

第一章概念：模式：指测量且需要分类的对象。模式类：把模式所属的类别或同一类别中模式的总体称为模式类（或简称类）。**样本**：所研究对象的一个个体。**样本集**：若干样本的集合。**特征**：指用于表征样本的观测。通常是数值表示的某些量化特征。如果存在多个特征，则它们就组成了特征向量。样本的特征组成了样本的特征空间。空间的维数就是特征的个数，而每个样本是特征空间中的一个点。某些情况下，对样本的原始描述可能是非数值形式的，需要采用一定的方法把这些特征转换成数值形式。**模式的基本属性**：{可观测性：待处理模式可由某种类型的传感器装置获取，这样的传感器装置可以是基于某种物理效应的传感器，也可以是基于某种化学效应或生物效应的传感器，甚至可以是某种数学模型的虚拟传感器。可区分性：不同模式类的观测样本之间应该具有可区分的特征。相似性：同一模式类的观测样本之间应该具有某种相似的特征。} **识别**：再认知，把对象分门别类的认出来。**模式识别**：用计算的方法根据样本的特征将样本划分到一定的类别中去。侧重于对人的认知行为进行模仿，把人的知识和经验转化为可以为机器所利用的一些规则和方法，赋予机器对被观测事物进行综合分析和自动分类的能力，使机器可以根据被观测事物往日的观测样本形成相应的分类规则并据此完成对新的观测样本进行分类的任务。难点：样本中不仅包含待识别事物的固有信息，也包含有害的环境信息。**模式采集**：模式的数字化表达。若干问题：1) 紧迫性 2) 特征选取 3) 相似性度量与分类 4) 性能评价。

(1) 相似性度量
测度函数-距离函数
在特征空间中用特征向量描述样本的属性，就是把相似性度量用距离度量（距离函数）表示。
定义紧迫性。紧迫性的性质：
(1) 临界点的数量与总的点数相比很少。
(2) 集合中任意两个内点可以用光滑曲线连接，在该连线上的点也属于这个集合。
(3) 每个内点都有一个足够大的邻域，在该领域中只包含同一集合中的点。



模式识别是一个过程，它将现实世界中的被测对象通过一系列的变换和处理映射到符号世界中被测对象的分类和描述。
预处理：提高输入模式的质量
✓ 去噪声：消除或减少模式采集中的噪声及其他干扰，提高信噪比
✓ 去模糊：消除或减少数据图像模糊（包括运动模糊）及几何失真，提高清晰度
✓ 模式结构转换：例如把非线性模式转变为线性模式，以利于后续处理等。

预处理的方法：滤波、变换、图像增强、图像恢复等
训练方法
训练是依据特征空间的分布，决定分类器的具体参数。分界线的方式一般由设计者决定，如可用直线、折线或曲线作为类别的分界线。
分界线的样式及其具体参数也可利用训练样本经训练过程确定。

特征提取与选择
✓ **度量特征**：长度、灰度、重量
✓ **属性特征**：性别、文化程度
✓ **基元特征**：形成待识别对象有效描述的一组度量或属性参数
✓ **基元特征**：形成待识别对象有效描述的无需再分割的基元子模式
✓ **基元特征**：边缘、轮廓、纹理、区域

分类决策
对待分类样本进行分类识别的过程
✓ 把未知类别属性的样本确定为类空间中的某一类型。
✓ 分类错误率越小越好。
✓ 分类错误率的分析与计算比较困难

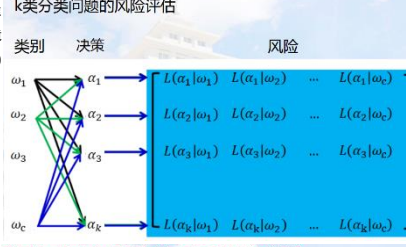
影响分类错误率的因素：
✓ 样本质量
✓ 提取的特征
✓ 分类方法
✓ 分类器设计
模式识别
- 特征向量表达：几何方法
- 统计模式识别：概率方法
- 结构表达：参数方法
- 有监督的模式识别：非参数方法
- 无监督的模式识别：各种聚类方法

【基于最小错误率的贝叶斯决策】
已知条件
类别：先验概率 $P(\omega_1), P(\omega_2)$ ，类条件概率密度 $p(X|\omega_1), p(X|\omega_2)$
问题：设计一个具有最小误分率的二类分类器
求解：设观测样本为 X ，则 X 的全概率密度为 $p(X) = P(\omega_1)p(X|\omega_1) + P(\omega_2)p(X|\omega_2)$
 X 属于两个类别的后验概率为 $P(\omega_1|X) = \frac{p(X|\omega_1)P(\omega_1)}{p(X)}$, $P(\omega_2|X) = \frac{p(X|\omega_2)P(\omega_2)}{p(X)}$
最大后验概率判决

概念：在模式分类问题中，人们往往希望尽量减少分类错误，使错误率最小的分类规则，称为基于最小错误率的贝叶斯决策。

最小错误率的贝叶斯准则：
(1) $P(\omega_1|X) = \max_{j=1,2} P(\omega_j|X)$ $\rightarrow X \in \omega_1$
(2) $P(\omega_1|X)P(\omega_1) > P(\omega_2|X)P(\omega_2)$ $\rightarrow X \in \omega_1$
(3) $l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$ 其中 $l(x)$ 为似然比。
(4) $h(x) = -\ln l(x) = -\ln p(x|\omega_1) + \ln p(x|\omega_2) < \ln \frac{P(\omega_2)}{P(\omega_1)}$ $\rightarrow X \in \omega_1$
 $h(x) = -\ln l(x) = -\ln p(x|\omega_1) + \ln p(x|\omega_2) > \ln \frac{P(\omega_2)}{P(\omega_1)}$ $\rightarrow X \in \omega_2$

可以证明，最大后验概率判决规则给出的判决结果的错误率是最小的



【基于最小风险的贝叶斯决策】
算法步骤：
第1步：已知类条件概率密度 $p(x|\omega_j)$ 和先验概率 $P(\omega_j)$, $j=1,2,\dots,c$ 以及给出待识别的 x 的情况下，根据贝叶斯公式计算后验概率：
 $P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{\sum_{j=1}^c p(x|\omega_j)P(\omega_j)}$, $j=1,2,\dots,c$
第2步：利用第1步计算的后验概率及决策损失表，计算采取 $\alpha_i, i=1,2,\dots,k$ 的条件风险 $R(\alpha_i|x)$ ：
 $R(\alpha_i|x) = \sum_{j=1}^c L(\alpha_i|\omega_j)P(\omega_j|x)$
第3步：对第2步得到的 k 个条件风险 $R(\alpha_i|x), i=1,2,\dots,k$ 进行比较，找出使条件风险最小的决策 α_j ，即
 $R(\alpha_j|x) = \min_k R(\alpha_k|x)$

多元正态分布下的最小错误率贝叶斯判别函数和决策面
1. 问题 正态分布下贝叶斯分类器设计
2. 多类判别函数 $g(x) = \ln p(x|\omega_1) + \ln P(\omega_1)$
3. 正态分布函数 $p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$
4. 判别函数 $g(x) = -\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_1| + \ln P(\omega_1)$
5. 决策面方程 $g(x) = g_j(x)$
即 $-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - \frac{1}{2} \ln |\Sigma_1| + \ln P(\omega_1) = 0$

参数估计分类
(1) **监督参数估计**：样本类别、类条件概率密度函数已知，但表征概率密度函数的某些参数未知。例如，只知道样本所属样本总体为正态分布，而正态分布的参数是未知的，参数估计的目的就是在样本类别标识已知的情况下，估计出这些参数。
(2) **非监督参数估计**：概率密度函数形式已知，样本类别标识未知。
(3) **非参数估计**：样本类别已知，但概率密度函数形式未知。此时，不能仅仅估计几个参数，而是用样本把概率密度函数数值化的估计出来。
(1) **统计量**：样本中包含总体的信息，希望通过样本集把有关信息抽取出来，针对不同要求构造出样本的某种函数，这种函数在统计学中称为统计量。
(2) **参数空间**：在参数估计中，假设样本的概率密度函数的形式已知，但其中参数是未知的，将这些参数的全部可容许值组成的集合称为参数空间。
(3) **点估计、估计量、估计值**：点估计问题就是构造一个统计量 $d(x_1, \dots, x_n)$ 作为参数 θ 的估计 $\hat{\theta}$ ，在统计学中称为 θ 的估计量；针对某一具体类别的样本估计出估计量的具体数值，该数值称为参数的估计值。
令 $l(\theta)$ 为样本集 X 的似然函数， $X = \{x_1, x_2, \dots, x_n\}$ 如果 $\hat{\theta} = d(X) = d(x_1, x_2, \dots, x_n)$ 是参数空间 Θ 中能似然函数 $l(\theta)$ 极大化的 θ 值，那么 $\hat{\theta}$ 是 θ 的极大似然估计量，记作 $\hat{\theta} = \arg \max_{\theta} l(\theta)$

贝叶斯估计原理
求解定理
如果损失函数为二次函数，即 $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ ，则 θ 的贝叶斯估计量 $\hat{\theta}$ 是在给定 x 时 θ 的条件期望，即
 $\hat{\theta} = E[\theta|x] = \int_{-\infty}^{\infty} \theta p(\theta|x) d\theta$
Input: $(N, C_{out}, H_{out}, W_{out})$ 或 (C_{in}, H_{in}, W_{in})
Output: $(N, C_{out}, H_{out}, W_{out})$ 或 $(C_{out}, H_{out}, W_{out})$, where
 $H_{out} = \begin{cases} H_{in} + 2 \times \text{padding}[0] - \text{dilation}[0] \times (\text{kernel_size}[0] - 1) - 1 \\ \text{stride}[0] \end{cases}$
 $W_{out} = \begin{cases} W_{in} + 2 \times \text{padding}[1] - \text{dilation}[1] \times (\text{kernel_size}[1] - 1) - 1 \\ \text{stride}[1] \end{cases}$

【贝叶斯估计】

求解步骤
1. 确定 θ 的先验分布 $p(\theta)$
2. 由样本集 $X = \{x_1, x_2, \dots, x_n\}$ 求出样本联合分布 $p(X|\theta)$ ，它是 θ 的函数
3. 利用贝叶斯公式，求出 θ 的后验分布
 $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_{\Theta} p(X|\theta)p(\theta)d\theta}$
4. 利用求解定理，求出贝叶斯估计量
 $\hat{\theta} = \int_{\Theta} \theta p(\theta|X) d\theta$

设 $p(\theta|X^0) = p(\theta)$ 无样本条件下概率密度估计，即先验概率密度。反复使用上式，可得到一系列对概率密度函数参数的估计 $p(\theta), p(\theta|x_1), p(\theta|x_1, x_2), \dots$
这称为参数估计的递推贝叶斯方法。如果随着样本数的增加，这个后验概率序列逐渐尖锐，逐步趋向于以 θ 的真实值为中心的一个尖峰，当样本无穷多时收敛于在参数真实值上的脉冲函数，就把这一过程称为贝叶斯学习。

如何完成非参数估计?
令 R 是包含样本 x 的一个区域，其体积为 V ，设有 n 个训练样本，其中有 k 个落在区域 R 中，则可对概率密度作出一个估计：
 $p(x) \approx \frac{k/n}{V}$
相当于用 R 区域内的平均性质来作为一点 x 的估计，是一种数据的平滑。
当 n 固定时， V 的大小对估计的效果影响很大，过大则平滑过多，不够精确；过小则可能导致在此区域内无样本点， $k=0$ 。
此方法的有效性取决于样本数量的多少，以及区域体积选择的合适。

$g(x) = c_0 + c_1x + c_2x^2$
即 $g(x) = a^T y$
其中 $y = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$, $a = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$
将 $g(x) = a^T y$ 称为广义线性判别函数
 a 为广义权向量

【Fisher线性判别分析】
假设集合 X 包含 N 个 d 维样本 x_1, x_2, \dots, x_N ，其中 N_1 和 N_2 分别表示 ω_1 和 ω_2 的样本子集 X_1 和 X_2 的个数。
在直线上的投影为 $y_j = w^T x_j, j=1,2,\dots,N$
(a) 在原始 d 维 X 样本空间中：
(1) 各类样本的均值向量 $m_1 = \frac{1}{N_1} \sum_{x \in X_1} x$
(2) 样本类内离散度矩阵 $S_1 = \sum_{x \in X_1} (x - m_1)(x - m_1)^T, j=1,2$
 $S_2 = S_1 + S_2$
(3) 类间离散度矩阵 $S_b = (m_1 - m_2)(m_1 - m_2)^T$

决定的。由于我们只关心 w 的方向，因此可以取 $w^* = S_b^{-1}(m_1 - m_2)$
这就是Fisher判别准则下的最优投影方向。阈值可以取为：
 $w_0 = -\frac{1}{2}(m_1 + m_2)^T S_b^{-1}(m_1 - m_2) - \ln \frac{P(\omega_2)}{P(\omega_1)}$
若 $g(x) = w^T x + w_0 \geq 0$ ，则 $x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

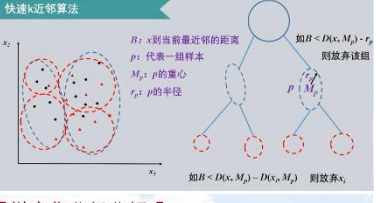
【感知器】
(2) 线性可分性
如果有一个线性机器能把每个样本正确分类，即如果存在一个权向量 a 使得对于任何 $x \in \omega_1$ 都有 $a^T x > 0$ ，对于任何 $x \in \omega_2$ 都有 $a^T x < 0$ ，则称这组样本集为线性可分，否则为线性不可分。
如果样本集 x_1, x_2, \dots, x_n 是线性可分的，则必存在某个权向量 a ，使得 $\begin{cases} a^T x_i > 0 & \text{对于 } -\omega_1 \in \omega_1 \\ a^T x_i < 0 & \text{对于 } -\omega_1 \in \omega_2 \end{cases}$
称 $x_i = \begin{cases} x_i & \text{对于 } -\omega_1 \in \omega_1 \\ -x_i & \text{对于 } -\omega_1 \in \omega_2 \end{cases}$
为规范化推广样本向量。
感知器算法：
1. 任意选择初始的权向量 a_1 ，设 $k=1$
2. 考察样本 x_i ，若错误分类，修正 a_k 如 $a_k^T x_i \leq 0$ $a_{k+1} = a_k + \eta x_i$
若正确分类， a_k 不修正
如 $a_k^T x_i > 0$ $a_{k+1} = a_k$
3. 考察另一个样本，重复第2步，直至所有样本都分类正确，即 $J_k = 0$

例题：有两类样本
 $\omega_1 = \{(x_1, x_2) = \{(1,0), (0,1)\}\}$
 $\omega_2 = \{(x_1, x_2) = \{(1,1), (0,0)\}\}$
解：先将四个样本规范化
 $y_1 = (1,0,1), y_2 = (1,0,1,1)$
 $y_3 = (-1,-1,-1,0), y_4 = (-1,0,-1,0)$
假设初始权向量 $a_1 = (1,1,1,1), p_k = 1$
第一次迭代：
 $a_1^T y_1 = (1,1,1,1) \cdot (1,0,1,1) = 4 > 3 > 0$ 所以不修正
 $a_1^T y_2 = (1,1,1,1) \cdot (1,0,1,1) = 4 > 3 > 0$ 所以不修正
 $a_1^T y_3 = (1,1,1,1) \cdot (-1,-1,-1,0) = -3 < 0$ 所以修正 y_3
 $a_2 = a_1 + y_3 = (1,1,1,1) + (-1,-1,-1,0) = (0,0,0,1)$
 $a_2^T y_1 = (0,0,0,1) \cdot (1,0,1,1) = 1 > 0$ 所以修正 a_2
 $a_3 = a_2 + y_1 = (0,0,0,1) + (1,0,1,1) = (1,0,1,2)$
第一次迭代后，权向量 $a_3 = (1,0,1,2)$ ，再进行第2,3...次迭代直到在一个迭代过程中权向量相同，训练结束。
可得： $a_6 = a_3 = (0,0,1,3)$
判别函数： $g(x) = -x_1 + 3x_2$

【感知器】
线性不可分样本集的分类(最近邻解)
对于线性可分的样本集，可以用上述方法得到正确分类的权向量。
当样本集线性不可分时，用上述方法求权值时算法不收敛。如果我们把循环的权向量取平均值作为待求的权向量，或取其中之作为权向量，一般可以得到较满意的近似结果。
【最小平方误差判别】
基本思想
平方误差准则函数
 $J(a) = \|a\|^2 - 2a^T \sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i^2 y_i^2$
然后找一个使 $J(a)$ 极小化的 a 作为问题的解。
 $\nabla_a J(a) = 2a - 2 \sum_{i=1}^n x_i y_i = 0 \rightarrow a = \sum_{i=1}^n x_i y_i$
令 $\nabla_a J(a) = 0$ ，得 $a^2 - 2a = 0$
即 $a(a-2) = 0$
很显然， a 依赖于 x, y 的取值方法不同，根据具体应用不同。
b. 选取问题：
如果对应同一类样本的 b_i 选择为相同的值，那么最小平方误差方法的解等价于Fisher线性判别解。
如果对所有样本都取 $b_i = 1$ ，那么当 N 趋近于无穷大时，最小平方误差方法的解是贝叶斯判别函数的最小平方误差逼近。

模式识别或者分类的基本方法有两大类：
一类是将特征空间划分成决策域，需要确定判别函数或确定决策面方程。
另一类是模板匹配：将待分类样本与标准模板进行比较，看跟哪个模板匹配度更好，从而确定待测试样本的类别。
近邻法在原理上属于模板匹配。
它将训练样本集中的每个样本都作为模板，用测试样本与每个模板做比较，看与哪个模板最相似（即为近邻），就以最相似的模板的类别作为自己的类别。

近邻法优缺点：
1) 原理简单、易于实现，在模板数量很大时其错误率低。
2) 计算量大，存储量大，要存储的模板很多，每个测试样本要对每个模板计算一次相似度。
当样本充分多时，近邻法错误率在贝叶斯错误率 J^* 和两倍贝叶斯错误率 $2J^*$ 之间。



【样本集分级分解】
步骤：将整个样本分成 t 个子集，每个子集又分为它的 t 个子集，如此进行若干次就能建立起一个样本集的树状结构。子集分解的原则是该子集中的样本尽可能聚成团。
节点参数：树状结构，每个节点表示一个样本子集，描述该子集的参数是：
 \mathcal{X}_p ：节点 p 对应的样本子集
 N_p ：样本子集 \mathcal{X}_p 中的样本数目
 M_p ：样本子集 \mathcal{X}_p 中的样本均值
 $r_p = \max_{x_i \in \mathcal{X}_p} D(x_i, M_p)$ ：从 M_p 到 \mathcal{X}_p 的最远距离

【样本集搜索规则】
规则1：如果 $B + r_p < D(x, M_p)$ 成立，则 $x_i \in \mathcal{X}_p$ 不可能是 X 的最近邻。
 B ：当前已经涉及到的样本集 \mathcal{X}_p 中的样本到 X 的最近距离。
规则2：如果 $B + D(x_i, M_p) < D(x, M_p)$ 成立，其中 $x_i \in \mathcal{X}_p$ ，则 x_i 不可能是 X 的最近邻。

搜索过程分析：
当搜索树状样本集结构由高层次向低层次深入时，对同一层次的所有节点，可以利用规则1排除掉一些不可能包含待识别样本最近邻的节点(样本子集)。但是这往往不能做到只留下唯一的搜索节点，因此必须选择其中某一个子集先深入搜索，以类似以深度优先的方法确定搜索路径直至最后节点。然而在该最后节点中找到的近邻并不能保证确实是全样本集中的最近邻者，所找到的该近邻样本需要在那些有可能包含最近邻的样本子集中核对与修正，直至找到真正的最近邻样本为止。

