第二章 统计判别

统计判别

■统计是什么?

- 当你买了一台电脑时，被告知三年内可以免费保修。你想过厂家凭什么这样说吗？说多了，厂家损失；说少了，失去竞争也是损失。
- 疾病传播时，如何能够通过被感染者入院前后的各种经历得到一个疾病传染方式的模型呢?
- 如何通过调查问卷来得到性别、年龄、职业、收入等各种因素与公众对某项事物如商品的态度的关系呢?

统计学（statistics）是用以收集数据，分析数据和由数据得出结论的一组概念、原则和方法。

2.1 作为统计判别问题的模式分类

- 模式识别的目的就是要确定某一个给定的模式样本属于哪一类。
- 可以通过对被识别对象的**多次**观察和测量，构成特征向量，并将其作为某一个判决规则的输入，按此规则来对样本进行分类。

如：科学竞赛平台Kaggle上面，已经有了一个COVID-19病例数据集，数据每天更新，内容包括患者年龄、患者居住地、何时出现症状、何时暴露、何时进入医院等等

2.1 作为统计判别问题的模式分类

- 在获取模式的观测值时，有些事物具有确定的因果关系，即在一定的条件下，它必然会发生或必然不发生。
  - 例如识别一块模板是不是直角三角形，只要凭“三条直线边闭合连线和一个直角”这个特征，测量它是否有三条直线边的闭合连线并有一个直角，就完全可以确定它是不是直角三角形。
  - 这种现象是**确定性**的现象

2.1 作为统计判别问题的模式分类

- 但在现实世界中，由许多客观现象的发生，就每一次观察和测量来说，即使在基本条件保持不变的情况下也具有不确定性。
- 只有在**大量重复**的观察下，其结果才能呈现出某种规律性，即对它们观察到的特征具有统计特性。
- 特征的值不再是一个确定的向量，而是一个**随机向量**。
- 此时，只能利用模式集的统计特性来分类，以使分类器发生错误的概率最小。

2.1 作为统计判别问题的模式分类

- 给定观测值x,判断其属于 ω_1 类还是 ω_2，作出某次判断时的错误率是：

$$P(error\,|\,\boldsymbol{x})=\begin{cases}P(\omega_1\,|\,\boldsymbol{x}),\boldsymbol{x}\in\omega_2\\ P(\omega_2\,|\,\boldsymbol{x}),\boldsymbol{x}\in\omega_1\end{cases}$$

结论：最小化误差概率条件下，决策规则为：

$P(\omega_1\,|\,\boldsymbol{x})>P(\omega_2\,|\,\boldsymbol{x})$则差别 $\boldsymbol{x}\in\omega_1$；否则判别 $\boldsymbol{x}\in\omega_2$

2.1 作为统计判别问题的模式分类

2.1.1 贝叶斯判别原则

- 两类模式集的分类

  •目的：要确定$\boldsymbol{x}$是属于ω_1类还是ω_2类，要看$\boldsymbol{x}$是来自于ω_1类的概率大还是来自ω_2类的概率大。

- 贝叶斯判别

$$P(\omega_1 \mid \boldsymbol{x}) \quad ? \quad P(\omega_2 \mid \boldsymbol{x}),$$

2.1 作为统计判别问题的模式分类

2.1.1 贝叶斯判别原则

■例子（地震预测）

•对某一地震高发区进行统计，地震以ω_1类表示，正常以ω_2类代表

•统计的时间区间内， 每周发生地震的概率为20%，即$P(\omega_1)=0.2$，当然$P(\omega_2)=1-0.2=0.8$

•在任意一周，要判断该地区是否会有地震发生。显然，因为$P(\omega_2)> P(\omega_1)$，只能说是正常的可能性大。如要进行判断，只能用其它观察现象来实现。

2.1 作为统计判别问题的模式分类

2.1.1 贝叶斯判别原则

■例子

- 通常地震与生物异常反应之间有一定的联系。
- 若用生物是否有异常反应这一观察现象来对地震进行预测，生物是否异常这一结果以模式x代表，这里x为一维特征，且只有x = "异常"和x = "正常"两种结果。

2.1 作为统计判别问题的模式分类

2.1.1 贝叶斯判别原则

■例子

- 假设根据观测记录，发现这种方法有以下统计结果：

  地震前一周内出现生物异常反应的概率=0.6，即$P(\boldsymbol{x}=\text{异常}|\ \omega_1)=0.6$

  地震前一周内出现生物正常反应的概率=0.4，即$P(\boldsymbol{x}=\text{正常}|\ \omega_1)=0.4$

  一周内没有发生地震但也出现了生物异常的概率=0.1，

  即$P(\boldsymbol{x}=\text{异常}|\ \omega_2)=0.1$

  一周内没有发生地震时，生物正常的概率=0.9，即$P(\boldsymbol{x}=\text{正常}|\ \omega_2)=0.9$

2.1 作为统计判别问题的模式分类

$P(\boldsymbol{x}=\text{异常}|\ \omega_1)=0.6$
$P(\boldsymbol{x}=\text{正常}|\ \omega_1)=0.4$
$P(\boldsymbol{x}=\text{异常}|\ \omega_2)=0.1$
$P(\boldsymbol{x}=\text{正常}|\ \omega_2)=0.9$

$P(\omega_1)=0.2$

2.1.1 贝叶斯判别原则

■问题

•若某日观察到明显的生物异常反应现象，一周内发生地震的概率为多少，即求$P(\omega_1\,|\,\boldsymbol{x}=\text{异常})=?$

•这里$P(\omega_1)$ 是根据以往的统计资料得到的地震发生的先验概率。现在经过观察，需要求出$P(\omega_1\,|\,\boldsymbol{x}=\text{异常})$，即观测到生物异常时一周内发生地震的概率，称为后验概率。

■[计算]

2.1 作为统计判别问题的模式分类

2.1.1 贝叶斯判别的推广

- **允许使用多于一个特征**：标量、向量、多种特征向量
- **允许多于两种类别状态的情形**
- **允许有其他行为而不是仅仅是判定类别**：如后验概率接近的情况下，拒绝做判决，如果因此付出的代价不大的话。
- **通过引入一个更一般的损失函数来替代误差概率**：损失函数精确阐述了每种行为所付出的代价的大小，分类错误代价相等是最简单的情况。

2.1 作为统计判别问题的模式分类

2.1.2 朴素贝叶斯

- 在特征 $\boldsymbol{x}=(x_1,x_2,x_3,\cdots,x_D)$ 是多维向量时，朴素贝叶斯算法是假设各个特征之间相互独立。

$$p(x_1,x_2,x_3,\cdots,x_D\,|\,\omega)=\prod_{i=1}^{D}p(x_i\,|\,\omega)$$

- 应用例子：社区账号的真实性检测。

- C=0表示真实账号，C=1表示不真实账号。
- 选择三个特征属性：x_1：日志数量/注册天数，x_2：好友数量/注册天数，x_3：是否使用真实头像。
- 下面给出划分：x_1：{x_1<=0.05, 0.05< x_1 <0.2, x_1 >=0.2}，x_2：{x_2<=0.1, 0.1< x_2 <0.8, x_2 >=0.8}，x_3：{x_3=0（不是），x_3=1（是）}。

2.1 作为统计判别问题的模式分类

2.1.2 朴素贝叶斯

$P(C=0)=0.89,\ P(C=1)=0.11$

$P(x_1<=0.05\,|\,C=0)=0.3$

$P(0.05<x_1<0.2\,|\,C=0)=0.5$

$P(x_1>=0.2\,|\,C=0)=0.2$

$P(x_1<=0.05\,|\,C=1)=0.8$

$P(0.05<x_1<0.2\,|\,C=1)=0.1$

$P(x_1>=0.2\,|\,C=1)=0.1$

$P(x_2<=0.1\,|\,C=0)=0.1$

$P(0.1<x_2<0.8\,|\,C=0)=0.7$

$P(x_2>=0.8\,|\,C=0)=0.1$

$P(x_2<=0.1\,|\,C=1)=0.7$

$P(0.1<x_2<0.8\,|\,C=1)=0.2$

$P(x_2>=0.8\,|\,C=1)=0.1$

$P(x_3=0\,|\,C=0)=0.2$

$P(x_3=1\,|\,C=0)=0.8$

$P(x_3=0\,|\,C=1)=0.9$

$P(x_3=1\,|\,C=1)=0.1$

某个账号使用非真实头像，日志数量与注册天数的比率为0.1，好友数与注册天数的比率为0.2。请问这个账号是否为真实账号？

- C=0表示真实账号，C=1表示不真实账号。

2.1 作为统计判别问题的模式分类

2.1.2 朴素贝叶斯

$$P(x_1, x_2, x_3 \mid C) = P(x_1 \mid C)P(x_2 \mid C)P(x_3 \mid C)$$

$$x_1 = 0.1, x_2 = 0.2, x_3 = 0$$

$$P(C=0 \mid x_1 = 0.1, x_2 = 0.2, x_3 = 0) = \frac{P(x_1 = 0.1, x_2 = 0.2, x_3 = 0 \mid C=0)P(C=0)}{P(x_1 = 0.1, x_2 = 0.2, x_3 = 0)}$$

$$P(C=1 \mid x_1 = 0.1, x_2 = 0.2, x_3 = 0) = \frac{P(x_1 = 0.1, x_2 = 0.2, x_3 = 0 \mid C=1)P(C=1)}{P(x_1 = 0.1, x_2 = 0.2, x_3 = 0)}$$

$$P(x_1 = 0.1, x_2 = 0.2, x_3 = 0 \mid C=0)P(C=0) = P(x_1 = 0.1 \mid C=0)P(x_2 = 0.2 \mid C=0)P(x_3 = 0 \mid C=0)$$

$$= 0.5 \times 0.7 \times 0.2 \times 0.89 = 0.0623$$

$$P(x_1 = 0.1, x_2 = 0.2, x_3 = 0 \mid C=1)P(C=1) = 0.1 \times 0.7 \times 0.9 \times 0.11 = 0.00693$$

C=0, 该账号为真实账号。

- C=0表示真实账号，C=1表示不真实账号。

2.1 作为统计判别问题的模式分类

2.1.3 贝叶斯最小风险判别

■意义

•对于自然属性是属于ω_i类的模式$\boldsymbol{x}$来说，它来自ω_i类的概率应为$P(\omega_i \mid \boldsymbol{x})$。

•如果分类器判别$\boldsymbol{x}$是属于ω_j类，但它实际上来自ω_i类，也就是说分类器失败，这时L_{ij}为失分，对应的条件风险为后验概率进行L_{ij}的加权运算。

•由于模式$\boldsymbol{x}$的自然属性可能来自M类中的任一类，因此可将观察样本指定为ω_j类的条件平均风险用$r_j(\boldsymbol{x})$的公式运算。

2.1 作为统计判别问题的模式分类

2.1.3 贝叶斯最小风险判别

- 当考虑到对于某一类的错误判决要比对另一类的判决更为关键时，就需要把最小错误概率的贝叶斯判别做一些修正，提出条件平均风险$r_j(\boldsymbol{x})$。
- M类分类问题的条件平均风险$r_j(\boldsymbol{x})$
  - •对M类问题，如果观察样本被判定属于ω_j类，则条件平均风险为：

$$\boxed{r_j = \sum_{i=1}^{M} L_{ij} P(\omega_i \mid x)}$$

  - •L_{ij}称为将本应属于ω_i类的模式判别成属于ω_j类的是非代价。

2.1 作为统计判别问题的模式分类

2.1.3 贝叶斯最小风险判别

■ L_{ij}的取值

- 若$i=j$，即判别正确，得分，L_{ij}可以取负值或零，表示不失分。
- 若$i \neq j$，即判别错误，失分，L_{ij}应取正值。

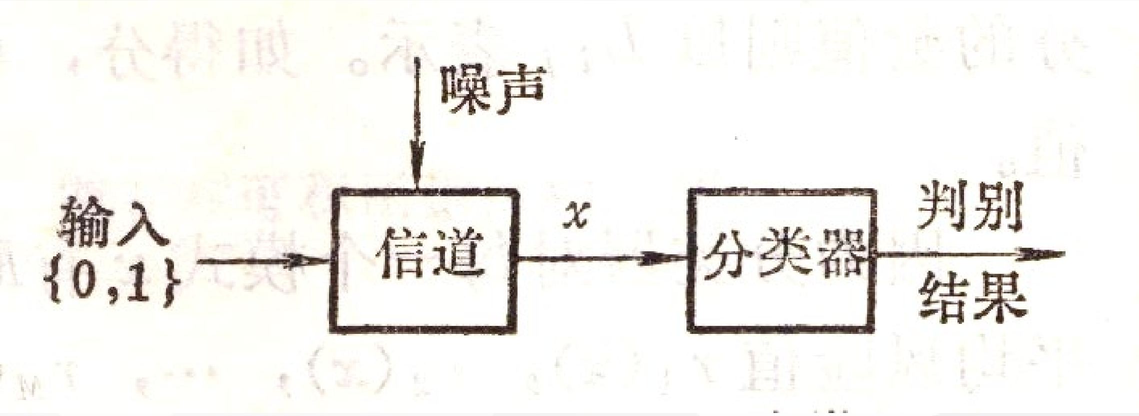
■ 最小平均条件风险分类器

- 分类器对每一个模式$\boldsymbol{x}$有M种可能的类别可供选择。
- 若对每一个$\boldsymbol{x}$计算出全部类别的平均风险值$r_1(x)$, $r_2(\boldsymbol{x})$,…, $r_M(\boldsymbol{x})$，并且将$\boldsymbol{x}$指定为是具有最小风险值的那一类，则这种分类器称为最小平均条件风险分类器。
- 表达式

2.1 作为统计判别问题的模式分类

2.1.3 贝叶斯最小风险判别

- 两类（M=2）的情况
- [例子]



- 一般多类（M类）的情况

作业（可选）

- 结合生活中的例子，出一道用贝叶斯判别及贝叶斯最小风险判别求解的题目。

2.2 正态分布模式的贝叶斯分类器

■出发点

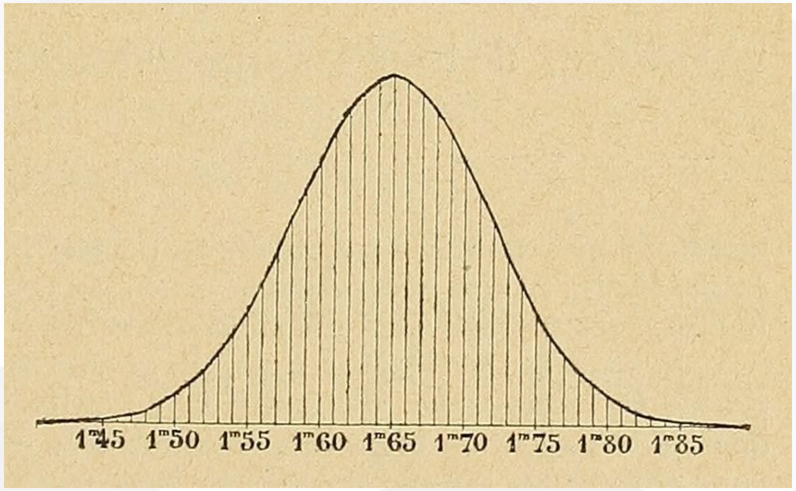
•当已知或者有理由设想类概率密度函数$p(x|\omega_i)$是多变量的正态分布时，上一节介绍的贝叶斯分类器可以导出一些简单的判别函数。

•由于正态密度函数易于分析，且对许多重要的实际应用又是一种合适的模型，因此受到很大的重视。

2.2 正态分布模式的贝叶斯分类器

■ 下面列出的变量的分布都比较接近正态分布：

- 人群的身高
- 成年人的血压
- 传播中的粒子的位置
- 测量误差
- 回归中的残差
- 人群的鞋码
- 一天中雇员回家的总耗时
- 教育指标



1893年人类身高分布图，作者：Alphonse Bertillon

2.2 正态分布模式的贝叶斯分类器

M种模式类别的多变量正态类密度函数

•判别函数是一个二次曲面。

$$d_i(\boldsymbol{x}) = \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{C}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)\boldsymbol{C}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i), i = 1, 2, \cdots, M$$

•对于正态分布模式的贝叶斯分类器，两个模式类别之间用一个二次判别界面分开，就可以求得最优的分类效果。

两类问题且其类模式都是正态分布的特殊情况

•当$C_1 \neq C_2$时的情况

显然，判别界面$d_1(\boldsymbol{x})$- $d_2(\boldsymbol{x})$=0是$\boldsymbol{x}$的二次型方程，即ω_1和ω_2两类模式可用二次判别界面分开。

当$\boldsymbol{x}$是二维模式时，判别界面为二次曲线，如椭圆、圆、抛物线或双曲线等。

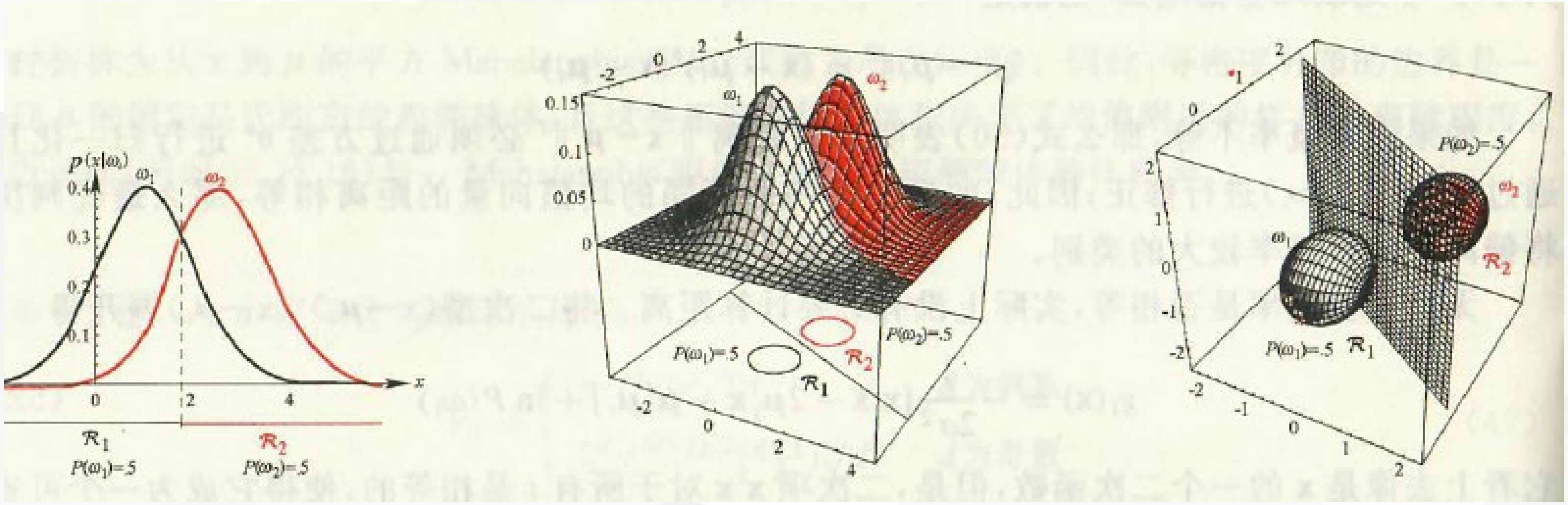
$$d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) = \ln P(\omega_1) - \ln P(\omega_2) + (\boldsymbol{m}_1 - \boldsymbol{m}_2)^T \boldsymbol{C}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{m}_1^T\boldsymbol{C}^{-1}\boldsymbol{m}_1 + \frac{1}{2}\boldsymbol{m}_2^T\boldsymbol{C}^{-1}\boldsymbol{m}_2 = 0$$

•当$C_1 = C_2 = C$时的情况

判别界面为$\boldsymbol{x}$的线性函数，为一超平面。

当$\boldsymbol{x}$是二维时，判别界面为一直线。

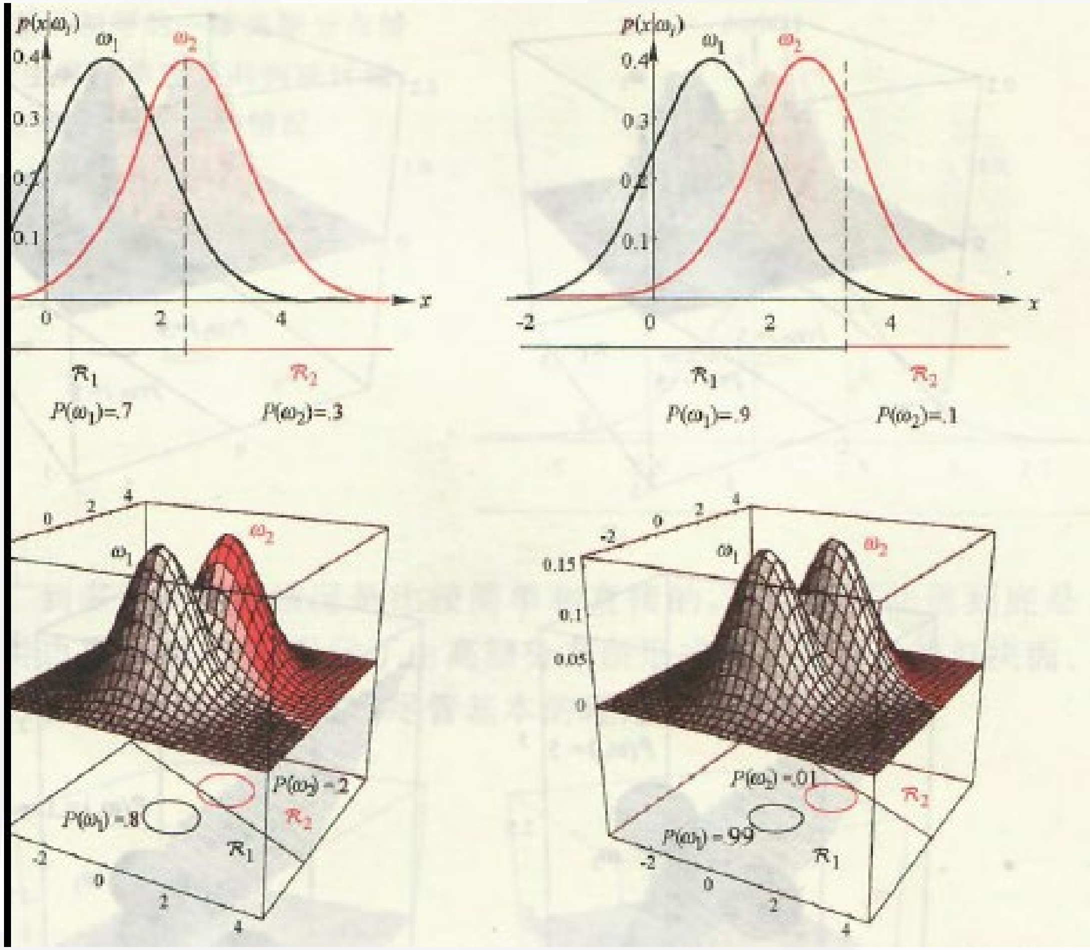
2.2 正态分布模式的贝叶斯分类器

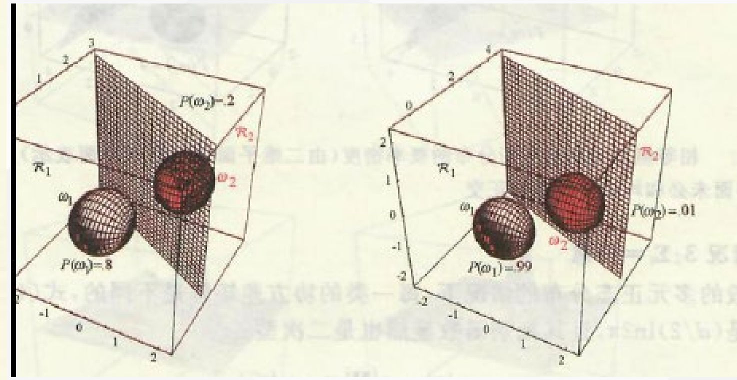


如果两种分布的协方差矩阵相等并且与单位阵成比例，且先验概率相等。则决策边界垂直于两个中心的连线。

$$\boxed{\begin{aligned} d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) = \ln P(\omega_1) - \ln P(\omega_2) + (\boldsymbol{m}_1 - \boldsymbol{m}_2)^T \boldsymbol{C}^{-1}\boldsymbol{x} - \\ \frac{1}{2}\boldsymbol{m}_1^T\boldsymbol{C}^{-1}\boldsymbol{m}_1 + \frac{1}{2}\boldsymbol{m}_2^T\boldsymbol{C}^{-1}\boldsymbol{m}_2 = 0 \end{aligned}}$$

2.2 正态分布模式的贝叶斯分类器





协方差矩阵相等，判决边界同样是超平面。随着先验概率的改变，判决边界也随之改变；对于差别较大的离散先验概率而言，判决边界不会落于中心点之间。

作业及编程（编程可选）

- 设以下模式类别具有正态概率密度函数：

ω_1： $\{(0\ 0)^T, (2\ 0)^T, (2\ 2)^T, (0\ 2)^T\}$

ω_2： $\{(4\ 4)^T, (6\ 4)^T, (6\ 6)^T, (4\ 6)^T\}$

（1）设$P(\omega_1)= P(\omega_2)=1/2$，求这两类模式之间的贝叶斯判别界面的方程式。

（2）绘出判别界面。

- 编写两类正态分布模式的贝叶斯分类程序。（可选例题或上述作业题为分类模式）

2.3 均值向量和协方差矩阵的参数估计

- 在贝叶斯分类器中，构造分类器需要知道类概率密度函数$p(\mathbf{x}|\omega_i)$。
- 如果按先验知识已知其分布，则只需知道分布的参数即可。
  - 例如：类概率密度是正态分布，它完全由其均值向量和协方差矩阵所确定。
- 对均值向量和协方差矩阵的估计即为贝叶斯分类器中的一种参数估计问题。

2.3 均值向量和协方差矩阵的参数估计

■参数估计的两种方式

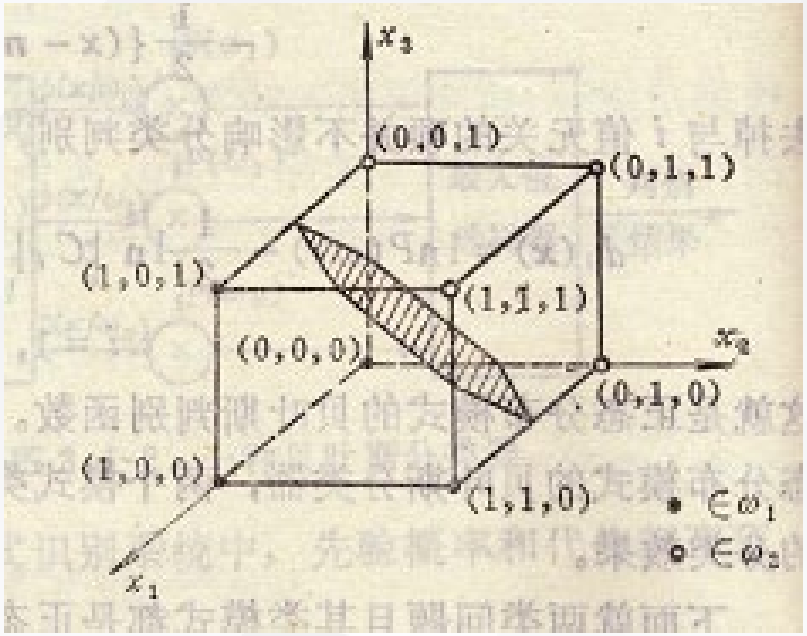
- 一种是将参数作为非随机变量来处理，例如矩估计就是一种非随机参数的估计。
- 另一种是随机参数的估计，即把这些参数看成是随机变量，例如贝叶斯参数估计。

2.3 均值向量和协方差矩阵的参数估计

- 均值和协方差矩阵的非随机参数的估计
  - 均值和协方差矩阵的估计量定义
  - 均值和协方差矩阵估计量的迭代运算

2.3 正态分布模式的贝叶斯分类器

■ [例子]



■ 讨论

- 贝叶斯分类规则是基于统计概念的。
- 如果只有少数模式样本，一般较难获得最优的结果。

2.3 均值向量和协方差矩阵的参数估计

■均值向量和协方差矩阵的贝叶斯学习

- 将概率密度函数的参数估计量看成是随机变量θ，它可以是标量、向量或矩阵。
- 按这些估计量统计特性的先验知识，可以先粗略地预选出它们的密度函数。
- 通过训练模式样本集$\{x^i\}$，利用贝叶斯公式设计一个迭代运算过程求出参数的后验概率密度$p(\theta|x^i)$。
- 当后验概率密度函数中的随机变量θ的确定性提高时，可获得较准确的估计量。

2.3 均值向量和协方差矩阵的参数估计

■均值向量和协方差矩阵的贝叶斯学习

- 一般概念
- 单变量正态密度函数的均值学习

2.3 均值向量和协方差矩阵的参数估计

贝叶斯参数估计

贝叶斯估计的基本步骤（基于平方误差损失函数）

1、 确定参数 θ 的先验分布 p(θ);

2、 由样本集 $D=\{x_1,x_2,\cdots x_N\}$ 求出样本联合分布 $p(D|\theta)$，它是 θ 的函数：

$$p(D|\theta)=\prod_{n=1}^{N}p(x_n|\theta)$$

3、 利用贝叶斯公式，求 θ 的后验分布：

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{\int_{\Theta}p(D|\theta)p(\theta)d\theta}$$

4、 求出贝叶斯估计值：

$$\hat{\theta}=\int_{\Theta}\theta p(\theta|D)d\theta$$

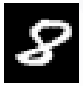
应用

- 手写数字识别
- Mnist:

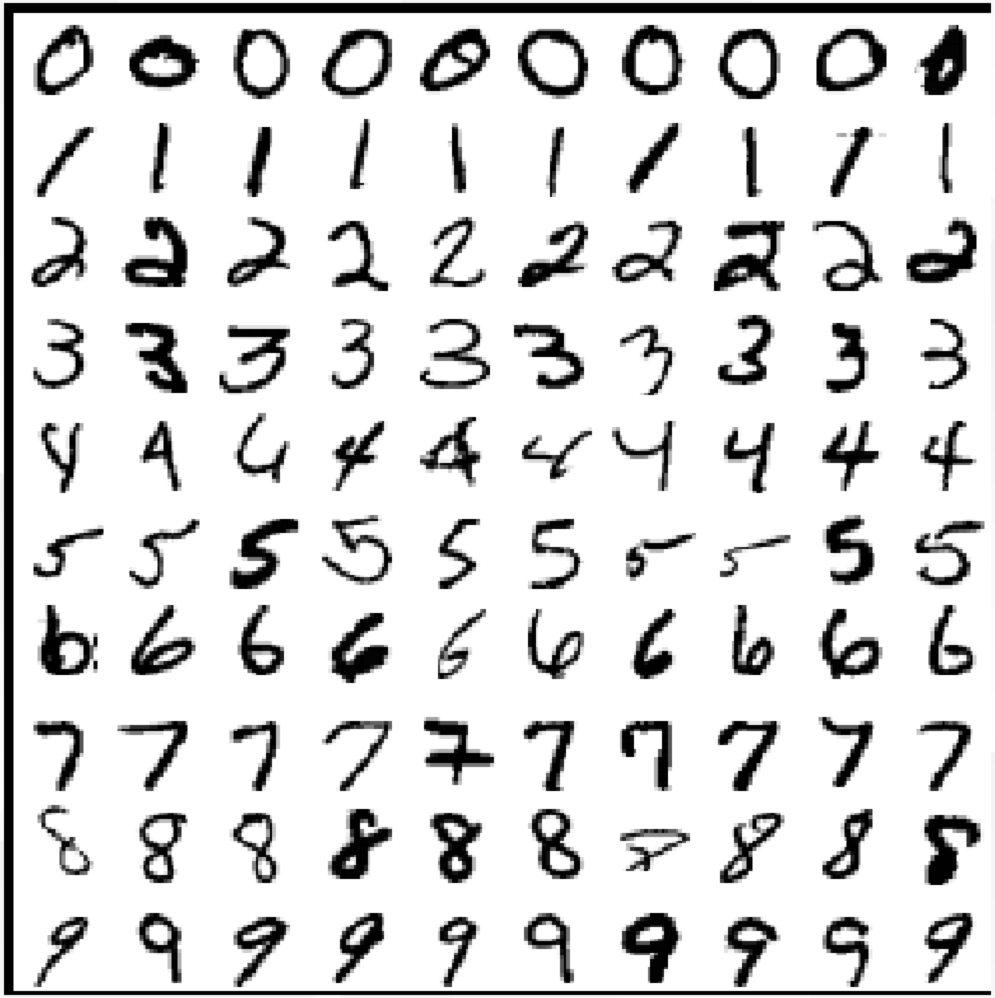
  每个数字为

  784（28*28）

  维的向量



$$p(y=j\,|\,x_1,x_2,\cdots,x_{784})$$

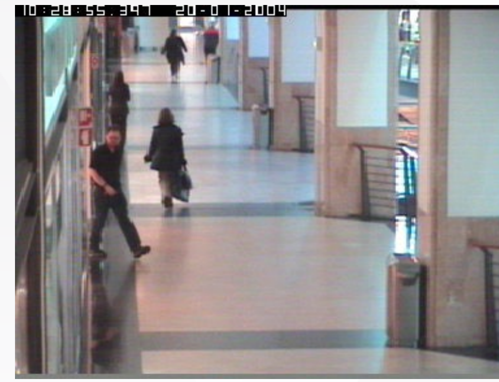


应用

■视频中的单目标跟踪

第一帧

第二帧

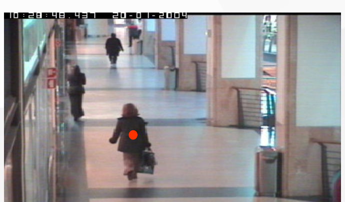


第十帧

- 均值和协方差矩阵的非随机参数的估计
- 贝叶斯判别

应用

■视频中的单目标跟踪



数据分布模型
（正态分布）

模型更新

贝叶斯判别

应用

■视频中的单目标跟踪