ki}{N}}$，
  - 而$c(k)$与$z(i)$之间的逆变换为
    - $c_k = \frac{1}{N} \sum_{i=1}^{N} z_i e^{-2\pi j \frac{ik}{N}}$。
- 为了进行傅里叶描述子的计算，就需要两步操作：
  - 一是从二值图像中得到一维复数数组，用以作为一维傅里叶变换的输入
  - 二是进行一维傅里叶正变换。

傅里叶描述子

- 为了验证所得到的傅里叶描述子是否正确，可以对傅里叶描述子进行傅里叶逆变换，然后从所得到的一维复数数组中恢复二值图像，通过比较新旧图像就可以看出所得到的傅里叶描述子是否正确。

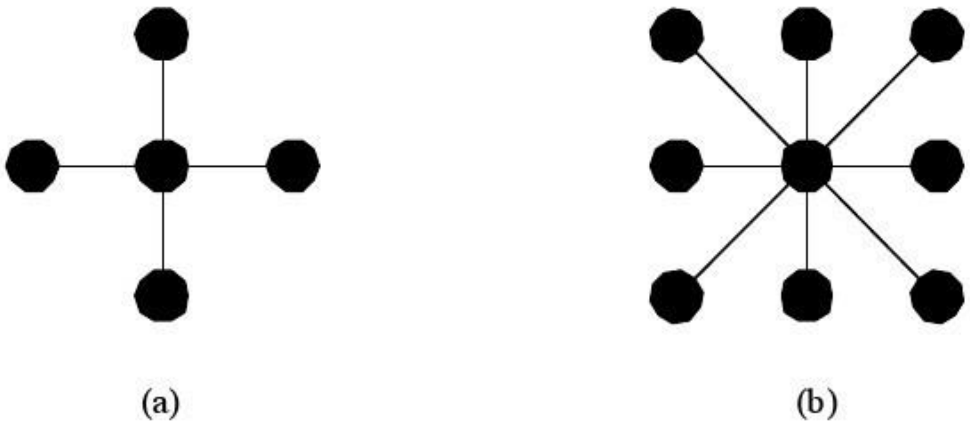


- 区域形状的描述
  - 常针对二值图
  - 属于手工特征

8.3.1几个基本概念

◆ 邻域与邻接

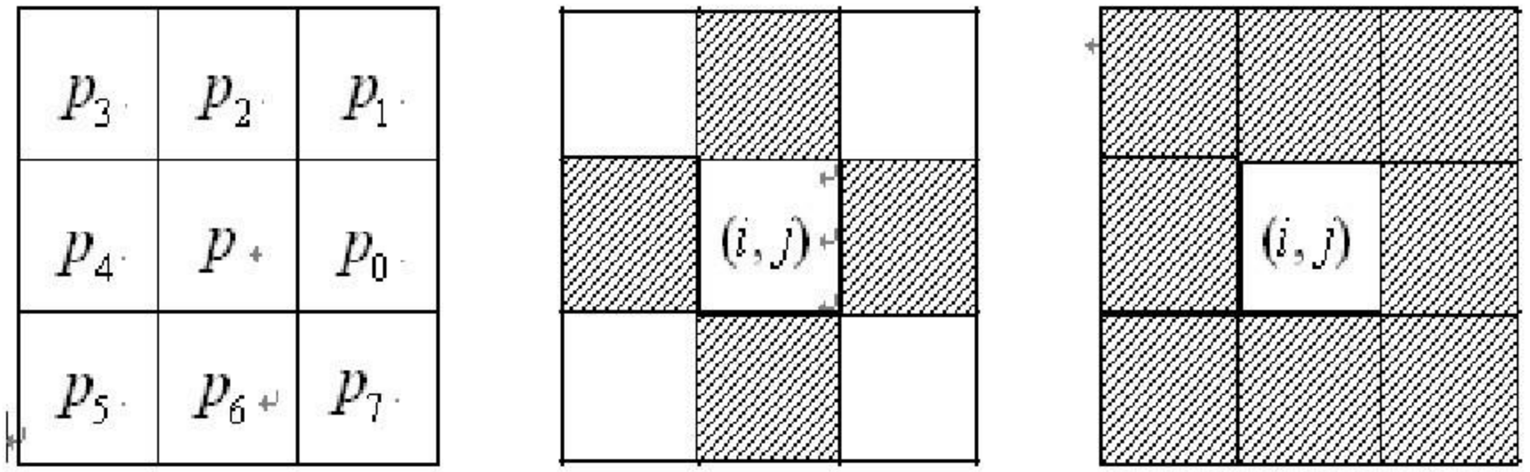
- 对于任意像素(i, j), (s, t)是一对适当的整数，则把像素的集合{(i+s, j+t)}叫做像素(i, j)的邻域.
- 直观上看，这是像素(i, j)附近的像素形成的区域.
- 最经常采用的是4-邻域和8-邻域



4-邻域和8-邻域

邻域与邻接

- 互为4-邻域的两像素叫4-邻接。
- 互为8-邻域的两像素叫8-邻接。



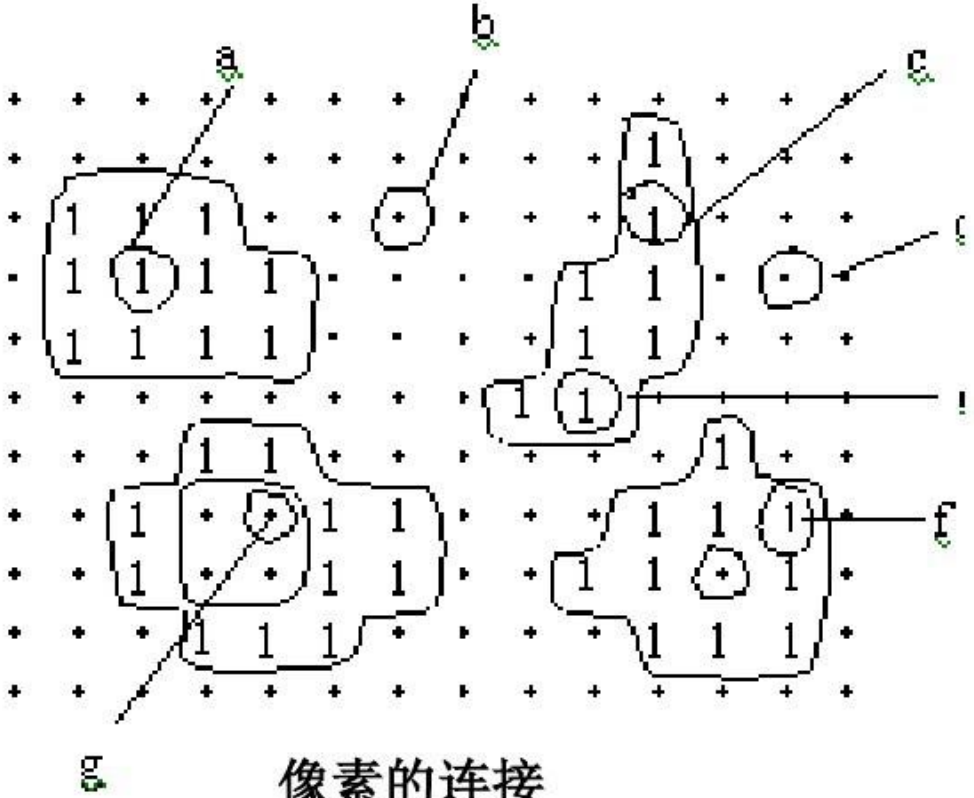
（a）　　（b）　　（c）

图 8-1 邻接像素的种类

（a）像素的编号　　（b）4-邻接　　（c）8-邻接

◆ 像素的连接

对于图像中具有相同值的两个像素A和B，如果所有和A、B具有相同值的像素序列 $L_0(=A), L_1, L_2, \Lambda, L_{N-1}, L_N(=B)$ 存在，并且 L_{i-1} 和 L_i 互为4-邻接或8-邻接，那么像素和叫做4-连接或8-连接，以上的像素序列叫4-路径或8-路径。



像素的连接

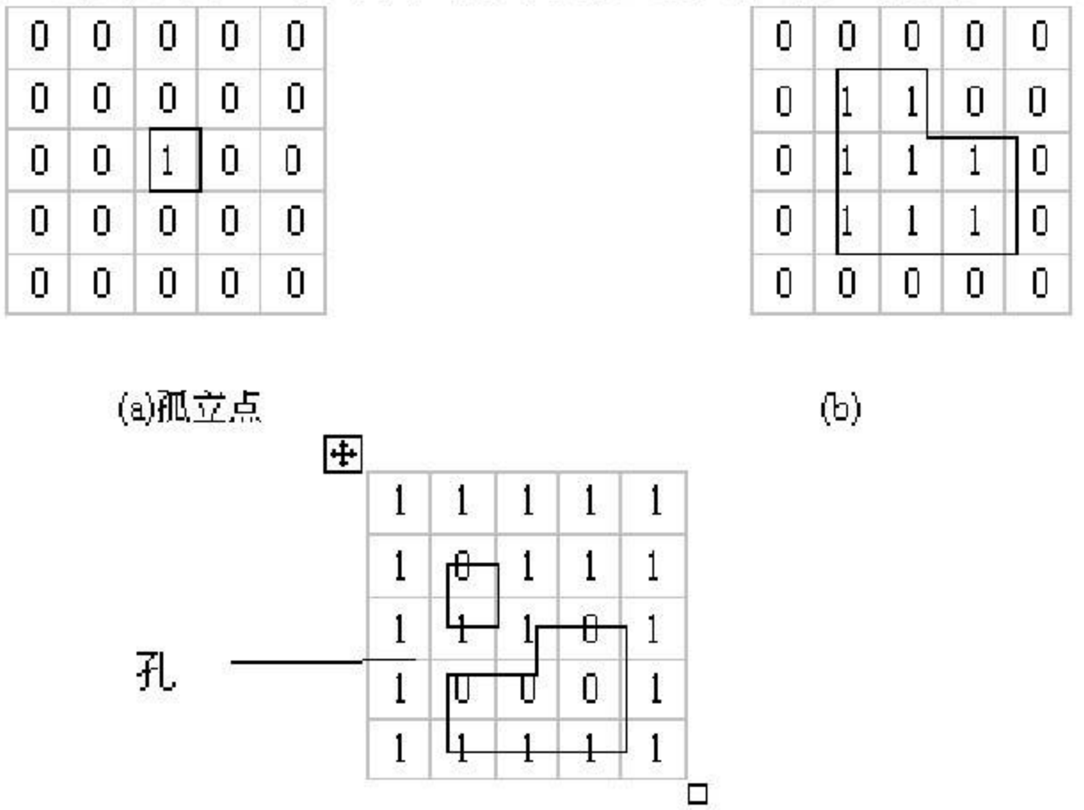
◆ 连接成分

- 在图像中，把互相连接的像素的集合汇集为一组，于是具有若干个0值的像素和具有若干个1值的像素的组就产生了。把这些组叫做连接成分，也称作连通成分。
- 在研究一个图像连接成分的场合，若1像素的连接成分用4-连接或8-连接，而0像素连接成分不用相反的8-连接或4-连接就会产生矛盾。
- 假设各个1像素用8-连接，则其中的0像素就被包围起来。如果对0像素也用8-连接，这就会与左下的0像素连接起来，从而产生矛盾。因此0像素和1像素应采用互反的连接形式，即如果1像素采用8-连接，则0像素必须采用4-连接。

0	0	0	0	0
0	1	1	1	0
0	1	0	1	0
0	0	1	1	0
0	0	0	0	0

连接性矛盾示意图

- 在0-像素的连接成分中，如果存在和图像外围的1行或1列的0-像素不相连接的成分，则称之为孔。不包含有孔的1像素连接成分叫做单连接成分。含有孔的l像素连接成分叫做多重连接成分。



(a)孤立点 (b) (c)

图 8-4 连接成分

(a)孤立点 (b)单连接成分(c)多重连接成分

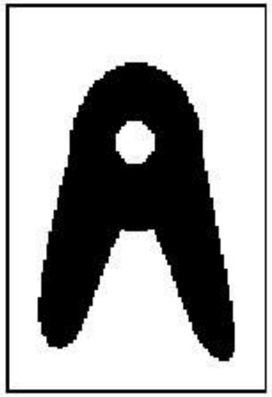
8.3.2区域内部空间域分析

◆ 区域内部空间域分析是不经过变换而直接在图像的空间域，对区域内提取形状特征。

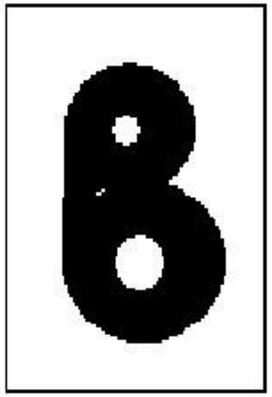
1. 欧拉数 欧拉数=连接成分数-孔数

✓ 图像的欧拉数是图像的拓扑特性之一，它表明了图像的连通性。下图（a)的图形有一个连接成分和一个孔，所以它的欧拉数为0，而下图（b）有一个连接成分和两个孔，所以它的欧拉数为-1。

✓ 可见通过欧拉数可用于目标识别。



(a)

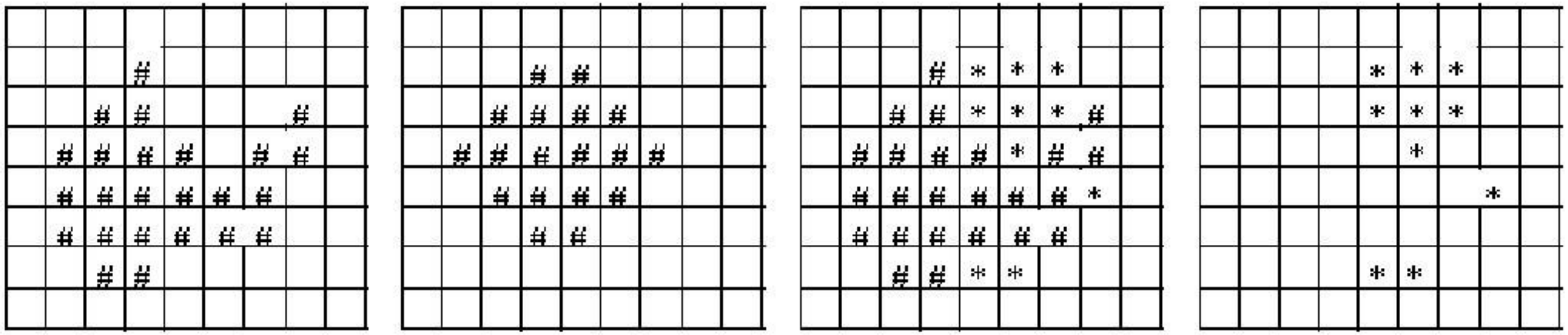


(b)

具有欧拉数为0和-1的图形

2.凹凸性

- 凹凸性是区域的基本特征之一，区域凹凸性可通过以下方法进行判别：区域内任意两像素间的连线穿过区域外的像素，则此区域为凹形。相反，连接图形内任意两个像素的线段，如果不通过这个图形以外的像素，则这个图形称为是凸的。任何一个图形，把包含它的最小的凸图形叫这个图形的凸闭包。
- 凸图形的凸闭包就是它本身。从凸闭包除去原始图形的部分后，所产生的图形的位置和形状将成为形状特征分析的重要线索。凹形面积可将凸封闭包减去凹形得到。



（a）凹形　（b）凸形　（c）(a)中凹形的凸封闭包　（d）凹形面积

区域的凹凸性

3.距离

- 距离在实际图像处理过程中往往是作为一个特征量出现，因此对其精度的要求并不是很高。所以对于给定图像中三点A,B,C，当函数D(A,B)满足下式的条件时，把D(A,B)叫做A和B的距离，也称为距离函数。

$$\begin{cases} D(A,B) \ge 0 \\ D(A,B) = D(B,A) \\ D(A,C) \le D(A,B) + D(B,C) \end{cases}$$

- 第一个式子表示距离具有非负性，并且当A和B重合时，等号成立；
- 第二个式子表示距离具有对称性
- 第三个式子表示距离的三角不等式。

✓ 计算点(i, j)和(h, k)间距离常采用的几种方法：

（1）欧氏距离，用d_e来表示。

$$d_e[(i,j),(h,k)] = [(i-h)^2 + (j-k)^2]^{1/2}$$

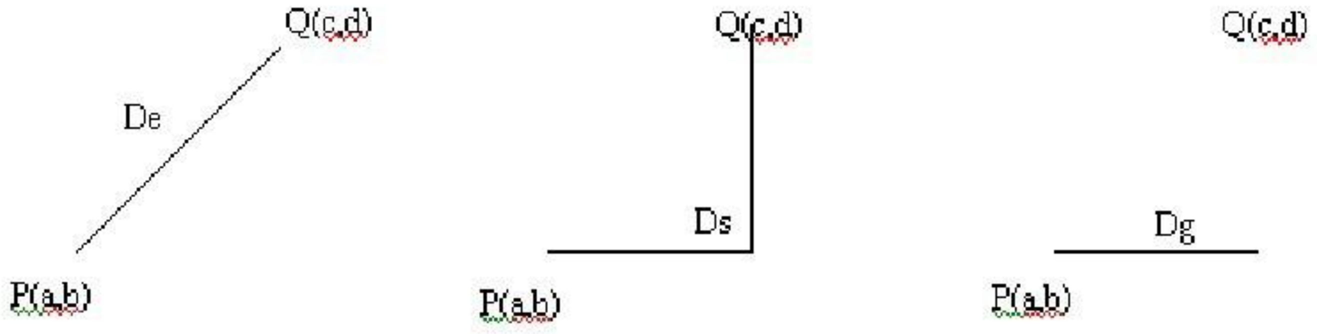
（2） 4-邻域距离，也称为街区距离。

$$d_e[(i,j),(h,k)] = [(i-h)^2 + (j-k)^2]^{1/2}$$

（3） 8-邻域距离，也称为棋盘距离。

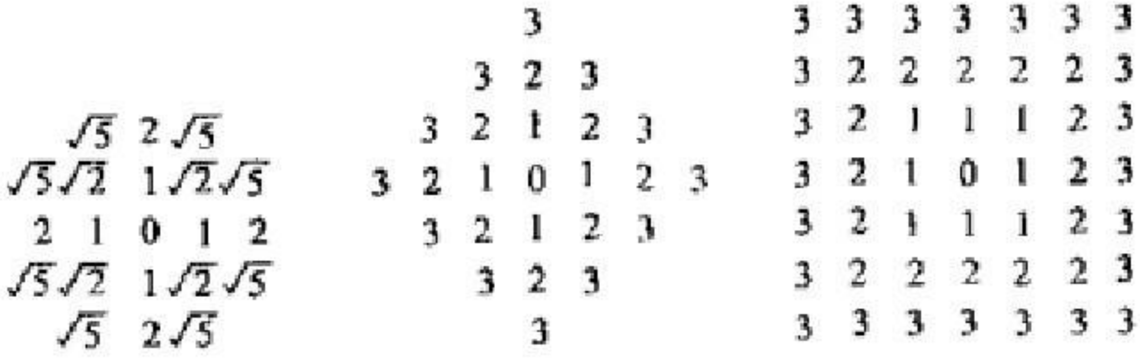
$$d_g[(i,j),(h,k)] = \max(|i-h|, |j-k|)$$

这三种距离之间的关系：$d_g \le d_s \le d_e$ ， 如图所示。街区距离和棋、盘距离都是欧式距离的一种近似。



（a）欧式距离　　（b）街区距离　　（c）棋盘距离

- 下图中表示了以中心像素为原点的各像素的距离。从离开一个像素的等距离线可以看出，在欧氏距离中大致呈圆形，在棋盘距离中呈方形，在街区距离中呈倾斜45度的正方形。街区距离是图像中两点间最短的4－连通的长度，而棋盘距离则是两点间最短的8－连通的长度。
- 此外，把4-邻域距离和8-邻域距离组合起来而得到的八角形距离有时也被采用，它的等距线呈八角形。



（a） （b） （c）

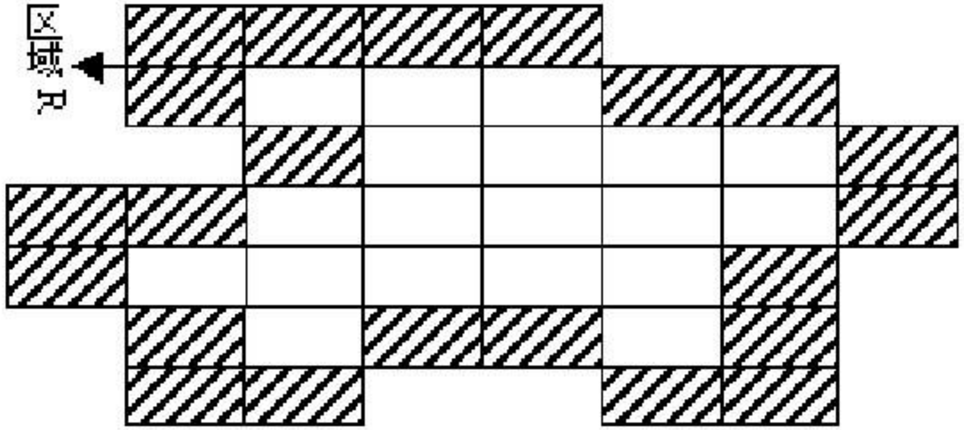
离开单个像素的距离

（a）欧几里德距离 （b）4-邻域距离 （c）8-邻域距离

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。

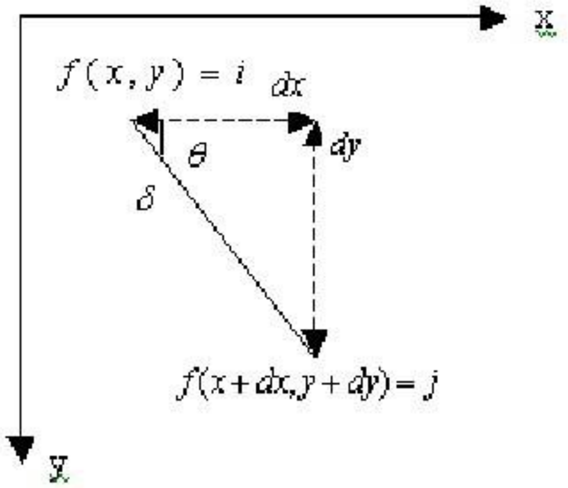
图像的纹理特征

- 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

I

1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

GLCM

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

- 分方向
- 构造矩阵
- 统计个数
- 形成特征向量

1. 0º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

$$\begin{matrix}
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3
\end{matrix}$$

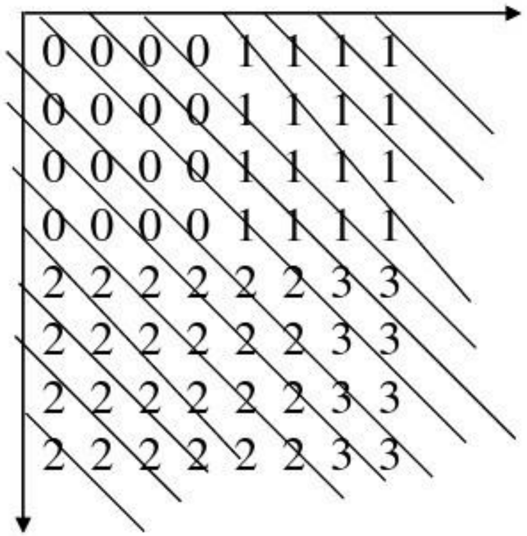
一幅数字灰度图像

$$\begin{matrix}
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\ \hline
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\ \hline
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\ \hline
2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\ \hline
\end{matrix}$$

0º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

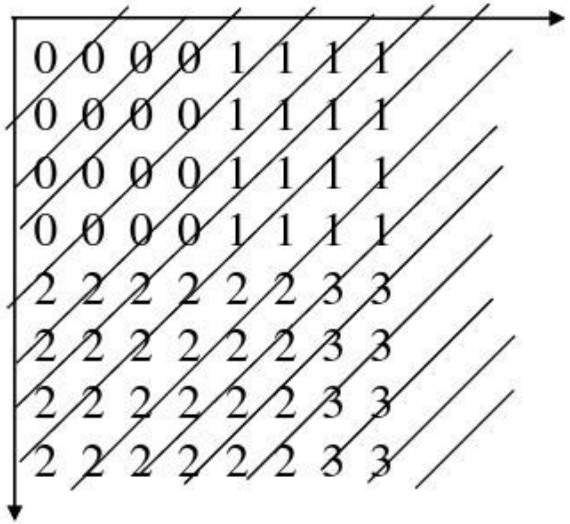
当 $\theta = 90^\circ$ 时，$dx = 0, dy = -1$。

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

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

(a)

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

(b)

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

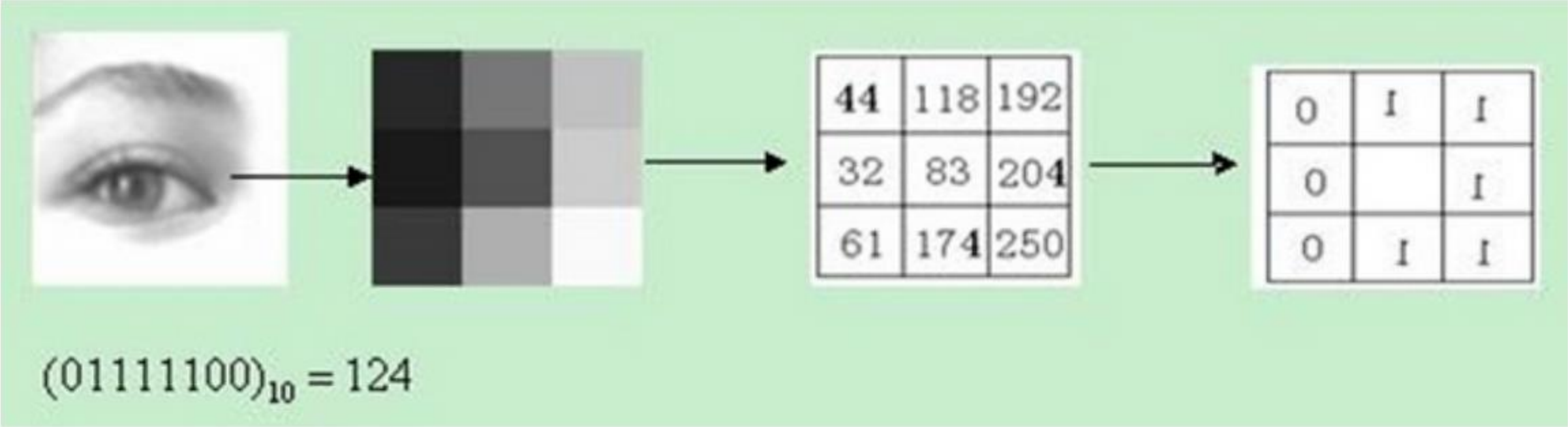
(c)

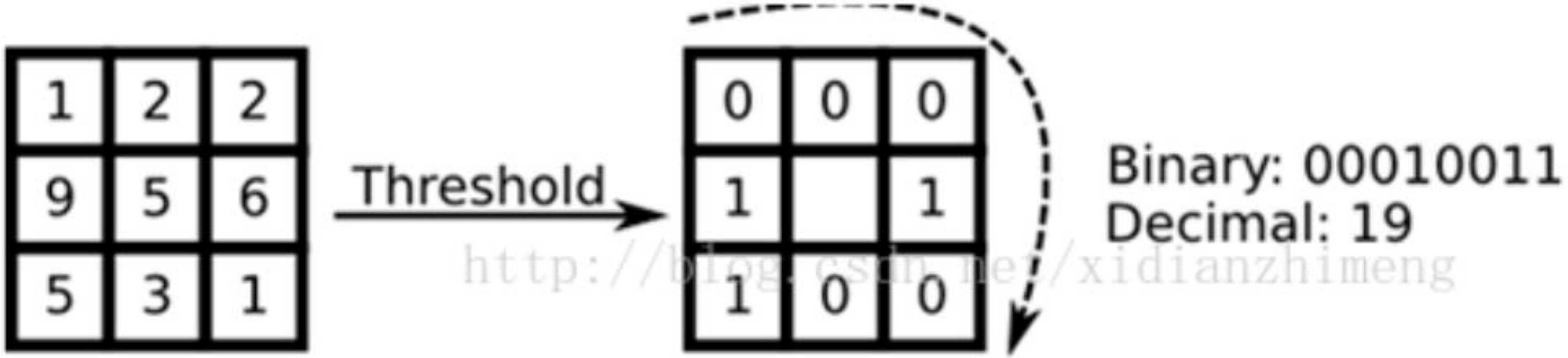
$$P(135°)=\begin{bmatrix}9 & 3 & 3 & 0\\ 3 & 9 & 3 & 1\\ 3 & 3 & 15 & 3\\ 0 & 1 & 3 & 3\end{bmatrix}$$

(d)

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

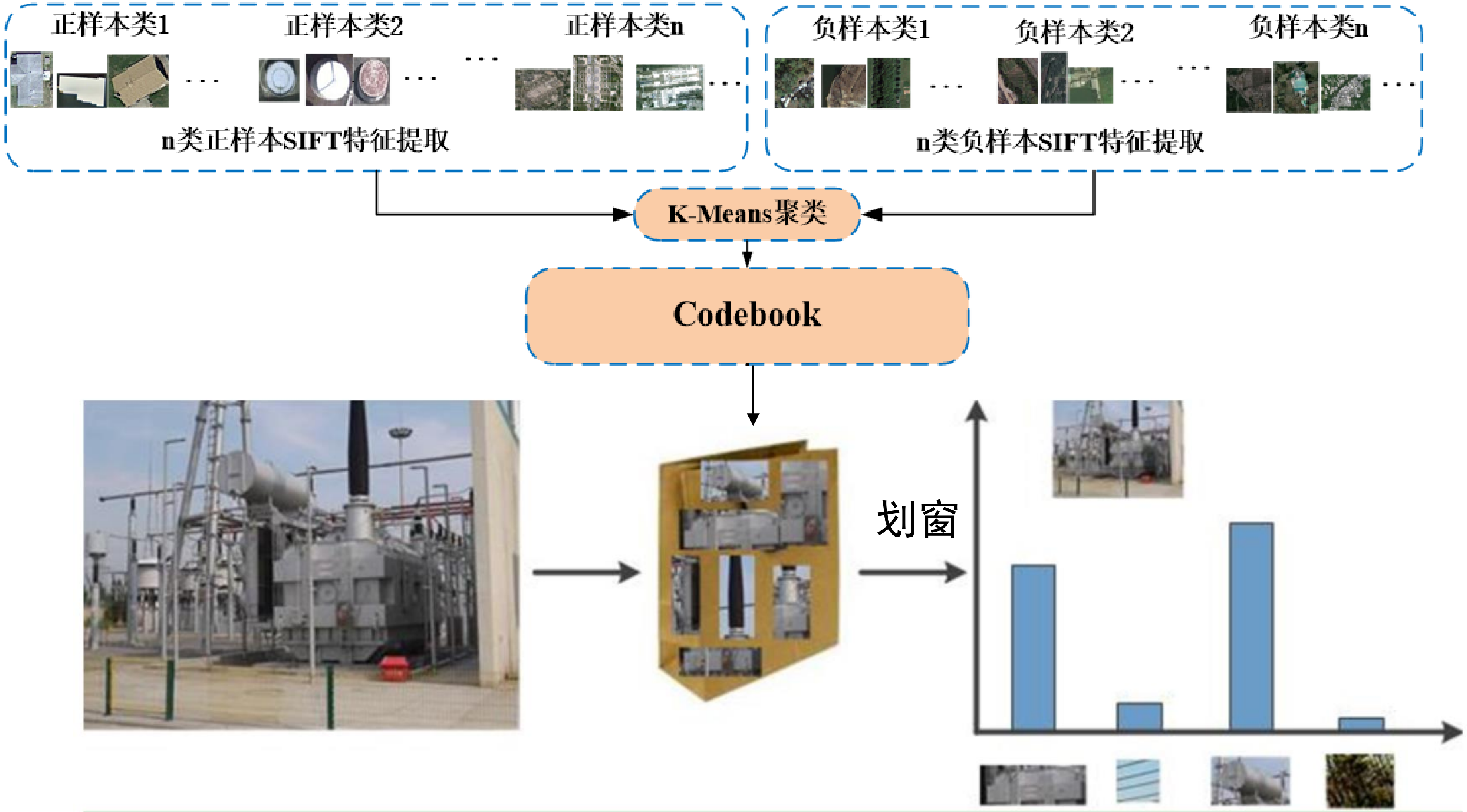
视觉词袋（Bag-of-words，BOW）

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。

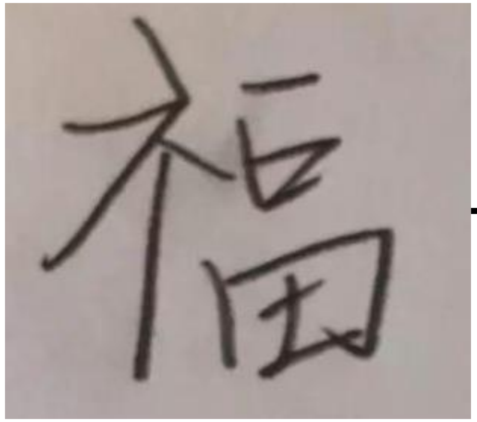
基于特征袋模型的分类器



图像　视觉词汇字典　直方图

大作业1：福字识别

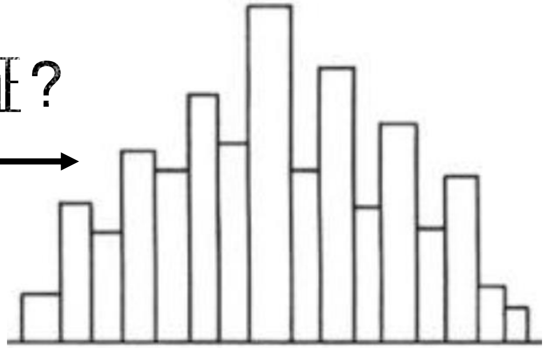
训练?



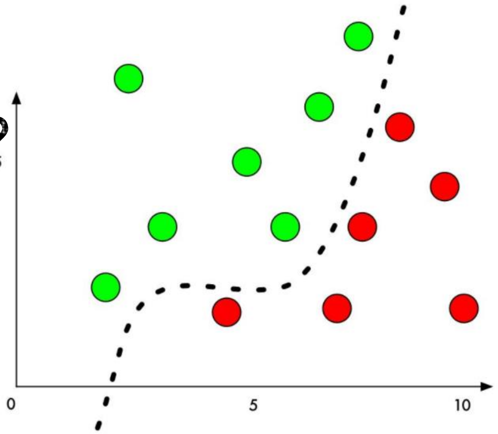
提供矩阵形式

预处理?

特征?



分类器?



线性分类器

- 1　线性分类器基础
- 2　垂直平分分类器
- 3　Fisher投影准则
- 4　感知准则
- 5　最小错分样本数准则
- 6　最小平方误差准则

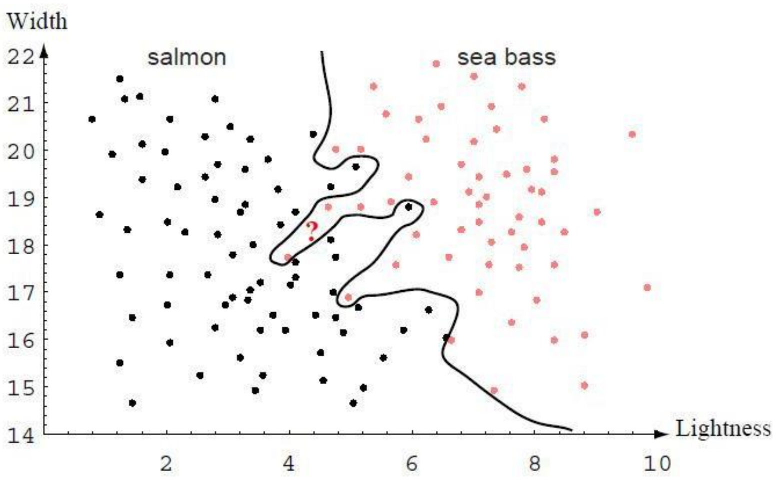
1　线性分类器基础

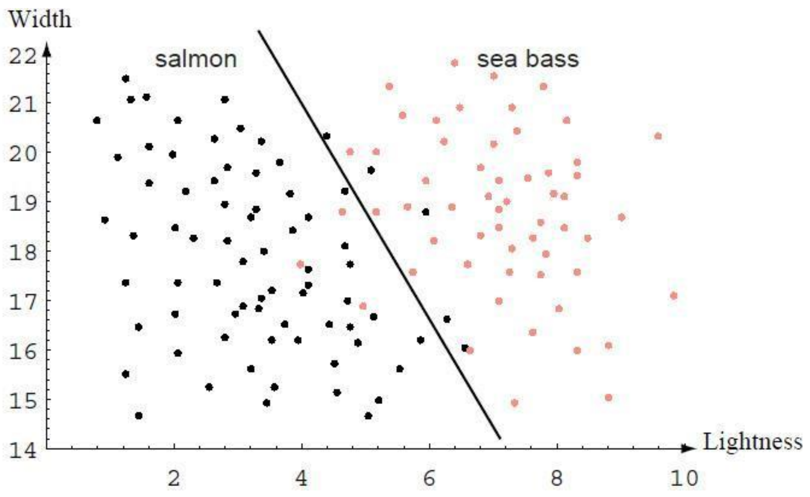
- 1.1　数学基础知识
- 1.2　线性分类器概念
- 1.3　线性判别函数
- 1.4　增广变换
- 1.5　相关概念归纳
- 1.6　线性分类器设计概述

1.1 数学基础知识

- **相关的数学基础包括**
  - 矩阵
  - 向量
  - 矩阵和向量的转置
  - 向量运算
  - 矩阵运算

1.2 线性分类器概念


Width
salmon
sea bass
?
Lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10


Width
salmon
sea bass
Lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10

1.2　线性分类器概念

- [线性分类器] 对于两类的分类问题，采用线性判别函数划分特征空间（即采用直线或平面等将两类样本在特征空间中的区域划分开），这样的分类器是线性分类器。

- 线性分类器特点：特征空间一分为二，适合于解决两类的分类问题

1.3 线性判别函数

- **两类二维问题**
  - C——类别数，D——维数，N——样本数
  - C = 2，D =2

  - 直线方程
    - 代数形式$w_1x_1 + w_2x_2 + w_0 = 0$
    - 向量形式$w^Tx + w_0 = 0$

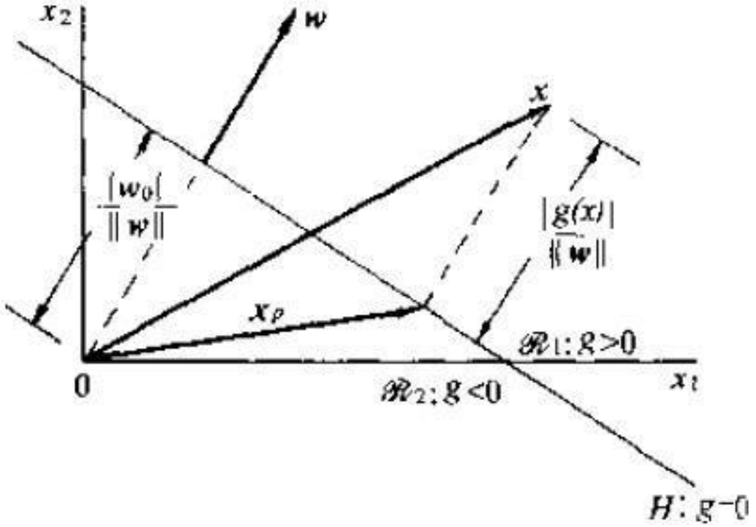
  - 定义线性判别函数
    - $g(x) = w^Tx + w_0$

1.3　线性判别函数

- **两类多维问题**
  - C = 2, D任意
  - 定义线性判别函数
    - $g(x) = w^T x + w_0$
    - w——权向量
    - w_0——阈值权

1.3 线性判别函数

- **线性判别函数的几何性质**
  - 法向量方向
  - 原点距离



1.4 增广变换

- **线性判别函数的增广变换**
  - 定义增广变换

$$g(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=1}^{d} a_i y_i = \boldsymbol{a}^T \boldsymbol{y}$$

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix}, \boldsymbol{a} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix},$$

1.4　增广变换

- **线性判别函数的增广变换**
  - 则线性判别函数为
    - $g(x) = a^T y$
    - y ——增广样本向量
    - a——增广（或广义）权向量

1.4 增广变换

- **线性判别函数的增广变换**
  - 增广变换的特点
    - 维数增加了一维：$D_G = D + 1$
    - 样本向量实际还是位于原D维子空间中
    - 样本间欧式距离保持不变
    - $a^T y = 0$ 是过原点的超平面H_G

1.5　相关概念归纳

- **概念回顾**
  - 线性判别函数，记为g(x)
  - 线性决策面，记为H
  - 线性决策面方程，令g(x) = 0

1.5 相关概念归纳

- **线性决策面法向量方向**
  - 线性决策面将特征空间分为两个区域。
    - 其中法向量方向区域称为正侧区域（简称正侧）
    - 法向量反方向的区域称为负侧区域（简称负侧）。
  - 设计时，通常使
    - 正侧对应ω_1类（甲类或A类）
    - 负侧对应ω_2类（乙类或B类）。

1.5 相关概念归纳

- **决策规则**
  - 已知判别函数
    - $g(x) = w^T x + w_0$，或$g(x) = a^T y$
  - 则决策规则为
    - 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

1.6　线性分类器设计概述

- **线性分类器的理论设计**
  - 设计线性分类器（理论上讲）
  - ↓
  - 设计决策规则
  - ↓
  - 设计线性判别函数$g(x) = w^T x + w_0$，或$g(x) = a^T y$
  - ↓
  - 求解权向量w和阈值权w_0，或增广权向量a

1.6 线性分类器设计概述

- **线性分类器设计常规步骤**
  - 给定类别已知的样本——训练样本集
  - ↓
  - 选择一个准则函数J，其值反映分类器性能（分类结果优劣）
  - ↓
  - 采用求最优解的数学方法求准则函数J的极值解，从而求得权向量w和阈值权w_0，或增广权向量a

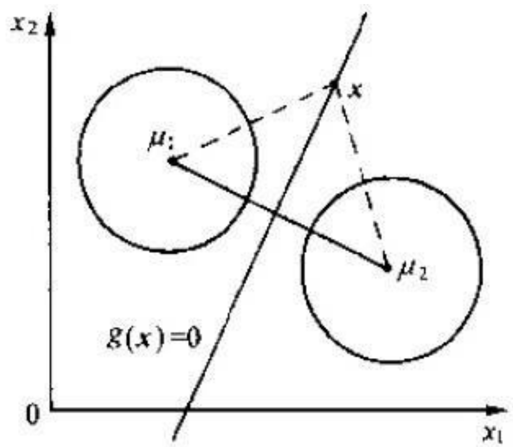
2　垂直平分分类器

- 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
    - $g(x) = (m_1 – m_2)^T x - (m_1 – m_2)^T (m_1 + m_2) / 2$
    - $= (m_1 – m_2)^T (x - (m_1 + m_2) / 2)$
  - **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

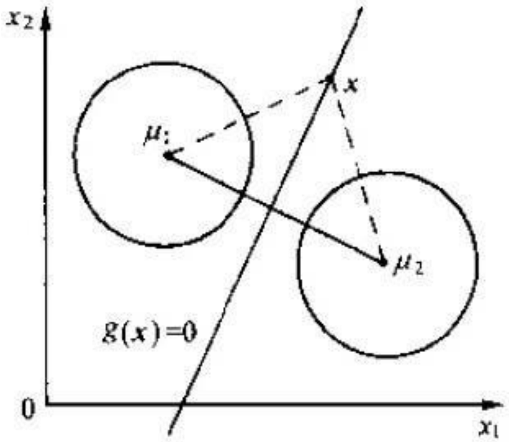
- **决策规则**
  - 已知垂直平分判别函数
    - $g(x) = (m_1 - m_2)^T (x - (m_1 + m_2) / 2)$
  - 垂直平分决策规则为
    - 对于未知样本x，若$g(x) > 0$，则x决策为ω_1类
    - 若$g(x) < 0$，则x决策为ω_2类

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

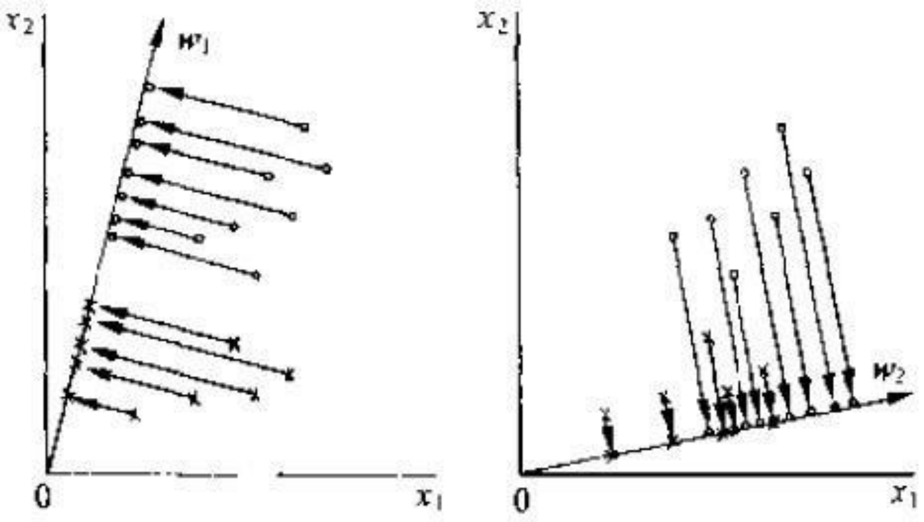
- 3.1　问题与思路
- 3.2　Fisher准则函数
- 3.3　准则函数化简
- 3.4　求极值解
- 3.5　特点
- 3.6　后续研究

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

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

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

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

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

- **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

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

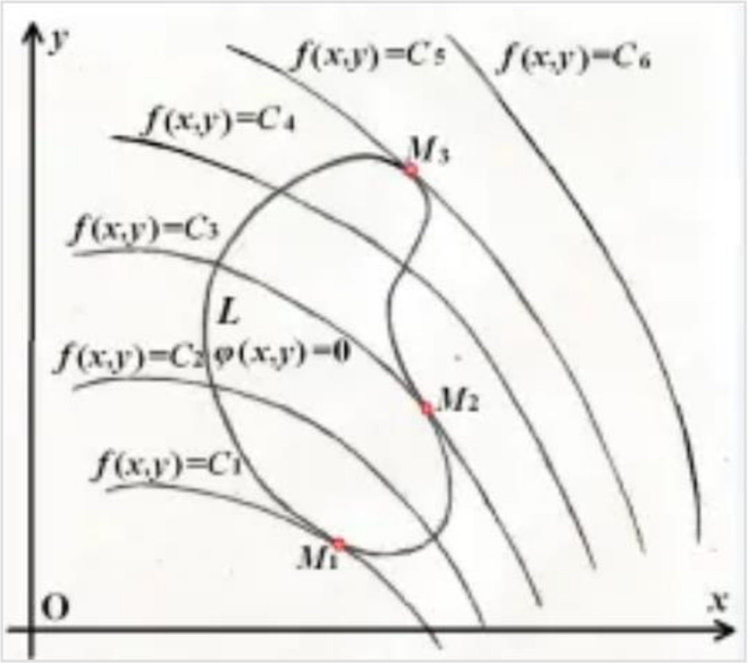
令偏导数为零，得

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

即

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

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



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

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

式中

$$R = (\boldsymbol{m}_1 - \boldsymbol{m}_2)^T w^*$$

为一标量，所以 $S_b w^*$ 总是在向量 $(\boldsymbol{m}_1 - \boldsymbol{m}_2)$ 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

$$\lambda w^* = S_w^{-1}(S_b w^*) = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)R$$

从而可得

$$w^* = \frac{R}{\lambda}S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{4-31}$$

忽略比例因子 R/λ，得

$$w^* = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{4-32}$$

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1：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

THE END !