聚类分析

相似性的测度

- **欧氏距离**

  设 $\boldsymbol{X}_1, \boldsymbol{X}_2$ 为两个 n 维模式样本，$\boldsymbol{X}_1 = [x_{11}, x_{12}, \cdots, x_{1n}]^T, \boldsymbol{X}_2 = [x_{21}, x_{22}, \cdots, x_{2n}]^T$

  $D(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| = \sqrt{(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T(\boldsymbol{X}_1 - \boldsymbol{X}_2)} = \sqrt{(x_{11} - x_{21})^2 + \cdots + (x_{1n} - x_{2n})^2}$

  使特征数据标准化，使其与变量的单位无关

- **马氏距离**

  $D(\boldsymbol{X}) = \sqrt{(\boldsymbol{X} - \boldsymbol{M})^T \boldsymbol{C}^{-1} (\boldsymbol{X} - \boldsymbol{M})}$

  $\boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ 为模式向量，$\boldsymbol{M} = \begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix}$ 为均值向量，$\boldsymbol{C}$ 为协方差矩阵

  $\boldsymbol{C} = E\left\{(\boldsymbol{X} - \boldsymbol{M})(\boldsymbol{X} - \boldsymbol{M})^T\right\}$

  表示的概念是各分量上模式样本到均值的距离，也就是在各维上模式的分散情况

  $d(\boldsymbol{X}, \boldsymbol{M}) = \sqrt{\left(\frac{x_1 - m_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - m_2}{\sigma_2}\right)^2 + \cdots + \left(\frac{x_n - m_n}{\sigma_n}\right)^2}$

  $d(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{(\boldsymbol{X} - \boldsymbol{Y})^T \boldsymbol{C}^{-1} (\boldsymbol{X} - \boldsymbol{Y})} = \sqrt{\left(\frac{x_1 - y_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - y_2}{\sigma_2}\right)^2 + \cdots + \left(\frac{x_n - y_n}{\sigma_n}\right)^2}$

- **明氏距离**

  $D_m(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left[\sum_{k=1}^{n} |x_{ik} - x_{jk}|^m\right]^{\frac{1}{m}}$

  $m = 1$ 时，$D_1(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$ 称为街区距离

  $m = 2$ 时，为欧氏距离

- **汉明距离**

  设 $\boldsymbol{X}_i, \boldsymbol{X}_j$ 为 n 为二值 $(1, -1)$ 模式样本向量

  $D_h(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{1}{2}\left(n - \sum_{k=1}^{n} x_{ik} \cdot x_{jk}\right)$

- **角度相似性函数**

  $S(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{{\boldsymbol{X}_i}^T \boldsymbol{X}_j}{\|\boldsymbol{X}_i\| \cdot \|\boldsymbol{X}_j\|}$

  为模式向量 $\boldsymbol{X}_i, \boldsymbol{X}_j$ 之间夹角的余弦

聚类准则

- **阈值准则**
- **函数准则**

  误差平方和 $J = \sum_{j=1}^{c} \sum_{\boldsymbol{X} \in S_j} \|\boldsymbol{X} - \boldsymbol{M}_j\|^2$

- $\{\boldsymbol{X}\}$ 为模式样本集
- $\{S_j, j=1,2,\cdots,c\}$ 为模式类别
- c 为聚类类别数目
- $\boldsymbol{M}_j = \frac{1}{N_j}\sum\limits_{\boldsymbol{X}\in S_j}\boldsymbol{X}$ 为属于 S_j 集的样本的均值向量
- N_j 为 S_j 中样本的数目

J 代表了分属于 c 个聚类类别的全部模式样本与其相应类别模式均值之间的误差平方和

适用于样本类数给定，各类样本密集且数目相差不多，而不同类间的样本又明显分开的情况

基于距离阈值的聚类算法

- **近邻聚类法**

  有 N 个待分类的模式 $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，要求按距离阈值 T 分类到以 $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots$ 为聚类中心的模式类中

  - 任取样本 $\boldsymbol{X}_i$ 作为第一个聚类中心的初始值，如 $\boldsymbol{Z}_1 = \boldsymbol{X}_1$
  - 计算样本 $\boldsymbol{X}_2$ 到 $\boldsymbol{Z}_1$ 的欧氏距离 $D_{21} = \|\boldsymbol{X}_2 - \boldsymbol{Z}_1\|$，若 $D_{21} > T$ 则定义一新的聚类中心 $\boldsymbol{Z}_2 = \boldsymbol{X}_2$，否则 $\boldsymbol{X}_2 \in$ 最近的聚类中心

  很大程度上依赖于距离阈值T的大小、第一个聚类 中心的位置选择、待分类模式样本的排列次序、以及样本分布的几何性质

- **最大最小距离算法**

  - 任取样本 $\boldsymbol{X}_i$ 作为第一个聚类中心的初始值，如 $\boldsymbol{Z}_1 = \boldsymbol{X}_1$
  - 选取离 $\boldsymbol{Z}_1$ 最远的样本作为第二聚类中心 $\boldsymbol{Z}_2$
  - 逐个计算各模式样本与已确定的所有聚类中心之间的距离，并选出其中的最小距离 $\min\limits_k(\|\boldsymbol{X}_i - \boldsymbol{Z}_k\|)$
  - 当最大值 $\max\limits_i \min\limits_k(\|\boldsymbol{X}_i - \boldsymbol{Z}_k\|)$ 达到 $\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|$ 的一定分数比值（阈值 T）以上，则选取新的聚类中心

层次聚类法

- **算法描述**
  - N 个初始模式样本自成一类 $G_1(0), G_2(0), \cdots, G_N(0)$

    计算各类之间的距离，得 $N \times N$ 维矩阵 $\boldsymbol{D}(0)$
  - 找出 $\boldsymbol{D}(n)$ 中的最小元素，将对应的两类合并为一类，建立新的分类 $G_1(n+1), G_2(n+1), \cdots$
  - 计算合并后新类别的距离 $\boldsymbol{D}(n+1)$
  - 当 $\boldsymbol{D}(n)$ 的最小分类超过阈值 T 或全部样本聚成一类时，算法停止
- **类间距离计算**
  - 最短距离法

    设 H, K 为两个聚类，距离 $D_{HK} = \min\{D(\boldsymbol{X}_H, \boldsymbol{X}_K)\}, \boldsymbol{X}_H \in H, \boldsymbol{X}_K \in K$

    若 K 由 I, J 合并而成，则 $D_{HK} = \min\{D_{HI}, D_{HJ}\}$
  - 最长距离法

    $D_{HK} = \max\{D(\boldsymbol{X}_H, \boldsymbol{X}_K)\}, \boldsymbol{X}_H \in H, \boldsymbol{X}_K \in K$

    $D_{HK} = \max\{D_{HI}, D_{HJ}\}$

- 重心法
- 类平均距离法

动态聚类法

- **K-means**

  设待分类的模式特征矢量集为 $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，类的数目 K 事先取定。S_j 为第 j 个聚类集，聚类中心为 Z_j，包含的样本数为 N_j

  - 任选 K 个模式特征向量 $\boldsymbol{Z}_1(1), \boldsymbol{Z}_2(1), \cdots, \boldsymbol{Z}_K(1)$ 作为初始聚类中心
  - 若 $D_j(k) = \min\{\|\boldsymbol{X} - \boldsymbol{Z}_i(k)\|\}, i = 1, 2, \cdots, K$，则 $\boldsymbol{X} \in S_j(k)$
  - $\boldsymbol{Z}_j(k+1) = \frac{1}{N_j} \sum\limits_{\boldsymbol{X} \in S_j(k)} \boldsymbol{X}, j = 1, 2, \cdots, K$

    $J_j = \sum\limits_{\boldsymbol{X} \in S_j(k)} \|\boldsymbol{X} - \boldsymbol{Z}_j(k+1)\|^2, j = 1, 2, \cdots, K$
  - 若 $\boldsymbol{Z}_j(k+1) = \boldsymbol{Z}_j(k), j = 1, 2, \cdots, K$，则结束

    否则 $K = K + 1$

- **ISODATA**

  - 确定控制参数、设置代表点

符号	内容
K	聚类期望数
θ_N	一个聚类的最少样本数
θ_C	类间距离控制参数
θ_S	标准偏差控制参数
L	每次迭代允许合并的最大聚类对数
I	允许迭代的次数
N_C	初始设定聚类数，聚类中心为 $\boldsymbol{Z}_i, i = 1, 2, \cdots, N_C$

  - 分类

    若 $D_j = \min\{\|\boldsymbol{X} - \boldsymbol{Z}_i\|\}, i = 1, 2, \cdots, N_C$，则 $\boldsymbol{X} \in S_j$
  - 撤销类内样本数过小类别

    若有样本 $S_j, N_j < \theta_N$，则舍去 S_j，令 $N_C = N_C - 1$，样本分配至其他类别
  - 更新均值向量

    $\boldsymbol{Z}_j = \frac{1}{N_j} \sum\limits_{\boldsymbol{X} \in S_j} \boldsymbol{X}, j = 1, 2, \cdots, N_C$
  - 计算类内平均距离

    $\tilde{D}_j = \frac{1}{N_j} \sum\limits_{\boldsymbol{X} \in S_j} \boldsymbol{X} - \boldsymbol{Z}_j, j = 1, 2, \cdots, N_C$

- 计算全部样本集到响应均值的平均距离

  $\tilde{D}=\frac{1}{N}\sum_{j=1}^{N_C}N_j\tilde{D}_j$

- 入口选择，判断分裂、合并

  若迭代次数为 I 则合并

  若 $N_C<\frac{K}{2}$，则分裂

  若迭代了偶数次，或者 $N_C\geq 2K$，则合并

  其余分裂

- 分裂
  - 求各聚类标准差

    $\sigma_j=[\sigma_{j1}\quad\sigma_{j2}\quad\cdots\quad\sigma_{jn}]^T$

    $\sigma_{ji}=\sqrt{\frac{1}{N_j}\sum_{\boldsymbol{X}_k\in S_j}(x_{ki}-z_{ji})^2}$
  - 求 $\sigma_{j\max},j=1,2,\cdots,C$
  - 执行分裂

    若 $\sigma_{j\max}>\theta_S$，且有 $\tilde{D}_j>\tilde{D},N_j>2\theta_N+1$ 或者 $N_C\leq\frac{K}{2}$，则把 S_j 分裂为两个聚类 $\boldsymbol{Z}_j^+,\boldsymbol{Z}_j^-$，$N_C=N_C+1$

    给定 $k,0<k<1$，令 $r_j=k\sigma_{j\max}$，$\boldsymbol{Z}_J^+=\boldsymbol{Z}_j+r_j,\boldsymbol{Z}_j^-=\boldsymbol{Z}_j-r_j$
- 合并
  - 计算聚类中心距离

    $D_{ij}=\|\boldsymbol{Z}_i-\boldsymbol{Z}_j\|,i=1,2,\cdots,N_C-1,j=i+1,\cdots,N_C$
  - 列出类间距离过近者

    $D_{i_1j_1}<D_{i_2j_2}<\cdots<D_{i_lj_l},l\leq L$
  - 执行合并

    将 $\boldsymbol{Z}_{i_l},\boldsymbol{Z}_{j_l}$ 合并得到 $\boldsymbol{Z}_l=\frac{1}{N_{i_l}+N_{j_l}}[N_{i_l}\boldsymbol{Z}_{i_l}+N_{j_l}\boldsymbol{Z}_{j_l}]$，$N_C=N_C-1$
- 结束步骤

  若为最后一次迭代则中止，否则迭代次数 +1

- **DBSCAN**
  - 核心点：在半径 Eps 内含有超过 MinPts 数目的点

    边界点：在半径 Eps 内点的数量小于 MinPts，但是在核心点的邻居

    噪音点：任何不是核心点或边界点的点
  - 直接密度可达：给定一个对象集合 D，如果 p 在 q 的 Eps 邻域内，而 q 是一个核心对象，则称对象 p 从对象 q 出发时是直接密度可达的

    密度可达：如果存在一个对象链 $p_1,p_2,\cdots,p_n,p_1=q,p_n=p$，对于 $p_i\in D,1\leq i\leq n$，p_{i+1} 是从 p_i 关于 Eps 和 MinPts 直接密度可达的，则对象 p 是从对象 q 关于 Eps 和 MinPts 密度可达的

    密度相连：如果存在对象 $O\in D$，使对象 p 和 q 都是从 O 关于 Eps 和 MinPts 密度可达的，那么对象 p 到 q 是关于 Eps 和 MinPts 密度相连

聚类结果的评价

- 聚类中心之间的距离：距离值大，通常可考虑分为不同类
- 聚类域中的样本数目：样本数目少且聚类中心距离远，可考虑是否为噪声
- 聚类域内样本的距离方差：方差过大的样本可考虑是否属于这一类

线性分类器

判别函数

- **定义**
  - 直接用来对模式进行分类的准则函数
  - 若分属于 ω_1, ω_2 的两类模式可用一方程 $d(\boldsymbol{X}) = 0$ 来划分，那么称为判 $d(\boldsymbol{X}) = 0$ 别函数，或称判决函数、决策函数
- **前提条件**
  - 样本所属类别未知
  - 样本类别数已知

线性判别函数

- **一般形式**

  $d(\boldsymbol{X}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} = \boldsymbol{W}_0^T \boldsymbol{X} + w_{n+1}$

  其中，$\boldsymbol{X} = [x_1, x_2, \cdots, x_n]^T$，$\boldsymbol{W}_0$ 为权向量

  增广形式：$d(\boldsymbol{X}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} = \boldsymbol{W}^T \boldsymbol{X}$

  其中，$\boldsymbol{X} = [x_1, x_2, \cdots, x_n, 1]^T$
- **分类方法**
  - 两类情况

    $$d(\boldsymbol{X}) = \boldsymbol{W}^T \boldsymbol{X} \begin{cases} > 0 & \boldsymbol{X} \in \omega_1 \\ < 0 & \boldsymbol{X} \in \omega_2 \end{cases}$$

    $d(\boldsymbol{X}) = 0$：不可判，可以 $\boldsymbol{X} \in \omega_1$ 或 $\boldsymbol{X} \in \omega_2$ 或拒绝
  - $\omega_i / \overline{\omega_i}$ 两分法

    $$d(\boldsymbol{X}) = \boldsymbol{W}^T \boldsymbol{X} \begin{cases} > 0 & \boldsymbol{X} \in \omega_i \\ < 0 & \boldsymbol{X} \in \overline{\omega_i} \end{cases} \quad i = 1, 2, \cdots, M$$

    若只有 $d_i(\boldsymbol{X}) > 0$，其余 $d(\boldsymbol{X}) < 0$，则判为 ω_i 类

    否则为失效区域（IR）
  - ω_i / ω_j 两分法

    $d_{ij}(\boldsymbol{X}) = \boldsymbol{W}_{ij}^T \boldsymbol{X}$

    $d_{ji} = -d_{ij}$

    若 $d_{ij}(\boldsymbol{X}) > 0, \forall j \neq i, i, j = 1, 2, \cdots, M$，则 $\boldsymbol{X} \in \omega_i$

    否则为 IR 区
  - ω_i / ω_j 两分法特例

当 $d_{ij}(\boldsymbol{X})=d_i(\boldsymbol{X})-d_j(\boldsymbol{X})$

若 $d_i(\boldsymbol{X})=\max\{d_k(\boldsymbol{X}),k=1,2,\cdots,M\}$，则 $\boldsymbol{X}\in\omega_i$

除边界区外没有不确定区域

广义线性判别函数

- **非线性多项式函数**

  设 $\{\boldsymbol{X}\}$ 在模式空间 $\mathbb{X}$ 中线性不可分

  $$d(\boldsymbol{X})=w_1f_1(\boldsymbol{X})+w_2f_2(\boldsymbol{X})+\cdots+w_kf_k(\boldsymbol{X})+w_{k+1}=\sum_{i=1}^{k+1}w_if_i(\boldsymbol{X})$$

  其中，$f_{k+1}(\boldsymbol{X})=1$

- **数学表达式**

  $\boldsymbol{X}^*=[x_1^*,x_2^*,\cdots,x_k^*,1]^T=[f_1(\boldsymbol{X}),f_2(\boldsymbol{X}),\cdots,f_k(\boldsymbol{X}),1]^T$

  $\mathbb{X}^*$ 空间的维数 k 高于 $\mathbb{X}$ 的维数 n

  $d(\boldsymbol{X})=\boldsymbol{W}^T\boldsymbol{X}^*=d(\boldsymbol{X}^*),\boldsymbol{W}=[w_1,w_2,\cdots,w_k,w_{k+1}]^T$

线性判别函数几何性质

- **模式空间与超平面**
  - 基本概念

    模式空间：以 n 维模式向量 $\boldsymbol{X}$ 的 n 个分类为坐标变量的欧氏空间

    超平面：$d(\boldsymbol{X})=\boldsymbol{W}_0^T\boldsymbol{X}+w_{n+1}=0$
  - 法向量

    $\boldsymbol{W}_0^T(\boldsymbol{X}_1-\boldsymbol{X}_2)=0$

    $\boldsymbol{W}_0$ 与超平面上任意向量正交，称为超平面的法向量，方向由超平面的负侧指向正侧

    单位法线向量 $\boldsymbol{U}=\frac{\boldsymbol{W}_0}{\|\boldsymbol{W}_0\|}$
  - $d(\boldsymbol{X})$ 取值

    $\boldsymbol{X}$ 在超平面上时，取值为 0

    否则，将 $\boldsymbol{X}$ 向超平面投影得到向量 $\boldsymbol{X}_p$，构造向量 $\boldsymbol{R}=r\cdot\boldsymbol{U}$，其中 r 为 $\boldsymbol{X}$ 到超平面的代数距离

    $\boldsymbol{X}=\boldsymbol{X}_p+\boldsymbol{R}$

    $d(\boldsymbol{X})=r\|\boldsymbol{W}_0\|\begin{cases}>0 & \text{超平面正侧}\\<0 & \text{超平面负侧}\end{cases}$

- **权空间和权向量**
  - 权空间

    以线性判别函数的权值为坐标变量的 $n+1$ 维欧氏空间
  - 规范化增广样本向量

    对于两类问题：$\omega_1=\{x_{11},x_{12},\cdots,x_{1p}\},\omega_2=\{x_{21},x_{22},\cdots,x_{2q}\}$，令

    $\boldsymbol{X}=\begin{cases}X_{1i} & i=1,2,\cdots,p\\-X_{2i} & i=1,2,\cdots,q\end{cases}$

    则称 $\boldsymbol{X}$ 为规范化增广样本向量

若 $p+q$ 个样本的判别函数均大于零，则为解区

感知器算法

- 选择 N 个分属于 ω_1 和 ω_2 的模式识别样本构成训练样本集，编号为 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N$，任取权向量初始值 $\boldsymbol{W}(1)$ 开始迭代
- 第 k 次迭代时，输入一个样本 $\boldsymbol{X}_i$

  $$\boldsymbol{W}(k+1)=\begin{cases}\boldsymbol{W}(k) & \boldsymbol{W}^T(k)\boldsymbol{X}_i>0\\ \boldsymbol{W}(k+1)=\boldsymbol{W}(k)+c\boldsymbol{X}_i & \boldsymbol{W}^T(k)\boldsymbol{X}_i\le 0\end{cases}$$

  c 为校正增量系数，$c>0$
- 若上一轮迭代未发生错误，则迭代结束

梯度法

- 选择 N 个分属于 ω_1 和 ω_2 的模式识别样本构成训练样本集，编号为 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N$，准则函数为 $J(\boldsymbol{W},\boldsymbol{X})$，任取权向量初始值 $\boldsymbol{W}(1)$ 开始迭代
- 第 k 次迭代时，输入一个样本 $\boldsymbol{X}_i$

  $\nabla J(k)=\frac{\partial J(\boldsymbol{W},\boldsymbol{X}_i)}{\partial \boldsymbol{W}}\Big|_{\boldsymbol{W}=\boldsymbol{W}(k)}$

  $\boldsymbol{W}(k+1)=\boldsymbol{W}(k)-c\nabla J(k)$
- 若对所有样本均有 $\nabla J=0$，迭代结束

最小平方误差法（LMSE)

选择准则函数 $J(\boldsymbol{W},\boldsymbol{X},\boldsymbol{B})=\frac{1}{2}\|\boldsymbol{XW}-\boldsymbol{B}\|^2=\frac{1}{2}\sum\limits_{i=1}^{N}(\boldsymbol{W}^TX_i-b_i)^2$，使得当 J 达到最小时，$\boldsymbol{XW}=\boldsymbol{B}$ 可得到最小二乘近似解

- 选择 N 个分属于 ω_1 和 ω_2 的模式识别样本构成训练样本集，将属于 ω_2 的样本乘以 -1，写出增广样本矩阵 $\boldsymbol{X}$
- 求 $\boldsymbol{X}$ 的伪逆矩阵 $\boldsymbol{X}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- 设置初值 c 和 $\boldsymbol{B}(1)$，c 为正的校正向量，$\boldsymbol{B}(1)$ 各分量大于零

  计算 $\boldsymbol{W}(1)=\boldsymbol{X}^{\#}\boldsymbol{B}(1)$
- 计算 $\boldsymbol{e}(k)=\boldsymbol{XW}(k)-\boldsymbol{B}(k)$
  - 若 $\boldsymbol{e}(k)=0$，则线性可分，解为 $\boldsymbol{W}(k)$，迭代结束
  - 若 $\boldsymbol{e}(k)>0$，则线性可分，进入下一步
  - 若 $\boldsymbol{e}(k)<0$，检查 $\boldsymbol{XW}(k)$。若 $\boldsymbol{XW}(k)>0$，有解；否则无解，迭代结束
- 计算 $\boldsymbol{W}(k+1)$ 和 $B(k+1)$
  - 方法1：计算 $\boldsymbol{W}(k+1)=\boldsymbol{W}(k)+c\boldsymbol{X}^{\#}|\boldsymbol{e}(k)|$，
    $\boldsymbol{B}(k+1)=\boldsymbol{B}(k)+c\left[\boldsymbol{e}(k)+|\boldsymbol{e}(k)|\right]$
  - 方法2：计算 $\boldsymbol{B}(k+1)=\boldsymbol{B}(k)+c[\boldsymbol{e}(k)+|\boldsymbol{e}(k)|]$，$\boldsymbol{W}(k+1)=\boldsymbol{X}^{\#}\boldsymbol{B}(k+1)$

  然后返回上一步

势函数法

- 势函数
  - Ⅰ型势函数：用对称的有限多项式展开，$K(\boldsymbol{X},\boldsymbol{X}_k)=\sum\limits_{i=1}^{m}\varphi_i(\boldsymbol{X}_k)\varphi_i(\boldsymbol{X})$

    $\varphi_i(\boldsymbol{X})$ 在模式定义域内为正交函数集
  - Ⅱ型势
  - 函数：双变量的对称函数 $K(\boldsymbol{X},\boldsymbol{X}_k)=K(\boldsymbol{X}_k,\boldsymbol{X})$
- 积累势函数初始值 $K_0(\boldsymbol{X})=0$
- 加入 $\boldsymbol{X}_1$

  $K_1(\boldsymbol{X})=\begin{cases}K_0(\boldsymbol{X})+K(\boldsymbol{X},\boldsymbol{X}_1) & \boldsymbol{X}_1\in\omega_1\\K_0(\boldsymbol{X})-K(\boldsymbol{X},\boldsymbol{X}_1) & \boldsymbol{X}_1\in\omega_2\end{cases}$
- 加入 $\boldsymbol{X}_{k+1}$

  $K_{k+1}(\boldsymbol{X})=K_k(\boldsymbol{X})+r_{k+1}K(\boldsymbol{X},\boldsymbol{X}_{k+1})$

  $r_{k+1}=\begin{cases}0 & \boldsymbol{X}_{k+1}\in\omega_1,K_k(\boldsymbol{X}_{k+1})>0\\0 & \boldsymbol{X}_{k+1}\in\omega_2,K_k(\boldsymbol{X}_{k+1})<0\\1 & \boldsymbol{X}_{k+1}\in\omega_1,K_k(\boldsymbol{X}_{k+1})\le 0\\-1 & \boldsymbol{X}_{k+1}\in\omega_2,K_k(\boldsymbol{X}_{k+1})\ge 0\end{cases}$

贝叶斯判别准则

研究对象及相关概率

- 条件概率常用公式
  - 概率乘法公式 $P(AB)=P(A)P(B\mid A)$
  - 全概率公式 $P(B)=\sum P(A_i)P(B\mid A_i)$
  - 贝叶斯公式 $P(A_i\mid B)=\frac{P(A_i)P(B|A_i)}{\sum P(A_i)P(B|A_i)}$
- 模式识别的概率
  - 先验概率 $P(\omega_i)$
  - 后验概率 $P(\omega_i\mid\boldsymbol{X})$
  - 先验概率 $P(\boldsymbol{X}\mid\omega_i)$

  $P(\omega_i\mid\boldsymbol{X})=\frac{P(\boldsymbol{X}|\omega_i)P(\omega_i)}{P(\boldsymbol{X})}=\frac{P(\boldsymbol{X}|\omega_i)P(\omega_i)}{\sum\limits_{i=1}^{M}P(\boldsymbol{X}|\omega_i)P(\omega_i)}$

贝叶斯决策

- 最小错误率贝叶斯决策
  - $P(\omega_i\mid\boldsymbol{X})=\max\{P(\omega_j\mid\boldsymbol{X})\}$，则 $X\in\omega_i$
  - $l_{12}(\boldsymbol{X})=\dfrac{P(\boldsymbol{X}\mid\omega_1)}{P(\boldsymbol{X}\mid\omega_2)}\begin{matrix}>\\<\end{matrix}\dfrac{P(\omega_2)}{P(\omega_1)}$，则 $\boldsymbol{X}\in\begin{cases}\omega_1\\\omega_2\end{cases}$

    $P(\boldsymbol{X}\mid\omega_1)P(\omega_1)\begin{matrix}>\\<\end{matrix}P(\boldsymbol{X}\mid\omega_2)P(\omega_2)$，则 $\boldsymbol{X}\in\begin{cases}\omega_1\\\omega_2\end{cases}$
- 最小风险贝叶斯决策

- 条件平均风险 $r_i(\boldsymbol{X}) = \sum_{j=1}^{M} L_{ij}(\boldsymbol{X}) P(\omega_j \mid \boldsymbol{X})$

  损失函数 $L_{ij}(\boldsymbol{X}) \begin{cases} \leq 0 & i = j \\ > 0 & i \neq j \end{cases}$，指将自然属性是 ω_j 类的样本你决策为 ω_i 类时的是非代价
- $l_{12}(\boldsymbol{X}) = \dfrac{P(\boldsymbol{X} \mid \omega_1)}{P(\boldsymbol{X} \mid \omega_2)} > \dfrac{(L_{21} - L_{22}) P(\omega_2)}{(L_{12} - L_{11}) P(\omega_1)} = \theta_{12}$，则 $\boldsymbol{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

- $(0-1)$ 损失最小风险贝叶斯决策
  - $P(\boldsymbol{X} \mid \omega_k) P(\omega_k) > P(\boldsymbol{X} \mid \omega_i) P(\omega_i)$，则 $\boldsymbol{X} \in \omega_k$
- 正态分布模式的贝叶斯决策
  - 相关知识
    - 二次型 $\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}$
    - 正定二次型 $\forall \boldsymbol{X} \neq \boldsymbol{0}, \boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X} > 0$
    - 单变量正态分布 $p(x) = \frac{1}{\sqrt{2\mu\sigma}} \exp\left(-\frac{1}{2} \frac{(x-\mu)}{\sigma}^2 \right)$
    - 3σ 规则
    - 多变量正态随机变量

      $p(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{M})^T \boldsymbol{C}^{-1} (\boldsymbol{X} - \boldsymbol{M})$

      $\boldsymbol{M}$ 为各维度的均值，协方差矩阵 $\boldsymbol{C} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1n}^2 \\ \vdots & & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}$
  - $p(\boldsymbol{X} \mid \omega_i) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{C}_i|^{1/2}} \exp\left(\frac{1}{2} (\boldsymbol{X} - \boldsymbol{M}_i)^T \boldsymbol{C}_i^{-1} (\boldsymbol{X} - \boldsymbol{M}_i) \right)$
  - $d_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2} \ln |\boldsymbol{C}_i| - \frac{1}{2} \left\{ (\boldsymbol{X} - \boldsymbol{M}_i)^T \boldsymbol{C}_i^{-1} (\boldsymbol{X} - \boldsymbol{M}_i) \right\}$

贝叶斯分类器的错误概率

- 性能指标

  准确率 $A = \frac{TP+TN}{P+N}$

  精确率 $P = \frac{TP}{TP+FP}$

  虚警概率 $FA = \frac{FP}{TP+FP}$

  漏警概率 $MA = \frac{FN}{TN+FN}$

  召回率 $R = \frac{TP}{TP+FN}$
- 先验概率相等时的错误概率

  $P(e) = \int_{r_{ij}/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) \mathrm{d}y$

  r_{ij} 为马氏距离
- 错误率估计
  - 先验概率未知 $\hat{\varepsilon} = \frac{k}{N}$
  - 先验概率已知 $\hat{\varepsilon}' = \sum_{i=1}^{2} P(\omega_i) k_i / N_i$

聂曼-皮尔逊决策

- 相关知识

  $\Phi(\lambda)=\int_{-\infty}^{\lambda}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x$

  $\Phi(\lambda)=\int_{-\infty}^{\lambda}\varphi(x)\mathrm{d}x$

  当 $\lambda<0$ 时，$\Phi(\lambda)=1-\Phi(-\lambda)$

- $\frac{P(\boldsymbol{X}\mid\omega_1)}{P(\boldsymbol{X}\mid\omega_2)}\begin{matrix}>\\<\end{matrix}\mu$，则 $\boldsymbol{X}\in\begin{cases}\omega_1\\\omega_2\end{cases}$

  $P_2(e)=\int_{-\infty}^{t(\mu)}p(\boldsymbol{X}\mid\omega_2)\mathrm{d}\boldsymbol{X}$

特征选择与特征提取

K-L 变换

- K-L 展开式

  对随机变量集合 $\{\boldsymbol{X}\}$ 做离散正交展开，正交向量系为 $\{\boldsymbol{u}_j\}$，得到 $\boldsymbol{X}=\sum_{j=1}^{n}a_j\boldsymbol{u}_j$

  $\boldsymbol{u}_i^T\boldsymbol{u}_j=\begin{cases}1 & j=i\\0 & j\neq i\end{cases}$

  将 $\{\boldsymbol{u}_j\}$ 的特征值排序 $\lambda_1>\lambda_2>\cdots>\lambda_m>\cdots>\lambda_n\geq0$

  $\boldsymbol{X}=\sum_{j=1}^{m}a_j\boldsymbol{u}_j$

- 利用 K-L 变换特征提取
  - 求总体自相关矩阵 $\boldsymbol{R}=E\left[\boldsymbol{X}\boldsymbol{X}^T\right]\approx\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{X}_j\boldsymbol{X}_j^T$
  - 求出并取前 m 大的特征值 λ_i，算出特征向量 $\boldsymbol{u}_j'$，归一化得到 $\boldsymbol{u}_j$
  - $\boldsymbol{X}^*=\boldsymbol{U}^T\boldsymbol{X}$

类别可分性测度

- 基于距离的可分性测度
  - 类内距离

    $\overline{D^2}=E\left\{\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2\right\}=E\left\{(\boldsymbol{X}_i-\boldsymbol{X}_j)^T(\boldsymbol{X}_i-\boldsymbol{X}_j)\right\}$

    若样本相互独立，则 $\overline{D^2}=2\sum_{k=1}^{n}\sigma_k^2$

  - 类内散布矩阵：该类分布的协方差矩阵

    越小越好

  - 类间距离

    $\overline{D_b^2}=\sum_{i=1}^{c}P(\omega_i)\|\boldsymbol{M}_i-\boldsymbol{M}_0\|^2=\sum_{i=1}^{c}P(\omega_i)(\boldsymbol{M}_i-\boldsymbol{M}_0)^T(\boldsymbol{M}_i-\boldsymbol{M}_0)$

    其中，$P(\omega_i)$ 为先验概率，$\boldsymbol{M}_i$ 为 ω_i 类的均值向量，$\boldsymbol{M}_0=E\{\boldsymbol{X}\}=\sum_{i=1}^{c}P(\omega_i)\boldsymbol{M}_i$

- 类间散布矩阵

  $\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{M}_i - \boldsymbol{M}_0)(\boldsymbol{M}_i - \boldsymbol{M}_0)^T$

  $\overline{D_b^2} = \operatorname{tr}\{\boldsymbol{S}_b\}$

  越大越好

- 多类情况距离

  $J_d = \frac{1}{2}\sum_{i=1}^{c} P(\omega_i) \sum_{j=1}^{c} P(\omega_j) \frac{1}{n_i n_j} \sum_{k=1}^{n_i}\sum_{l=1}^{n_j} D^2\left(\boldsymbol{X}_k^i, \boldsymbol{X}_l^j\right)$

  其中，$\boldsymbol{X}_k^i$ 为 ω_i 类的第 k 个样本，n_i 为 ω_i 的样本数

  $J_d = \sum_{i=1}^{c} P(\omega_i)\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{X}_k^i - \boldsymbol{M}_i\right)^T\left(\boldsymbol{X}_k^i - \boldsymbol{M}_i\right) + (\boldsymbol{M}_i - \boldsymbol{M}_0)^T(\boldsymbol{M}_i - \boldsymbol{M}_0)\right]$

- 多类情况散布矩阵

  多类类间：$\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)\,(\boldsymbol{M}_i - \boldsymbol{M}_0)(\boldsymbol{M}_i - \boldsymbol{M}_0)^T$

  多类类内：$\boldsymbol{S}_w = \sum_{i=1}^{c} P(\omega_i) E\left\{(\boldsymbol{X} - \boldsymbol{M}_i)(\boldsymbol{X} - \boldsymbol{M}_i)^T\right\}$

  多类总体：$\boldsymbol{S}_t = E\left\{(\boldsymbol{X} - \boldsymbol{M}_0)(\boldsymbol{X} - \boldsymbol{M}_0)^T\right\} = \boldsymbol{S}_b + \boldsymbol{S}_w$

  $J_d = \operatorname{tr}(\boldsymbol{S}_t)$

- 基于概率分布的可分性测度

  - 散度

    ω_i 类对 ω_j 类的散度 $J_{ij} = \int_X [p(\boldsymbol{X} \mid \omega_i) - p(\boldsymbol{X} \mid \omega_j)] \ln \frac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)} \mathrm{d}\boldsymbol{X}$

    越大越好

    - $J_{ij} = J_{ji}$
    - $J_{ij} \geq 0$
    - 具有可加性
    - 值越大，错误率越小

  - 两个正态分布模式类的散度

    $J_{ij} = (\boldsymbol{M}_i - \boldsymbol{M}_j)^T \boldsymbol{C}^{-1}(\boldsymbol{M}_i - \boldsymbol{M}_j)$

    一维正态分布时 $J_{ij} = \frac{(m_i - m_j)^2}{\sigma^2}$

特征选择

神经网络