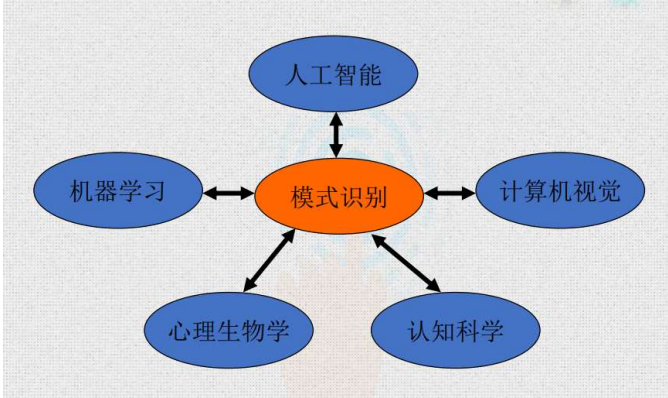
第一章绪论

模式识别就是采用机器（计算机）模仿人脑对现实世界各种事物进行描述、分类、判断和识别的过程的方法

广义地说，存在于时间和空间中可观察的物体，如果我们可以区别它们是否相同或是否相似，都可以称之为模式。模式所指的不是事物本身，而是从事物获得的信息，因此，模式往往表现为具有时间和空间分布的信息。

模式的直观特性:（1）可观察性（2）可区分性（3）相似性

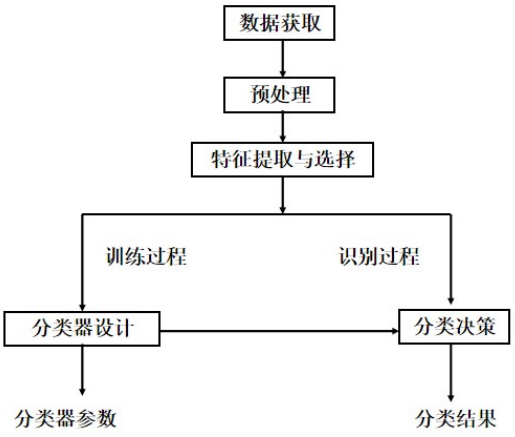


模式识别的相关学科

模式识别的应用：（1）图像分类（2）字符识别（3）医疗诊断（4）语音识别（5）指纹识别（6）遥感图像分类

模式识别系统

基本过程（1）训练过程（2）识别过程



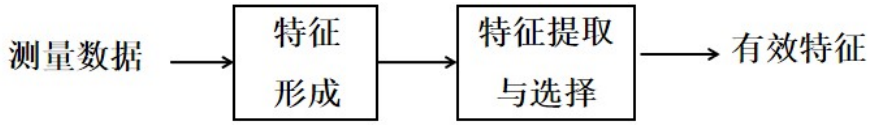
数据获取主要是由不同形式的传感器构成，实现信息获取与信息在不同媒体之间的转换。

预处理主要是指去除所获取信息中的噪声，增强有用的信息，及一切必要的使信息纯化的处理过程。

特征选择和提取将所获取的原始量测数据转换成能反映事物本质，并将其最有效分类的特征表示。

输入：原始的测量数据(经过必要的预处理)。

输出：将原始测量数据转换成有效方式表示的信息，从而使分类特征形成根据这些信息决定样本的类别。



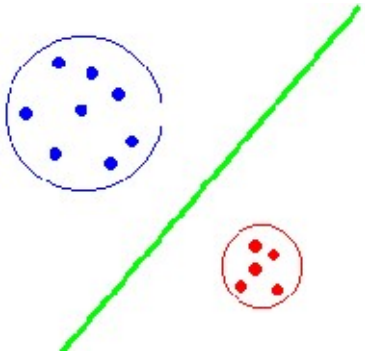
分类器设计：将该特征空间划分成由各类占据的子空间，确定相应的决策分界。
分类决策：是指分类器在分界形式及其具体参数都确定后，对待分类样本进行分类决策的过程

基本概念
模式类与模式：所见到的具体事物称为模式，而将他们的归属类别称为模式类。
样本：所研究对象的个体
样本集：若干样本的集合
类别：样本所属的事物类别，样本集中的子集。在同一类的样本在某种所关心的性质上是不可区分的。即具有相同的模式。
模式识别：将某一样本正确地归入某一模式。
特征（feature）：用于表征样本的量。

模式识别的主要方法
基于知识的方法：主要以专家系统为代表的方法，基本思想是：根据人们已知的（专家收集整理的）关于研究对象的知识，整理出若干描述特征与类别间关系的准则，建立一套计算机推理系统对未知样本进行决策分类。典型方法：句法模式识别
基于数据的方法：收集一些已知类别的样本，用这些样本作为训练集（training set） 来训练一个模式识别系统（分类器），达到能够对未知样本正确分类。这种方法可以看做是基于数据的机器学习（machine learning）的特殊情况。基础：**统计模式识别**

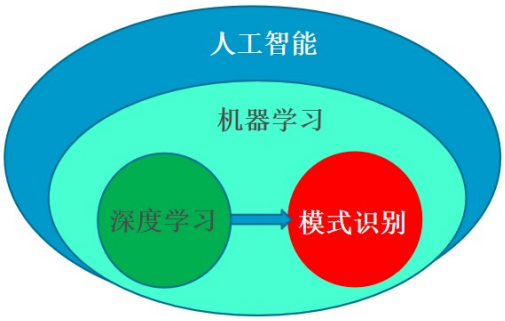
模式的描述方法
定量描述：一种是对事物的属性进行度量，属于定量的表示方法。
定性描述：另一种则是对事物所包含的成分进行分析，称为定性的描述或结构性描述

定量的表示方法及统计模式识别
特征：反映事物的本质属性的量。 f1,f2,…
特征向量：所有特征组合起来形成的向量[f1,f2,f3,...,fn]
特征空间: 特征向量张成的空间。特征向量即是特征空间中的一个样本点。
统计模式识别: 在特征空间与特征向量这种表示模式的方法前提下，讨论模式识别的基本理论与基本方法。
从数学上来说，即是对特征空间的划分或确定特征向量属于哪个特征空间的问题。或者说是寻找能将特征空间划分为不同子空间的边界问题。寻找这个边界或确定不同类别的子空间过程即是分类器设计过程。

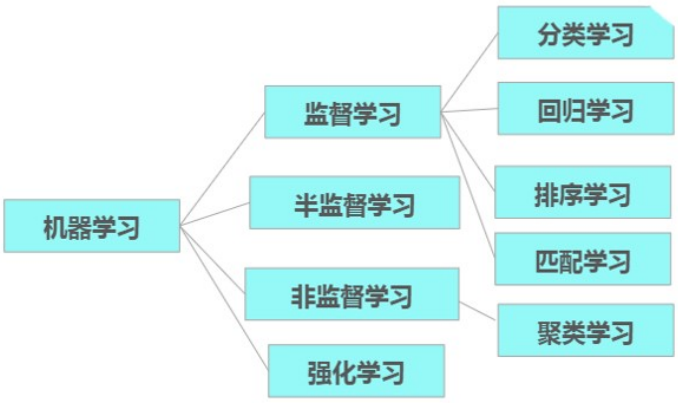


定性的描述方法及结构模式识别
结构性描述：由事物组成成分与相互关系表示的表示方法。常用的有串、树、图等
结构模式识别： 基于结构性描述方法讨论模式识别的基本理论与基本方法

机器学习、人工智能、深度学习、模式识别关系



机器学习（模式识别）算法分类



有监督学习：通过已知数据（样本）以及其对应的输出（类别属性----训练样本）来训练，得到一个最优模型（分类器），再利用这个模型（分类器）将所有新的数据样本映射为相应的输出结果，对输出结果进行简单的判断从而实现分类。

无监督学习：指在未加标签的数据中（无训练样本），根据数据本身之间的属性对数据进行分类，相似相近的数据分在同一类；不相似或不相近的数据分在不同的类中。

半监督学习：是监督学习与无监督学习相结合的一种学习方法。半监督学习使用大量的未标记数据，以及同时使用标记数据，来进行模式识别工作。

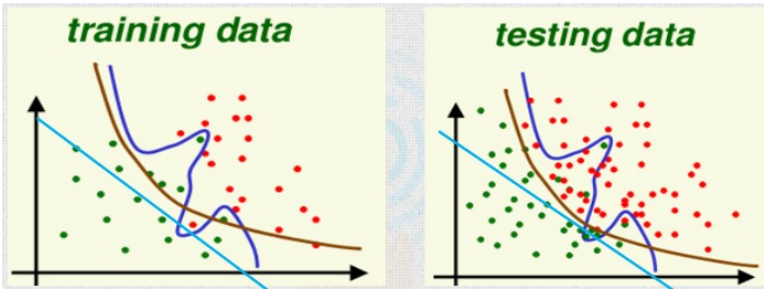
模型评价

偏差：算法的期望预测与真实值之间的偏差程度，反映了模型本身的拟合能力。

方差：方差度量了同等大小训练集的变动导致学习性能的变化，刻画了数据扰动所导致的影响。

误差（错误率）：训练误差（训练错误率、视在错误率），泛化误差（测试误差）

泛化能力：学得的模型适用于新样本的能力称为泛化能力，机器学习的目标是使学得的模型能够很好地适用于新的样本，而不是仅仅在训练样本上工作的很好。



欠拟合：模型过于简单(线性)，训练误差/错误率及测试误差（泛化误差）均较大。推广性差，泛化能力差模型过于简单

适中：模型较简单（二次曲线），训练误差适中，泛化误差错也不高。推广性好，泛化能力强

过拟合：模型复杂（复杂曲线），训练误差小（过拟合）。泛化误差大，推广性不好

模式识别的一些基本问题

1.学习

机器的学习过程（1）使用包含各种类别的训练样本（2）勾画出各类事物在特征空间分布的规律性（建立数学公式描述分类器）（3）建立准则及采用优化方法确定分类器的具体数学公式中的参数。

数学式子中参数的确定:是一种学习过程。如果当前采用的分类函数会造成分类错误，利用错误提供应如何纠错的信息，纠正分类函数

分类器设计：求解优化问题的过程

模式识别中的学习与训练是从训练样本提供的数据中找出某种数学式子的最优解，这个最优解使分类器得到一组参数，按这种参数设计的分类器使人们设计的某种准则达到极值。分类器参数的选择或者学习过程得到的结果取决于设计者选择什么样的准则函数。不同准则函数的最优解对应不同的学习结果，得到性能不同的分类器。

2.模式的紧致性

分类器设计难易程度与模式在特征空间的分布方式有密切关系

形象说法：不要混迭，分界面干净利索

特征提取：改善数据紧致性

3.相似性度量

在特征空间中用特征向量描述样本的属性。样本间的关系则采用相似性度量来描述，一般采用距离度量表示。同类样本应具有聚类性，或紧致性好不同类别样本应在特征空间中显示出具有较大的距离。统计模式识别各种方法实际上都是直接或间接以距离度量为基础的。

第 2 章　有监督模式识别——贝叶斯决策理论

模式识别是一种分类问题，即根据识别对象所呈现的观察值，将其分到某个类别中去。

统计决策理论是处理模式分类问题的基本理论之一，对模式分析和分类器的设计起指导作用。

贝叶斯决策理论是统计模式识别中的一个基本方法。是一种将特征空间划分为子空间的方法，对模式分析和分类器的设计起指导作用。

贝叶斯决策理论的核心是当给定具有特征向量 X 的待识别样本时，它属于某一类的可能性有多大。

分类问题的描述

已知总共有 c 类样本 ω i(i=1,2,……,c)，其先验概率为 P(ω i)，条件概率密度函数为 p(X| ω i)，样本分布在 n 维特征空间，则对于待识别样本，如何确定其所属类别？

由于属于不同类的待识别对象存在着呈现相同观察值的可能，即所观察到的某一样本的特征向量为 X，而在 c 类中又有不止一类可能呈现这一 X 值，这种可能性可用 P(ω i|X)(i=1,2,……,c)表示。

如何做出合理的判决就是贝叶斯决策理论所要讨论的问题。

贝叶斯决策理论方法所讨论的问题是：

已知:总共有 c 类物体，以及先验概率 P(ωi)及类条件概率密度函数 P(x|ωi)

问题: 如何对某一样本按其特征向量分类的问题。

先验概率 P(ωi)是指ωi(i=1,2,……,c)出现的可能性，不考虑其它任何条件。

类条件概率密度函数 P(X|ωi)是指ωi 条件下在一个连续的函数空间出现 X 的概率密度，也就是第ωi 类样本的特征 X 是如何分布的。它是指在某种确定类别条件下，样本 X 出现的概率密度分布函数，常用 p(X|ωi)(iϵ1,2,……,c)来表示。

后验概率是指在某个具体的模式样本 X 条件下，某种类别出现的概率，以 P(ωi|X)表示。

后验概率可以根据贝叶斯公式计算得到。

$$P(\omega_i|X)=\frac{p(X|\omega_i)P(\omega_i)}{p(X)}, \qquad p(X)=\sum_{i=1}^{c}p(X|\omega_i)P(\omega_i)$$

假设一个待识别的物理对象用其 d 个属性观察值描述，称之为 d 个特征，每个观察值即是一个特征。这 d 个特征组成一个 d 维的向量，叫特征向量。记为 x = [x1, x2, …, xd]T，d 维待征所有可能的取值范围则组成了一个 d 维的特征空间。

最大先验概率分类

$$\left.\begin{array}{l} P(\omega_1) > P(\omega_2), x \in \omega_1 \\ P(\omega_1) < P(\omega_2), x \in \omega_2 \end{array}\right\}$$

这种分类决策与具体观测 X 无关，没有意义，表明由先验概率所提供的信息太少。按先验概率大小分类不合理

最大后验概率决策——基于最小错误率的贝叶斯决策

分类识别中为什么会有错分类？

当某一特征向量值 X 只为某一类物体所特有，对其作出决策是容易的，也不会出什么差错，问题在于出现模棱两可的情况，c 类中有不止一类可能呈现这一 X 值，任何决策都存在判错的可能性。

后验概率: P(ω1|x)和 P(ω 2 |x) 同一条件 x 下，比较ω1 与ω2 出现的概率 两类ω1 和ω2，则有 P(ω1|x)+P(ω2|x)=1 如 P(ω1|x)> P(ω2|x)则可以下结论，在 x 条件下，事件ω1 出现的可能性大	类条件概率密度: p(x|ω1)和 p(x|ω2) 是在不同条件下讨论的问题 即使只有两类ω1 与ω2，p(x|ω1)+p(x|ω2)≠1 p(x|ω1)与 p(x|ω2)两者没有联系

基于最小错误概率的贝叶斯决策理论就是按后验概率的大小作判决的

这是考虑了 X 属于哪类的概率，可以作为分类的准则。----最大后验概率准则。

具体规则如下：

$$\text{若：}\ P(\omega_i \mid X) = \max_{j=1,\ldots,c} P(\omega_j \mid X)\ \text{则：}\ X \in \omega_i$$

$$P(\omega_i \mid X) = \frac{p(X \mid \omega_i)P(\omega_i)}{p(X)} = \frac{p(X \mid \omega_i)P(\omega_i)}{\sum_{i=1}^{c} p(X \mid \omega_i)P(\omega_i)}$$

最大后验概率决策的其他形式

(1) 用先验概率及类条件概率密度函数表示

$$\text{若：}\ p(X \mid \omega_i)P(\omega_i) = \max_{j=1,2} p(X \mid \omega_j)P(\omega_j)\ \text{则：}\ X \in \omega_i$$

(2) 用比值的方式表示-----似然比

$$\text{如果} l(x) = \frac{p(X \mid \omega_1)}{p(X \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} = \lambda \quad \text{则：} \quad X \in \omega_1$$

(3) 对数似然比

$$h(x) = -\ln[l(x)]$$

$$= -\ln p(X \mid \omega_1) + \ln p(X \mid \omega_2) < \ln \frac{P(\omega_1)}{P(\omega_2)} \quad \text{则：} X \in \omega_1$$

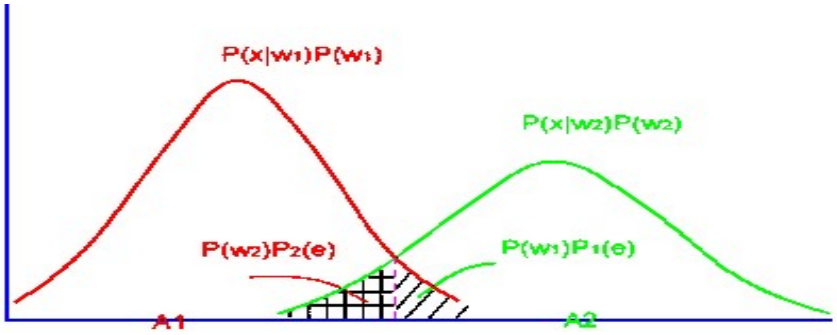
最大后验概率决策即是最小错误率决策的证明

把作出ω1 决策的所有观测值区域称为 R1，则在 R1 区内的每个 x 值，条件错误概率为 p(ω2|x)。另一个区 R2 中的 x,条件错误概率为 p(ω1|x)。

$$P(e)=\int_{R1}P(\omega_2\mid x)p(x)dx+\int_{R2}P(\omega_1\mid x)p(x)dx$$

$$P(e)=P(\omega_2)\int_{R1}p(x\mid\omega_2)dx+P(\omega_1)\int_{R2}p(x\mid\omega_1)dx$$

$$=P(\omega_2)P_2(\mathbf{e})+P(\omega_1)P_1(\mathbf{e})$$



在 R1 区内任一个 x 值都有 P(ω2|x)＜P(ω1|x)，或 P(ω2)p(x|ω2)<P(ω1)p(x|ω1)

在 R2 区内任一个 x 值都有 P(ω2|x)>P(ω1|x) ，或 P(ω2)p(x|ω2)>P(ω1)p(x|ω1)

错误率在每个 x 值处都取小者，因而平均错误率 P(e)也必然达到最小

因而，按最大后验概率作出的决策，其平均错误率为最小

最小风险的贝叶斯决策

使错误率最小并不一定是一个普遍适用的最佳选择。错误的代价(损失)不同，宁可扩大一些总的错误率，但也要使总的损失减少。引进一个与损失有关联的，更为广泛的概念——风险。在作出决策时，要考虑所承担的风险

(1)自然状态与状态空间。

自然状态：指待识别对象的自然类别: ωi

状态空间Ω：由所有自然状态所组成的空间， Ω={ω1， ω2，…， ωc}

(2)决策与决策空间。

对分类问题所作的判决，称之为决策， αi 。

由所有决策组成的空间称为决策空间。 A={α1， α2,….., αK}

决策不仅包括根据观测值将样本划归哪一类别(状态)，还可包括其它决策，如“拒绝”等，因此决策空间内决策总数 K 可以不等于类别数 c

(3)损失函数λ(αi|ωj)或写成λ(αi,ωj)或λij 表示对自然状态ωj 作出决策αi 时所造成的损失。

(4)条件风险：将观测值 X 下决策为αi 的期望损失

$$R(\alpha_i\mid X)=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j\mid X)\qquad i=1,2,...,K$$

（5）最小风险贝叶斯决策规则(条件风险最小)：

若： $R(\alpha_i\mid X)=\min_{j=1,2,...,K}R(\alpha_j\mid X)$ 则： $X\in\alpha_i$

似然比决策规则

如果 $l(x)=\frac{p(X\mid\omega_1)}{p(X\mid\omega_2)}>\frac{P(\omega_2)(\lambda_{12}-\lambda_{22})}{P(\omega_1)(\lambda_{21}-\lambda_{11})}$， 则 $X\in\omega_1$

两种决策方法之间的关系：0-1 损失时 R(αi|X)最小即是 P(ωi|X)最大，最小错误率贝叶斯决策就是 0-1 损失函数下的最小风险贝叶斯决策，基于最小错误率的决策是基于最小风险决策的特例。

决策域：各类别在特征空间内所占的区域。
分类决策：待识别的特征向量落在哪个决策域，该样本就被判为哪一类。
判别函数：用于表达决策规则的某些函数则称为判别函数。
决策面及决策面方程：决策域的边界面就是决策面，在数学上用解析形式表示成决策面方程。
决策面是一种统称，特征空间是一维时决策面是一个点。二维特征空间里是一条曲线。三维则是一曲面。超过三维的空间，决策面是一个超曲面。

正态分布时的统计决策
（1）正态分布在数学上比较简便 （2）物理上的合理性
在模式识别中，单变量正态分布不是先验概率 P(ωi)，也不是后验概率 P(ωi|X)，而是 p(x|ωi)。
单变量正态分布概率密度函数定义为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ表示随机变量 x 的数学期望

$$\mu=E\{x\}=\int_{-\infty}^{+\infty}xp(x)dx$$

σ2 为其方差，而σ则称为标准差

$$\sigma^2=\int_{-\infty}^{+\infty}(x-\mu)^2p(x)dx$$

多元正态分布
多维向量：是一个随机向量，每一个分量都是随机变量，服从正态分布。每一个随机向量不仅要求考虑每个分量单独的分布，还要考虑两个随机变量之间的关系----相关性。

$$p(X)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}$$

$$\mu=E\{X\}=[\mu_1,\mu_2,...,\mu_d]^T$$

$$\Sigma=E\{(X-\mu)(X-\mu)^T\}$$

Σ是 d×d 维协方差矩阵
Σ是对称非负定矩阵，在此我们只考虑正定阵，即|Σ|＞0
衡量相关性 ，协方差越大，说明两个变量的相关度越高。非对角元素正表示了两个分量之间的相关性。主对角元素则是各分量本身的方差。是一个正定的对称矩阵

$$\Sigma=\begin{bmatrix}E[(x_1-\mu_1)^2] & E[(x_1-\mu_1)(x_2-\mu_2)]\\ E[(x_2-\mu_2)(x_1-\mu_1)] & E[(x_2-\mu_2)^2]\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \sigma_{12}^2\\ \sigma_{21}^2 & \sigma_2^2\end{bmatrix}$$

多元正态分布的性质
(1)参数μ与Σ对分布具有决定性。记作 p(X)～N(μ，Σ)。
(2)等密度点分布在超椭球面上 $(x-\mu)^T\Sigma^{-1}(x-\mu)$＝常数，二维时表示一个椭圆，在三维表示椭球，在高维是表示超椭球，这是一个二次型问题。
(3)多元正态分布的离散程度由参数$|\Sigma|^{1/2}$决定----与单变量时由标准差σ决定是对应一致的
(4)不相关性等价于独立性 ；不相关：E[xixj]＝E[xi]·E[xj] ；独立：p(xi,xj)＝p(xi)p(xj)
两个随机变量不相关，不意味着它们一定独立；相互独立的随机变量，它们之间是不相关的
正态分布中不相关性等价于独立性
(5)边缘分布和条件分布的正态性
多元正态分布的边缘分布和条件分布仍然是正态分布

(6)线性变换的正态性

这是指多元正态分布的随机向量的线性变换仍然是多元正态分布的随机向量

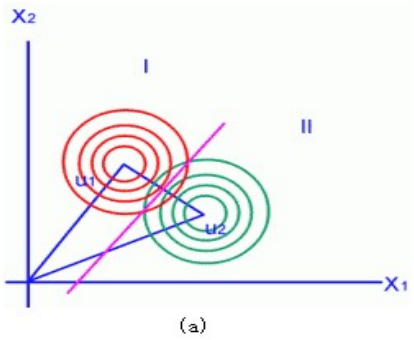
$$Y = AX \Rightarrow p(Y) \sim N(A\mu, A\Sigma A^T)$$

(7)线性组合的正态性

这是指多元正态分布的随机向量，在经过线性组合后得到的一维随机变量也是正态分布的

$$y = \alpha^T X \Rightarrow p(y) \sim N(\alpha^T \mu, \alpha^T \Sigma \alpha)$$

在正态分布条件下，基于最小错误率贝叶斯决策只要能做到两类协方差矩阵是一样的，那么无论先验概率相等还是不相等，都可以用线性分界面实现。



(a)

最小欧氏距离分类器则要求正态分布协方差矩阵为单位阵，先验概率相等。

$\Sigma i=\sigma^2 I$

决策面方程 $W^T(X-X0)=0$

决策面为一超平面

其法线方向为 $W=\mu i-\mu j$

通过 X0 点

当 $P(\omega i)=P(\omega j)$时该超平面过$(\mu i+\mu j)/2$ 点(最小距离分类器)

判别函数：$g_i(X)=\|X-\mu_i\|^2$，$i=1,2,.....,C$，这里μ_i是ω_i类的均值向量。

决策规则：$\|X-\mu_i\|^2=\min_{j=1,2,...,C}\{\|X-\mu_j\|^2\}$若，则$X\in\omega_1$

在二维情况下，就是过μi 与μj 连线的垂直平分线

当 $P(\omega i)\neq P(\omega j)$时，该超平面的位置要向远离先验概率大的方向偏，但超平面方向不变。

最小马氏距离分类器则要求正态分布协方差矩阵相等，先验概率相等

$\Sigma i=\Sigma$

决策面方程 $W^T(X-X0)=0$

$$W=\Sigma^{-1}(\mu_i-\mu_j)$$

当 $P(\omega i)=P(\omega j)$时该超平面过$(\mu i+\mu j)/2$ 点(最小马氏距离分类器)

$(X-\mu)^T\Sigma^{-1}(X-\mu)$称为向量X到向量μ的马氏距离的平方，

即$r^2=(x-\mu)^T\Sigma^{-1}(x-\mu)$

分类器错误率/正确率的实验估计方法

采用测试样本集（已知类别属性的样本）通过估计分类器错误率/正确率来测试分类器的性能。一般会将已知类别属性的样本分成训练样本及测试样本集两部分。训练样本集用于训练（设计）分类器。测试样本集用于测试分类器性能

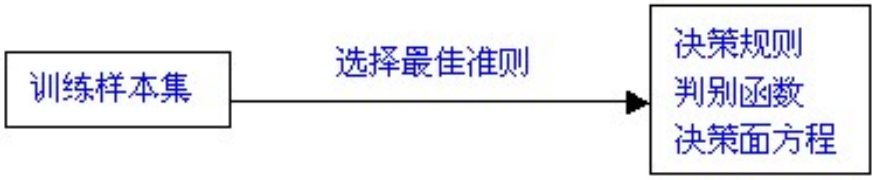
贝叶斯决策理论：要设法获取样本统计分布的资料,要知道先验概率，类条件概率密度函数等。类条件概率密度函数的确定是通过确定其函数形式 $p(x|\omega i)$并对其参数估计来完成的。因此，以贝叶斯决策方法为基础的方法称为参数判别方法

根据训练样本集直接进行分类器设计。这种方法绕过统计分布状况的分析，绕过参数估计这一环，而企图对特征空间实行划分，称为非参数判别分类法，即不依赖统计参数的分类法。

非参数判别分类方法两个过程

（1）确定使用什么典型的分类决策方法，即决定判别函数类型（如线性判别函数）及优化准则

（2）利用训练样本集提供的信息及优化准则（Fisher 准则、感知函数准则等）确定这些函数中的参数。 相对最小错误率及最小风险决策（最优分类器）而言，是次优方法，但在所提准则下，是最好的。



线性分类器

判别函数是线性判别函数的分类器称为线性分类器

主要工作：用训练样本去估计线性判别函数的参数

线性判别函数的一般形式

$$g(X) = W^T X + w_0$$

$$X = \begin{pmatrix} x_1 \\ \cdots \\ x_2 \end{pmatrix} \quad W = \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_d \end{pmatrix}$$

如果

$g(X)>0$，则决策 $X \in \omega_1$

$g(X)<0$，则决策 $X \in \omega_2$

$g(X)=0$，可将其任意分类或拒绝。

g(X)＝0 就是相应的决策面方程，在线性判别函数条件下它对应 d 维空间的一个超平面

W 是该超平面（分界面）的法线向量

g(X)是 d 维空间任一点 X 到决策面 H 的距离的代数度量

符号表明类别，数值是距离的代数度量

g(x)>0, X 在 H 的正侧；X→ ω1

g(x)<0, X 在 H 的负侧；X→ ω2

w0 体现该决策面在特征空间中的位置

(1) w0=0 时，该决策面过特征空间坐标系原点

(2)否则，R0=w0/||W||表示坐标原点到决策面的距离

设Xp是X在H上的投影向量，

r是x到H的垂直距离

W/||W||是W方向的单位向量，则

$g(X) = W^T x + w_0 = W^T(Xp + rW/||W||) + w_0$

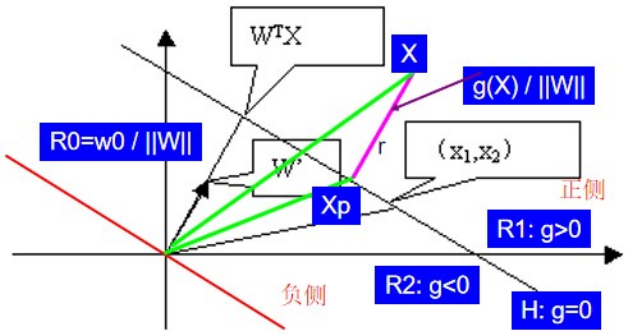
$= (W^T Xp + w_0) + rW^T W/||W||$

$= r||W||$

$r = g(X)/||W||$

若X=0(原点)，则$g(X) = W^T x + w_0 = w_0$

否则：$r0 = g(0)/||W|| = w_0/||W||$



广义线性判别函数

选择一种映射 X→Y，将原样本特征向量 X 映射成另一向量 Y，从而可以采用线性判别函数的方法。

线性判别函数的齐次简化

增广样本向量　　增广权向量 (广义权向量)

$$Y=\begin{bmatrix}1\\x_1\\x_2\\\vdots\\x_d\end{bmatrix}=\begin{bmatrix}1\\X\end{bmatrix}\qquad \bar{a}=\begin{bmatrix}w_0\\w_1\\w_2\\\vdots\\w_d\end{bmatrix}=\begin{bmatrix}w_0\\W\end{bmatrix}$$

将 g(x)中的 W 向量与 w0 统一表示成

$$g(x)=w_0+\sum_{i=1}^{d}w_ix_i=\bar{a}^{T}Y$$

它使特征空间增加了一维，但保持了样本间的欧氏距离不变，对于分类效果也与原决策面相同，只是在 Y 空间中决策面是通过坐标原点的，这在分析某些问题时具有优点，因此经常用到。

线性分类器设计任务

在给定样本集 X={X1,X2,…,XN}条件下， 确定线性判别函数的各项系数,w1,w2,…,wd ，以期对待测样本进行分类时，能满足相应的准则函数 J 为最优的要求。

关键问题:确定所需的准则函数，然后用最优化技术求解使准则函数的极值解 w*及 w0*（或增广权向量 a*）

Fisher 线性判别函数基本原理

线性判别函数可以看成是把 d 维空间映射到 1 维空间

$g(X)= W^TX+ w0$　　→　　$g(X) = W^TX > - w0$　则

不考虑 w0，则 $g(X)=W^TX$ 是各样本向量与向量 W 的点积。（又可看作各样本向量在向量 W 上的投影）。然后在一维上找 w0

如果在二维空间中一条直线能将两类样本分开，或者错分类很少，则同一类别样本数据在该直线的单位法向量上的投影的绝大多数都应该超过某一值。而另一类数据的投影都应该小于(或绝大多数都小于)该值，则这条直线就有可能将两类分开。

Fisher 准则就是要找到一个最合适的投影轴，使两类样本在该轴上投影的交迭部分最少，从而使分类效果为最佳。

分析 w1 方向之所以比 w2 方向优越，可以归纳出这样一个准则 ：

向量 W 的方向选择应能使

（1）两类样本投影的均值之差尽可能大些

（2）类内样本的离散程度尽可能小

Fisher 选择投影方向 W 的原则： y=WTX

类间分布尽可能分开，

类内样本投影尽可能密集的要求

Y 空间

各类样本均值

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in Y_i} y, \qquad i=1,2$$

样本类内离散度 和总类内离散度

$$\tilde{S}_i^2 = \sum (y-\tilde{m}_i)^2, \qquad i=1,2$$

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

评价投影方向 W 的函数（目标函数）

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

X 空间

各类样本均值向量

$$m_i = \frac{1}{N_i}\sum_{x\in X_i} x \qquad i=1,2$$

样本类内离散度矩阵

$$S_i = \sum_{X\in S_i} (X-m_i)(X-m_i)^T, \qquad i=1,2$$

总类内离散度矩阵 Sw

$$S_w = S_1 + S_2$$

样本类间离散度矩阵 Sb

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

$$\tilde{m}_i = \frac{1}{N}\sum_{y\in Y_i} y = \frac{1}{N}\sum_{y\in Y_i} W^T X = W^T m_i \qquad i=1,\ 2$$

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

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

$$\tilde{S}_i^2 = \sum_{y\in Y_i} (y-\tilde{m})^2 = \sum_{x\in S_i} (W^T X - W^T m_i)^2 = W^T \left[\sum (X-m_i)(X-m_i)^T\right] W = W^T S_i W$$

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

J_F 进一步化为 W 的显函数

$$J_F(W) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2} = \frac{W^T S_b W}{W^T S_w W}$$

最佳 W 值的确定:　求取使目标函数 J_F 达极大值时的 w*。

由于 W 乘以比例因子，不影响 J_F 的的取值，W 乘以一个常数 k 以使分母等于一常数 C，则将求 J_F 的无约束极大值的问题化为有约束的条件极大值问题

$$\max_{W}\{W^T S_b W\} \quad s.t \quad W^T S_W W = c$$

可采用***拉格朗日法***求解。

$$L(W,\lambda) = W^T S_b W - \lambda(W^T S_W W - c)$$

$$\frac{\partial L}{\partial W} = (Sb + Sb')W - \lambda(Sw + Sw')W$$

$$= 2SbW - 2\lambda SwW = 0$$

$$SbW = \lambda SwW$$

$$SbW = (m1 - m2)(m1 - m2)^T W$$

$R = (m1 - m2)^T W$是标量

即$(m1 - m2)R = \lambda SwW$

$$W = \frac{R}{\lambda} Sw^{-1}(m1 - m2)$$

$\frac{R}{\lambda}$是常数不影响方向，将其忽略

$$W = Sw^{-1}(m1 - m2)$$

W0 的确定

$$W_0 = -\frac{\tilde{m}_1 + \tilde{m}_2}{2}$$

两类正态分布且具有相同的协方差矩阵Σ时，按最小错误率的贝叶斯决策得到的

$$W=\Sigma^{-1}(u_1-u_2)$$

如果 P(ωi)=P(ωj)，则最佳分界线就是两类概率密度函数值相等的点的集合。

因此，Sw=N1Σ1+ N2Σ2=(N1+N2)Σ~ Sw= Σ ，

按 Fisher 准则求得：

$$W=\Sigma^{-1}(u_1-u_2)$$

可见：若两类样本的离散度矩阵相近，也就是说两类分布的形式很相近，按 Fisher 准则，错分率就应比较小（接近最小错误率），Fisher 准则的合理性可以在这里体现

感知器算法

特点：随意确定判别函数的初始值，在对样本分类训练过程中逐步修正直至最终确定。

感知准则函数使用增广样本向量与增广权向量

$$g(X) = w_0 + \sum_{i=1}^{d} w_i x_i = a^{-T} Y$$

$$g(X) = a^{-T} y \begin{cases} > 0, & X \in \omega_1 \\ < 0, & X \in \omega_2 \\ = 0, & \text{由设计者选择决策} \end{cases}$$

线性可分性

若训练样本集是线性可分的，则必存在一个权向量 a，可使该训练样本集中的每个样本正确分类。

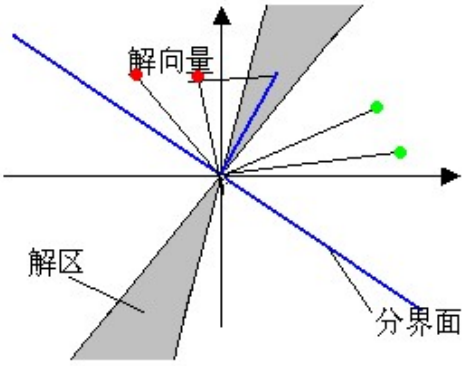
样本规范化

$$Y' = \begin{cases} Y & \text{若} Y \in \omega_1 \\ -Y & \text{若} Y \in \omega_2 \end{cases}$$

Y’ 称为规范化的增广样本向量。则合适的 a 能使所有的 Y'满足 $a^T Y' > 0$

满足 $a^T Y' > 0$ 的权向量 a 称为解向量。

解向量存在无穷多个，解向量组成的区域称为解区



令被错分类的规范化增广样本组成的集用 y^k 表示，并定义一准则函数 Jp(a)

$$J_p(a) = \sum_{y \in y^k} (-a^T y) \geq 0$$

对线性可分情况:

最佳的 a 应能将该样本集中所有样本正确分类，即 y^k 是空集

$$J_p(a) = \sum_{y \in y^k} (-a^T y) = 0; \qquad y^k = \Phi$$

确定向量 a 的问题变为求 Jp(a)的极小值的问题

梯度下降算法

$$\nabla J_p(a) = \frac{\partial J_p(a)}{\partial a} = \frac{\partial [\sum_{y \in y^k} (-a^T y)]}{\partial a} = \sum_{y \in y^k} (-y)$$

$$a(k+1) = a(k) - \rho_k \nabla J_p = a(k) + \rho_k \sum_{y^k} y, \quad \rho_k > 0$$

可以证明，对于线性可分的样本集，经过有限次修正，一定可以找到一个解向量 ，即算法能在有限步内收敛。

收敛速度取决于初始权向量 a(0) 和系数 ρ k 。

感知器方法是一种利用错分类样本对现决策权向量进行修正直至收敛的方法。

感知器方法只对线性可分样本集有效

对线性不可分的样本集，该算法不能收敛。其它方法，如最小错分样本数准则等。

这一节对感知器算法的讨论，只是很初步的，并且只讨论了线性可分的情况。但这种利用错误提供的信息，进行自修正的思想意义是十分深远的。

这种只解决线性分类的感知器称为单层感知器，由它基础上发展起来的多层感知器在原理上能解决非线性分类、多类划分，以及非线性拟和非线性映射等多种功能，是人工神经元网络的基础。

非线性判别函数

非线性判别函数可用分段线性判别函数近似

每一类的样本数据在特征空间中的分布呈复杂分布时，使用线性判别函数就会产生很差的效果

如果能将它们分割成子集，而每个子集在空间聚集成团，那么子集与子集的线性划分就可以取得比较好的效果。

同一类样本可以用若干个子类来描述，子类的数目就可作为确定分段段数的依据。

因此分段线性判别的主要问题是如何对数据划分成子集的问题（聚类方法）。

样本分布及子类划分已定或仅已知子类数目的情况下，设计分段线性判别函数的问题

样本分布及合适子类划分并不知道，则采用聚类的方法首先将样本进行划分 ，分成不同子类，然后按上述方法处理

分段线性分类器设计的一般考虑

一、已知样本的子类划分（分段线性分类器等）

把每个子类看成一个类，然后按多类线性判别函数算法将各子类分开。

二、子类数目已知，但子类划分未知

采用错误修正算法(略)----感知准则函数的扩展

三、子类数目、子类划分均未知

方法很多，典型方法是树状分段线性分类器

线性回归

回归分析的一种，目标值（因变量）与特征值（自变量）之间用线性关系描述 ：$Y=W^TX+e$，e 为误差，服从 0 均值高斯（正态）分布。

线性回归就是利用已知的数据对（X，Y）,建立其关系， $Y=W^TX$，使得误差 e 最小（目标）。

多元线性回归 $y_i=\theta^T X_i+e$

具有两个或两个以上的自变量 $\mathrm{X}_i=\left[x_i^1, x_i^2, \cdots, x_i^n\right]$

拟合值：$h_\theta(\mathrm{X}_i)=\theta_0+\theta_1 x_i^1+\theta_2 x_i^2+\cdots+\theta_n x_i^n=\theta^T X_i$

其中：$\theta=\left[\theta_0 \quad \theta_1 \quad \cdots \quad \theta_n\right]^T$

$X_i=\left[1 \quad x_i^1 \quad x_i^2 \quad x_i^3 \cdots \quad x_i^n\right]^T \quad i=1,2,\ldots,m$

目标函数：$J=\frac{1}{2}\sum_{i=1}^{m}(\theta^T \mathrm{X}_i-y_i)^2=\frac{1}{2}(\mathrm{X}\theta-\mathrm{Y})^T(\mathrm{X}\theta-\mathrm{Y})$

$Y=\left[y_1 \quad y_2 \quad \cdots \quad y_m\right]^T \qquad X=\begin{bmatrix}1 & x_1^1 & & x_1^n\\ 1 & x_2^1 & & x_2^n\\ & & \ddots & \\ 1 & x_m^1 & & x_m^n\end{bmatrix}$

$J=\frac{1}{2}\sum_{i=1}^{m}(\theta^T \mathrm{X}_i-y_i)^2=\frac{1}{2}(\mathrm{X}\theta-\mathrm{Y})^T(\mathrm{X}\theta-\mathrm{Y})$

令：$\frac{\partial J}{\partial \theta}=X^T X\theta-X^T \mathrm{y}=0$

得：$\theta=(X^T X)^{-1}X^T \mathrm{y}$

近邻法

非参数法中最重要的方法之一；原理上属于模板匹配

近邻法的优点：在模板数量很大时其错误率小

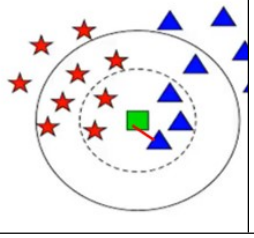
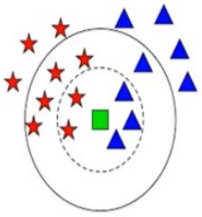
近邻法的缺点：计算量大，存储量大

改进的办法：剪辑近邻法与压缩近邻法

基于距离的分段线性函数的极端情况

每个样本是一个子类

分类方法：以全部训练样本作为“代表点”，计算测试样本与这些“代表点”（即所有样本）的距离，并以最近邻者（最近似的”代表点”）的类别作为分类的类别。这种决策方法就是近邻法

最近邻	K 近邻
对一个C类别问题，每类有N_i个样本，i＝1，…，则第i类ωi的判别函数 $g_i(X)=\min_k \|X-X_i^k\|$，k=1，…，$N_i$ X_i^k表示是ω_i类的第k个样本 决策规则 若： $g_j(X)=\min_i g_i(X)$，i=1，…，C， 则：X∈ωj 	在所有N个样本中找到与测试样本的k个最近邻其中各类别所占个数表示成k_i,i＝1，…，c 则决策规划是： 如果 $k_j(X)=\max_i k_i(X)$，i=1，…，c 则：X∈ω_j 

错误率分析

当 N→∞时，最近邻法的渐近平均错误率的下界是贝叶斯错误率

在其它条件下，最近邻法的错误率要高于贝叶斯错误率，可以证明以下关系式成立

$$P^* \le P \le P^*\left(2-\frac{C}{C-1}P^*\right)$$

一般情况下 P*很小 ，则：P*<P<2P*

当 k 增大时错误率是单调递减的。因此,在 N→∞的条件下，k-近邻法的错误率要低于最近邻法。

朴素贝叶斯

基于最小错误率贝叶斯决策与特征相互独立假设的分类方法

「朴素」指的是假设特征向量中的各个特征相互独立。这当然不是一个完美的假设，所以才「朴素」。

假设某个体有n项特征（Feature），分别为F_1、F_2、…、F_n。现有m个类别（Category），分别为C_1、C_2、…、C_m。贝叶斯分类器就是计算出概率最大的那个分类，也就是求下面这个算式的最大值：

$$P(C|F1F2...Fn)=P(F1F2...Fn|C)P(C)/P(F1F2...Fn)$$

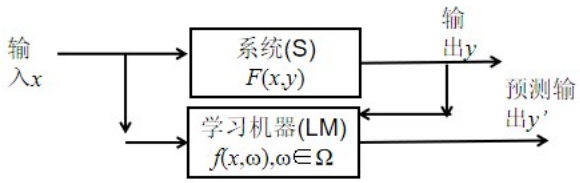
（特征相互独立时）

$$P(F1F2...Fn|C)P(C)$$

类条件概率密度与先验概率乘积

$$P(F1F2...Fn|C)P(C)=P(F1|C)P(F2|C)...P(Fn|C)P(C)$$

机器学习的基本模型



机器学习的目的：根据已知的训练样本求取系统对的输入(x)输出(y)的关系 F(x,y)的估计 f，使其能够对某一输入的未知输出做出尽可能准确的估计。

模型求解：风险最小化准则 损失函数 L(y,f(x,ω))：用 f(x,ω) 对输出 y 预测带来的损失 风险函数（期望风险）： $R(\omega)=\int L(y,f(x,\omega))dF(x,y)$ 学习目标：根据 F(x,y)及训练数据集（xi,yi），寻找函数 f(x,ω0) ，使它在函数类 {f(x,ω), ω ∈Ω } 最小化风险 R(ω)。 期望风险最小：需要已知 F(x,y) 实际： F(x,y) 未知，无法计算期望风险。	实际处理：经验风险最小化(ERM) 经验风险（实际情况：有限样本）： $R_{emp}(\omega)=\frac{1}{N}\sum_{i=1}^{N}L(y,f(x,\omega))$ 经验风险最小化:根据训练数据集（xi,yi），寻找函数 f(x,ω0) ，使它在函数类 {f(x,ω), ω ∈Ω } 最小化经验风险 $R_{emp}(\omega)$。

VC 维：衡量函数集 f(x, ω) （模型）的复杂性的指标。用 h 表示，h 是整数。模型越复杂，h 越大

$$R(\omega)\le R_{emp}(\omega)+\Phi(n/h)$$

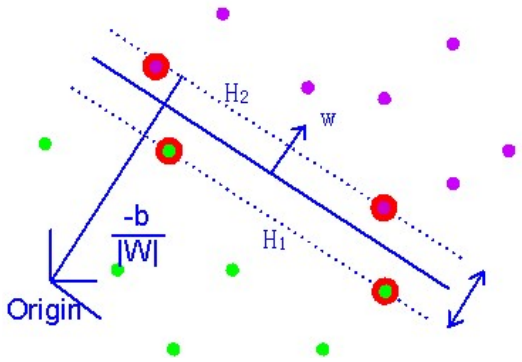
Φ(n/h)： 称置信范围或 VC 信任， 是随 n/h 增大而减小的函数。对于特定的问题，n 一般固定。因此，VC 维 h 越大，真实风险与期望风险的误差越大。因此，设计分类器时，不仅要考虑使经验风险最小，而且，要考虑机器的复杂性，使 VC 维尽量小。这样，期望风险才会小。

最优的学习机器：应使得 Remp(ω)和 Φ(n/h)同时最小。

SVM 的思路：

由于两类别训练样本线性可分，因此在两个类别的样本集之间存在一个隔离带。

如图示，H 是将两类分开的分界面，而 H1 与 H2 与 H 平行，H 是其平分面，H1 上的样本是第一类样本到 H 最近距离的点，H2 的点则是第二类样本距 H 的最近点。



支持向量: 处在隔离带的边缘上的样本点，称为支持向量，它们决定了这个隔离带。

分界面：不止一个，H1、 H2 的方向也随之改变，H1 与 H2 之间的间隔会发生改变。显然使 H1 与 H2 之间间隔最大的分界面 H 是最合理的选择，因此最大间隔准则就是支持向量机的最佳准则。

对需要非线性分类界面的情况，支持向量机采用的方法与前面提到的方法很不相同，支持向量机是利用***特征映射方法***，使非线性分类的问题可以利用线性分类的计算框架来实现。

由于特征进行了映射，从 x 变成了 f(x)，因此问题是在另一个映射后的空间讨论的。

设原空间维数为 d，即 $X\in R^d$ ，而新空间为 m 维，即 $f(X)\in R^m$ ，一般 m 维要比 d 维大得多。

第四章 特征的选择与提取

一般说来要对初始的特征空间进行优化是为了降维。即初始的特征空间维数较高。能否改成一个维数较低的空间，称为优化，优化后的特征空间应该更有利于后续的分类计算

特征选择：已有 D 维特征向量空间，Y={y1,y2,…,yD} ，从原有的 D 维特征空间，删去一些特征描述量，从而得到精简后的特征空间。

特征提取：找到一个映射关系 A:Y→X 使新样本特征描述维数比原维数降低。其中每个分量 xi 是原特征向量各分量的函数，即 Xi=fi(y1,y2,…,yD)

类别可分离性判据

可分性判据应满足的要求

（1）与错误率有单调关系，这使判据取最大值时错误率也较小

（2）当特征独立时有可加性：

$$J_{ij}(x_1,x_2,...,x_d)=\sum_{k=1}^{d}J_{ij}(x_k)$$

Jij 是第 i 类与第 j 类的可分性准则

（3）度量特性：

$$J_{ij}>0,i\neq j;J_{ij}=0,i=j;J_{ij}=J_{ji}$$

（4）单调性：加入新的特征时，判据不减小

$$J_{ij}(x_1,x_2,...,x_d)\le J_{ij}(x_1,x_2,...,x_d,x_{d+1})$$

基于距离的可分性判据

欧氏距离下的可分性判据

$$\delta(x_k^{(i)},x_l^{(j)})=(x_k^{(i)}-x_l^{(j)})^T(x_k^{(i)}-x_l^{(j)})$$

$$m_i=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k^{(i)}$$

$$m=\sum_{i=1}^{c}P_im_i$$

平均距离（判据）、

$$J_d(x)=\sum_{i=1}^{c}P_i[\frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^{(i)}-m_i)^T(x_k^{(i)}-m_i)+(m_i-m)^T(m_i-m)]$$

$$\tilde{S}_b=\sum_{i=1}^{c}P_i(m_i-m)(m_i-m)^T,\qquad \tilde{S}_w=\sum_{i=1}^{c}P_i\frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^{(i)}-m_i)(x_k^{(i)}-m_i)^T$$

判据的矩阵形式

$$J_d(x)=tr(\tilde{S}_b+\tilde{S}_w)$$

按欧氏距离度量的特征提取方法

基于距离可分性判据的特征优化过程是通过一个线性变换实现的。

设在原特征空间一个样本向量表示成 X(D 维)而在优化特征空间中，样本向量表示成 Y(d 维)而 X 与 Y 之间的关系是：

$Y=W^TX$

其中 W 是一个 D×d 维矩阵（d<D）

目的：利用判据找出一种线性变换 W，它可实现这种判据 J(Y)=J(W)的极值化。

J2 判据下的特征提取

$$J_2(x)=tr(S_w^{-1}S_b)$$

将原特征空间 X(D 维)通过线性映射 $Y=W^TX$ 降维到特征空间 Y 中，若 X 空间的类内离散度矩阵和类间离散度矩阵分别为 S_W ,Sb，则按 J2 判据最后特征提取矩阵 W 是按如下方式构造的：

若矩阵 $S_W^{-1}S_b$ 的本征值λi 按大小顺序列为

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_D$$

则选择前 d 个本征值所对应的本征向量组成变换矩阵 WD*d，都可使这些判据 J2(W)达到最大值。

证明：

因为： $Y=W^TX$，

设：X 的类内和类间离散度矩阵分别为 S_W ,Sb

则：Y 的类内和类间离散度矩阵分别为 S_W',Sb'为

$S_W'=W^TS_WW$, $Sb'=W^TSbW$

在使用 J2 判据下，将其 Y 的可分性判据表示成变换 W 的函数 $J2(Y)=tr[(SW')^{-1}Sb']$

则：

$J2(Y)=tr[(W^TS_WW)^{-1}(W^TSbW)]$

$=tr[W^TS_W^{-1}（W^T）^{-1}W^TSbW]$

$=tr[W^TS_W^{-1}SbW]$

设 $S_W^{-1}Sb$ 的本征值为λ1 >λ2 >λ3 >……>λD ，对应的本征向量矩阵为 U=[u1，u2,….,uD]

则 $U^TS_W^{-1}SbU=\Lambda$

可以证明：在不降维条件下，W 是 D*D 维的，

令 W=U=[u1，u2,….,uD]

则 J2 判据不变 J2(Y)= J2(X)

$J2(Y)=tr[W^TS_W^{-1}SbW]=tr[S_W^{-1}SbWW^T]=tr[S_W^{-1}Sb]=J2(X)$

则 $J2(W)=tr[W^TS_W^{-1}SbW]=tr[U^TS_W^{-1}SbU]=\sum_{i=1}^{D}\lambda_i=J2(Y)=J2(X)$

降维条件下

W=Ud =[u1，u2,….,ud]

$$J2(W)=tr[U_d^TS_W^{-1}S_bU_d]=tr\begin{bmatrix}\lambda_1&0&0&0\\0&\lambda_2&0&0\\0&0&\ddots&0\\0&0&0&\lambda_d\end{bmatrix}=\sum_{i=1}^{d}\lambda_i$$

设 $S_W^{-1}Sb$ 的本征值为λ1 >λ2 >λ3 >……>λD，那么由对应于 d 个最大的本征值的本征向量所组成的矩阵 W(D×d)，就能使所得到的 d 维特征满足 J2 判据最大的要求

利用 W 向量对原始的两类两维样本进行线性变换得到新的一维分布，特征空间从二维降到一维，并满足 J2 判据。该特征空间实质上就是对应于 Fisher 准则求得的线性分类器的法向量。

如果讨论的是多类别 C 问题，则优化后的维数至多为(C-1)

特征选择

特征选择在概念上十分简单，即对原有特征进行删选优化。

选择特征的标准：也就是选择前面讨论过的可分离性判据，以这些判据为准则，使所选择的 d 维子空间具有最大的可分离性。

要找出较好的特征选择方法，以在允许的时间内选择出一组最优的特征。所谓最优的特征组，就是要找到合适的特征的组合。如果从逐个特征配组进行性能比较的话，即穷举的算法，特征配组的数量可能极大，组合配置的数目按下式计算

$$q = C_D^d = \frac{D!}{(D-d)!d!}$$

任何非穷举的算法都不能保证所得结果是最优的，因此要得到最优解就必须待用穷举法。

“自上而下”是指，从 D 维特征开始，逐步将其中某些特征删除，直到剩下所要求的 d 维特征为止。

“自下而上”则是从零维特征空间开始，逐个地从 D 维持征中选择特征，直至达到预定的维数指标为止。

在选择的过程中，“自上而下”算法做到筛选剩下的特征组在每一步上都是最优的，而“自下而上”则在每一步都生成最优的特征空间。

最优搜索算法

“分支定界”算法：至今能得到最优解的唯一快速算法 。属于“自上而下”算法，但是具有回溯功能，可使所有可能的特征组合都被考虑到。其核心问题是通过合理组合搜索过程，可以避免一些计算而仍能得到最优的结果。关键是利用了判据的单调性。

分支定界算法虽然比盲目穷举法节省计算量，但计算量仍可能很大而无法实现，因此人们还是常用次优搜索法。

次优搜索法

（1）单独最优特征组合：这是一种最简单的方法，即将各特征按单独使用计算其判据值，然后取其前 d 个判据值最大的特征作为最优特征组合。这种做法的问题在于即使各特征是统计独立的，也不一定得到最优结果。

但如果可分性判据可写成如下形式，则用这种方法可以选出一组最优的特征。

$$J(X) = \sum_{i=1}^{D} J(x_i)$$

$$J(X) = \prod_{i=1}^{D} J(x_j)$$

（2）顺序前进法（SFS）：这是最简单的自下而上搜索方法。

首先计算每个特征单独进行分类的判据值，并选择其中判据值最大的特征，作为入选特征。然后每次从未入选的特征中选择一个特征，使得它与已入选的特征组合在一起时所得的 J 值为最大，直到特征数增至 d 个为止。

顺序前进法与单独特征最优化组合相比，一般说来，由于考虑了特征之间的相关性，在选择特征时计算与比较了组合特征的判据值，要比前者好些。

主要缺点是，一旦某一特征被选入，即使由于后加入的特征使它变为多余，也无法再把它剔除

推广至每次入选 r 个特征，而不是一个，称为广义顺序前进法。

（3）顺序后退法（SBS）：这是一种自上而下的方法。

从现有的特征组中每次减去一个不同的特征并计算其判据，找出这些判据值中之最大值，如此重复下去直到特征数达到予定数值 d 为止。

主要缺点是，一旦某一特征被剔除，无法再把它加入

与 SFS 相比，此法计算判据值是在高维特征空间进行的，因此计算量比较大。

此法也可推广至每次剔除 r 个，称为广义顺序后退法

（4）增 L 减 r 法

前面两种方法都有一个缺点，即一旦特征入选(或剔除)，过程不可逆转

其原理是对特征组在增加 L 个特征后，转入一个局部回溯过程，又用顺序后退法，剔除掉 r 个特征。这种方法既可能是“自上而下”方法，也可能是“自下而上”的，这取决于 L 与 r 的数据大小。

当 L>r 时，入选特征数逐渐增加，属“自下而上”型，反之属“自上而下”型。

K-L 变换进行特征提取

K-L 展开式的性质

（1）K-L 变换的展开系数是互相无关的

（2）K-L 变换后的协方差阵为对角阵，表明经过 K-L 变换后，原向量各分量之间存在的相关性已被消除

利用类均值向量提取特征

为了估计各分量（特征）对于分类的单独作用，先按类内离散度矩阵 Sw 作为产生矩阵产生相应的 K-L 坐标系统，从而把包含在原向量中各分量的相关性消除，并得到在新坐标系中各分量离散的程度。

$$u_j^T S_w u_j = \lambda_j$$

然后对均值向量在这些新坐标中分离的程度作出判断，决定在各坐标轴分量均值向量所能提供的相对可分性信息。据此选取 K-L 变换的基向量作为特征提取变换矩阵。

$$u_j^T S_b u_j = S_b^j$$

在 uj 轴上，原第 i 类特征向量 Xik 投影为： $X_{ik}^j = u_j^T X_{ik}$ 其类内离散度： $S_w{}^j = u_j{}^T S_w u_j = \lambda_j$ $S_w = \sum_i P_i \Sigma_i = \sum_i P_i \frac{1}{n_i} \sum_k (X_{ik} - m_i)(X_{ik} - m_i)^T$	类间离散度 $S_b{}^j = u_j{}^T S_b u_j$ $S_b = \sum_i P_i (m_i - m)(m_i - m)$

判据 J(Xi) ：为类间离散度与类内离散度在 uj 坐标的分量之比:

$$J(X_i) = \frac{S_b^j}{S_w^j} = \frac{u_j^T S_b u_j}{u_j^T S_w u_j} = \frac{u_j^T S_b u_j}{\lambda_j}$$

J(Xi)越大，表明在新坐标系中该坐标抽包含较多可分性信息。

为了降低特征空间的维数，可以将各分量按 大小重新排列，使 J(X1)≥ J(X2) ≥ J(X3) ….≥ J(XD)

取与前面 d 个最大的 J(Xi)≥值相对应的本征向量 uj，j=1,...,d; 作为特征空间的基向量 W=[u1,u2,…,ud]

基于类平均向量中判别信息的最优压缩

具体说来这种方法分成两步：

1）白化处理

1、先用原坐标系中 Sw 作为产生矩阵，实行 K-L 变换，将原有数据的相关性消除掉。(Y1=UTX)

所得到的 K-L 坐标系中的 $S'_w=\Lambda$ 是一个对角矩阵，其相应的 K-L 坐标系为 U，由原 Sw 本征值对应的本征向量组成。

即：$S_w U=U\Lambda$，或 $U^T S_w U=\Lambda$

2、进一步实行变换：$Y2=\Lambda^{-1/2} Y1$，

使 S'_w 矩阵变为单位矩阵：$Sw''=(\Lambda^{-1/2})^T Sw' \Lambda^{-1/2}=(\Lambda^{-1/2})^T U^T Sw\ U\ \Lambda^{-1/2}=\Lambda^{-1/2}\Lambda\Lambda^{-1/2}=I$

即 $(U\Lambda^{-1/2})^T S_w (U\Lambda^{-1/2})=I$

3、上式中，令: $\boldsymbol{B=U\Lambda^{-1/2}}$（白化矩阵），则：$B^T S_w B=I$（白化）

此时， $Y2=\Lambda^{-1/2} Y1=\Lambda^{-1/2} U^T X=(U\Lambda^{-1/2})^T X=B^T X$

经过 B 变换后的类间离散度矩阵 S'b 应有 $\boldsymbol{S'b=B^T S_b B}$

2）特征提取

采用上节方法求最佳的变换 $S_w' U=U\Lambda$

因 $S'_w=I$,则$\Lambda=I$

而且 U 可以是任何正交矩阵。即 $U^T I U=U^T U=I$

$$J(X_i)=\frac{S_b^j}{S_w^j}=\frac{u_j^T S_b' u_j}{u_j^T S_w' u_j}=\frac{u_j^T S_b' u_j}{1}=u_j^T S_b' u_j$$

由于 U 可以是任何正交矩阵，因此 S'b 可作为产生矩阵，作第二次 K-L 变换，得到正交矩阵

$\boldsymbol{Sb'V=\Lambda'V}$

$$J(X_i)=u_j^T S'_b u_j=\lambda_j'$$

因此，按判据最大选择 uj，就是按 Sb' 正交分解的本征值λ'j 最大选择。

由于 S'b 的秩最多是 ***c-1***，所以最多只有 c-1 个非零本征值。选择最大的 d（≤c-1）个非零本征值 ，则该 d 个非零本征值就可表示类均值向量所包含的全部信息。设这 d 个本征向量系统用 V 表示，

即 V=[v1,v2,….,vd] （d≤c-1）

则变换为:$Y=V^T Y2=V^T B^T X=(BV)^T X$

因此变换矩阵为：$\boldsymbol{W=BV}=U\Lambda-1/2V$

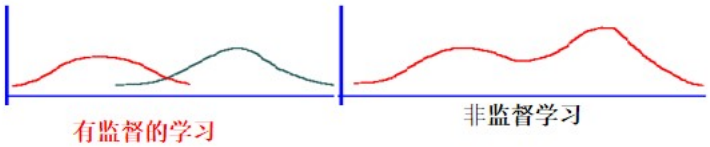
第五章 非监督学习法

以前讨论的分类器设计方法都是在样本集中的类别标签已知的条件下进行的，这些样本称为训练样本。

在样本类别标签已知的情况下，可以统计出各类训练样本不同的描述量，如其概率分布，或在特征空间分布的区域等，利用这些参数进行分类器设计，称为有监督的学习方法。

然而在实际应用中，不少情况下无法预先知道样本的标签，也就是说没有训练样本，因而只能从没有样本标签的样本集进行分类器设计，这就是非监督学习方法。

有监督学习中，样本集分布呈现交迭情况，而无监督学习方法由于没有类别样本指导，无法确定它们的交迭情况，只能按分布的聚类情况进行划分



非监督学习与有监督学习的不同点

1． 有监督学习方法必须要有训练集与测试样本。在训练集中找规律，而对测试样本使用这种规律；而非监督学习没有训练集，只有一组数据，在该组数据集内寻找规律。

2． 有监督学习方法的目的就是识别事物。识别的结果表现在给待识别数据加上了标号，因此训练样本集必须由带标号的样本组成。

非监督学习方法的目的是寻找数据集中的规律性。不一定要达到划分数据集的目的，也就是说不一定要“分类”。只有要分析的数据集本身，预先没有什么标号。如果发现数据集呈现某种聚集性，则可按自然的聚集性分类，但不以与某种预先的分类标号对上号为目的。

非监督学习方法可以分成两大类

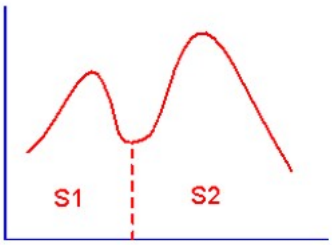
一类为基于概率密度函数估计的直接方法，指设法找到各类别在特征空间的分布参数再进行分类。

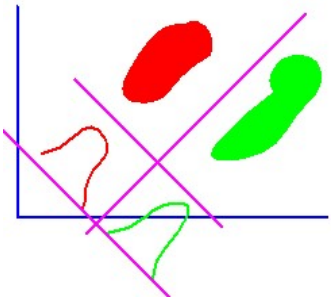
另一类称为基于样本间相似性度量的间接聚类方法，其原理是设法定出不同类别的核心或初始类核，然后依据样本与这些核心之间的相似性度量将样本聚集成不同类别。

单峰子类的分离方法

样本概率密度分布在特征空间的分布是多峰的。

每个单峰区域则被看作不同的决策域。落在同一单峰区域的待分类样本就被划分成同一类，称为单峰子类。





高维空间寻找概率密度的“峰”是困难的。

一维空间寻找概率密度的“峰”较容易。

因此，可通过将高维空间样本投影到不同的一维空间 ui 上，$xi=ui^{T}Y$；然后，在此一维空间上估计边缘概率密度 p(xi)，并在此概率密度上寻找各个峰，并确定每个峰的范围（即每个聚类），各个聚类的分解面与该坐标轴 ui 垂直，交点则是两个峰值之间的最小点。称为***投影法***

动态聚类方法

(1)C 均值算法

准则函数—误差平方和准则

这个准则函数是以计算各类样本到其所属类均值点误差平方和为准则。

$$J_c = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \| y - m_i \|^2$$

最佳的聚类是使 Jc 为最小的分类。这种类型的聚类通常称为最小方差划分。

如果原属Γk 中的一个样本 y 从Γk 移入Γj 时，它会对误差平方和产生影响, Γk 类在抽出样本 y 后其相应均值为

$$\tilde{m}_k = m_k + \frac{1}{N_k - 1}[m_k - y]$$

而样本 y 新加盟的Γj 集合均值

$$\tilde{m}_j = m_j + \frac{1}{N_j+1}[y - m_j]$$

由于 y 的移动只影响到 k 与 j 这两类的参数改动，因此，计算 Jc 值的变动只要计算相应两类误差平方和的变动即可，此时

$$\tilde{J}_k = J_k - \frac{N_k}{N_k-1}\|y-m_k\|^2 \qquad \tilde{J}_j = J_j + \frac{N_j}{N_j+1}\|y-m_j\|^2$$

总误差变化：

$$\Delta J_e = (\tilde{J}_k - J_k) + (\tilde{J}_j - J_j) = -\frac{N_k}{N_k-1}\|y-m_k\|^2 + \frac{N_j}{N_j+1}\|y-m_j\|^2$$

$$\frac{N_k}{N_k-1}\|y-m_k\|^2 > \frac{N_j}{N_j+1}\|y-m_j\|^2$$

$$\Delta J_e < 0$$

即将样本 y 从Γk 移入至Γj 就会使误差平方总和 Jc 减小，它表明样本变动是合乎准则要求的

算　法

(1) 选择某种方法把样本分成 C 个聚类的初始划分，计算每个聚类的均值 m1,…,mc 和 Jc

(2) 从第一个样本开始，按顺序选择一个备选样本 y，设其在 wi 中。

(3) 若 Ni=1，则转(2)（样本只有 1 个，不移出），否则继续下一步。

(4) 计算

$$e_j = \begin{cases} \dfrac{N_j}{N_j+1}\|y-m_j\|^2 & , \quad j \neq i \\ \dfrac{N_i}{N_i-1}\|y-m_i\|^2 & , \quad j = i \end{cases}$$

(5) 若 ek=min(ej)<ei, 则将 y 从 wi 移到 wk 中。（否则， ei=min(ej), 不用移。转（7））

(6) 重新计算 mi 和 mk，并修改 Jc 。

$$J'c = Jc + \Delta J_e = Jc - \frac{N_k}{N_k-1}\|y-m_k\|^2 + \frac{N_j}{N_j+1}\|y-m_j\|^2$$

(7)选择下一个样本。重复（3）-（7）。直到最后一个样本。

(8)重复（2)-(7)。连续迭代 N 次(即所有样本都运算过) Jc 不变，则停止，否则转到 2。

C—均值算法比较简单，但它的自我调整能力也比较差。这主要表现在类别数不能改变，受代表点初始选择的影响也比较大。

(2)ISODATA 算法

全称‘迭代自组织数据分析技术’（Iterative Self-Organizing Data Analysis Technique　Algorithm）。

ISODATA 算法的功能与 C—均值算法相比的改进。

1. 不是每调整一个样本的类别就重新计算一次各类均值（逐个样本修正），而是每次把全部样本都调整完毕后再重新计算样本均值（成批样本修正） 。

2.考虑了类别的合并与分裂，因而有了自我调整类别数的能力。从而可以得到较为合理的类别数。

分级聚类方法

分级聚类方法的目的并不把 N 个样本分成某一个预定的类别数 C，而是把样本集按不同的相似程度要求分成不同类别的聚类。

合并：开始时，每个样本自成一类，然后通过合并类别减少类

分裂：开始时，所有样本是一类，然后通过分裂类别样本增加类

基于合并的分级聚类算法

初始时设置Γj， j ∈I, I={j|j=1,2,…,N}及距离阈值 d（可以取无穷大）。 Γj 表示各个聚类集合，N 是样本数，初始时每个样本自成一类。

步骤 1： 在集合Γj， j ∈I 中找到一对满足下列条件的聚类集合Γi 与Γk。

$\Delta(\Gamma i,\Gamma j)=\min\{\Delta(\Gamma i,\Gamma j), i, j \in I\}$

步骤 2： 若距离超过设定的阈值 d，则算法终止；

否则，把Γi 并入Γk ，并去掉Γi。

步骤 3： 把 i 从指标集 I 中除掉，若 I 的基数仅等于 2 时，则终止计算，否则转向步骤 1。

最近距离作相似性度量时聚类结果

1）对长条形分布聚类正确，可以出现细长条的类。

2）聚类结果对噪声或数据点波动非常敏感。

最远距离作相似性度量时聚类结果

1）分布紧密的两类，交叠区无点，正确聚类

2）分布紧密的两类，但交叠区有点（p1,p2），也能正确聚类

3）长条形分布：聚类错误（说明个别远离点对聚类结果影响大，该度量方法不能得到长条形的聚类

不同聚类算法的比较

单峰子集类的分离法（直接法）

优点：样本数多时，反映数据的概率结构，结果能反映数据构造的真实情况

缺点：1）要对概率密度函数进行估计，计算的工作量大。

2）在进行概率估计时要选定一些参数，因此估计的结果也会受到参数选择的较大影响。

3）是在有噪声的情况下，具有局部最大值的概率密度函数的峰点都会发生变化，从而不能正确反映数据中的单峰子集数。

4）样本数较少的情况下，由于没有可能对概率密度函数进行估计。这种方法完全失去意义。

间接的动态聚类算法

优点：一般情况下计算效率很高

缺点：选定的模型常常不能反映数据的概率结构，因此用这些方法得到的结果不能反映数据构造的真实情况。只有通过选择各种各样的核函数以及分析这些核函数所得所到的聚类结果来部分地解决这个问题。

分级聚类算法

样本数较少的情况下是特别有用