10.3 分类

① 分类的定义

利用已知类别的训练数据输入机器学习算法模型进行训练，用于将未知类别的测试数据划分成已知类别。标称离散值

② 原理

输入：设空间 $\mathcal{X} \subseteq R^N$ $(N \geq 1)$

输出：设输出空间 $\mathcal{Y} = \{c_1, c_2, \cdots c_K\}$ $K \in C$(正整) c_i表示类别

设 $D = \mathcal{X} \times \mathcal{Y}$（笛卡尔乘积）是一个未知的分布

目标函数：$f: \mathcal{X} \to \mathcal{Y}$，$f$为分类函数，将 $\mathcal{X}$ 映射到 $\mathcal{Y}$

训练集：$S = \{(x_i, y_k) \mid x_i \in \mathcal{X}, y_k = c_k \in \mathcal{Y}, i \in [1, m], k \in [1, K]\}$

其中 $(x_i, y_k) \in D$

设假设函数集为是 $\mathcal{X} \to \mathcal{Y}$ 的函数集合，目标：通过训练得到 $\hat{f} \in H$：泛

总：$\hat{f}: \mathcal{X} \to \mathcal{Y}$，使 $\hat{f}$ 和 f 有最小的泛化误差和经验误差

$$\hat{f} = \arg\min_{h \in H} P(h) = P_{\mathcal{X}}[h(x) \neq f(x)]$$

再用 $\hat{f}$ 对未知数据进行分类

给定具有 m' 个未知类别的 $T \subseteq \mathcal{X}$，$T = \{x_j \mid x_j \in \mathcal{X}, j \in [1, m']\}$

则 $\hat{f}(T) = \{(x_j, y_k) \mid j \in [1, m'], y_k = h(x_j) = c_k \in \mathcal{Y}, h \in [1, k]\}$

已知类别的标注数据 → 分类器 $H(X)$ —$\arg\min_{h \in H} P(h)$→ $\hat{f}(x)$ 分类函数

令 $\hat{f} = h$.

未知类别的数据 → $\hat{f}(x)$ 分类函数 —$\hat{y}_i = \hat{f}(x_i)$→ 分类结果

$\hat{y} = \hat{f}(x)$

③ 线性与非线性 ： 主要区别 线性为类器 与非线性为类器

1°. 线性为类器 特点 ： 线性判别函数 . 线性决策边界

(i) 线性判别函数 ： $y(x) = w \cdot x + b = \sum_{i=0}^{N} w_i \cdot x_i + b$

其中 $w \in R^N$ ： $w = (w_1, \cdots w_N)$

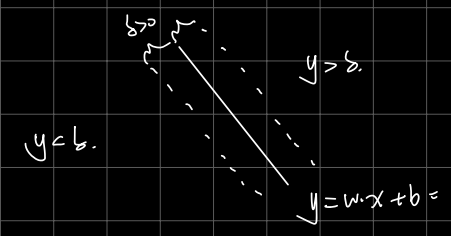
$x \in \chi$ ： $x = (x_1, \cdots x_N)^T$

$b \in R$ ： 偏差

(ii) 线性决策边界

对 $\forall b > 0$ ： （目标函数 $y(x) = w \cdot x + b$ 含 $y(x) = 0$）

$|y(x)| < b$ 时 $y(x) \to 0$



2° 非线性为类器 ： 特点 ： 基于非线性决策边界 （可为非曲线）

$y(x) = w_0 + w_1 x + w_2 x^2 + \cdots w_N x^N = \sum_{i=1}^{N} w_i x^i$

立则此后 ： $y(x) = \sum_{i=1}^{N} w_i \phi(x_i) = w^T \Phi(x)$

其中 $\Phi(x) = (x, x^2, \cdots x^N)$ 核函数

④ 类别与个数

类别个数 = 为类个数 ， $y_k(x) = w_k \cdot x + b$

判别准则 取决于超平面的维度 .

10.4. 回归

① 定义

回归用于估计因变量与自变量之间的关系，预测连续值

② 形式

空间：R^N ($N \geq 1$) 非负实数集

特征空间：$\mathcal{X} \subseteq R^n$　　其中 $D = \mathcal{X} \times \mathcal{Y}$ 为求和空间

标记空间：$\mathcal{Y}$

目标函数：$f: \mathcal{X} \to \mathcal{Y}$

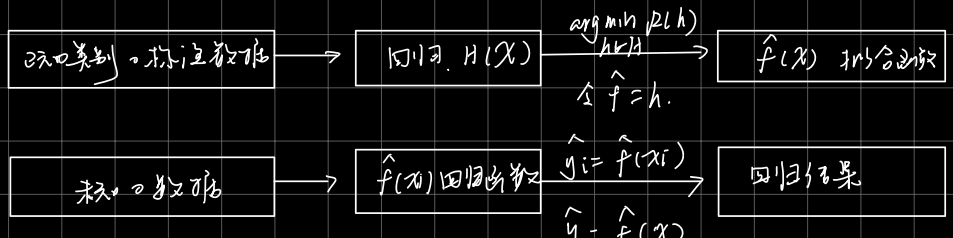
训练数据：$S = \{(x_i, y_i) \mid (x_i, y_i) \in D, i = [1, m]\}$

估计函数：$\hat{f} = \underset{h \in H}{\operatorname{argmin}}\, R(h) = \underset{h \in H}{\operatorname{argmin}}\, E_x[L(h(x), f(x))]$

其中 $h \in H$，h 为假设函数

未知数据：$T = \{x_j \mid x_j \in \mathcal{X}, j \in [1, m']\}$

预测数据：$\hat{f}(T) = h(T) = \{y_j = h(x_j) \mid j \in [1, m'], h(x_j) \in \mathcal{Y}\}$



③ 线性与非线性

1° 线性回归函数为线性组合　$y(x) = \sum_{i=0}^{N} w_i \cdot x_i + \varepsilon$　(ε 为残差) $\varepsilon \sim N(\mu, \sigma^2)$

$p(y|x, \theta) = N(y \mid \mu(x), \sigma^2(x))$　$\mu(x) = w^T x$　$\sigma^2(x) = \sigma^2$

$\theta = (w, \sigma^2)$　即 $p(y|x, \theta) = N(y \mid w^T x, \sigma^2)$

3° 非线性： $y(x)=\sum_{i=0}^{n} w_i x^i \Rightarrow \sum_{i=1}^{N} w_i \phi_i(x) = w^T \cdot \phi(x)$ （基为非线性的）

$$p(y|x,\theta)=N(y \mid w^T \phi(x), \sigma^2)$$

10.5.

（不同）	函数	输出	训练的目标函数	求解方法	目的
分类	决策函数	离散	$\arg\max_{h\in H} P(h)$	与决策平面距离最小	判断成功分类的
回归	回归函数	连续	$\arg\max_{\theta} p(y\|x,\theta)$	后验概率最大	估计（预测）未知值

（相同）：模式回归都是基于统计的方法，训练和预测流程类似

10.6. 聚类

① 定义：

将未标注数据通过训练寻找数据之间内在关系，将聚成不同一类簇

② 形式：

空间：R^N $(N\geq 1)$ 非负实数集

特征空间：$\mathcal{X} \subseteq R^n$ 其中 $n=\mathcal{X}$ 表示特征

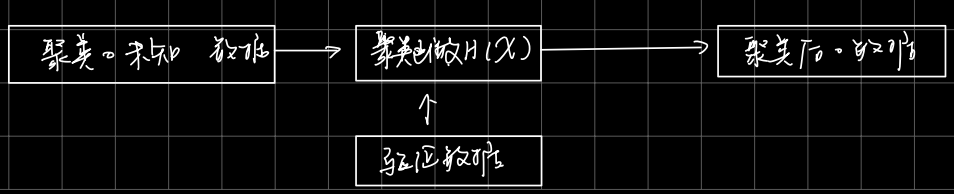
输出空间：$\mathcal{Y}$

最佳类聚函数：$f: \mathcal{X} \to \mathcal{Y}$ 且 $f\in H$（H为假设函数集）

输入数据（未标注）：$T=\{x_i \mid x_i \in \mathcal{X}, i\in[1,m]\}$

输出结果：$y'=\{c_1, c_2, \cdots c_k\}$ $k\leq K$ K为预定

$y'=f(T)=\{y_i=h(x_i) \mid x_i\in\mathcal{X}, i\in[1,m], y_k=c_k\in\mathcal{Y}, k\in[1,k]\}$



③ 聚类类型 { 硬性　硬性划分类簇，在该类簇计算特征值.
　　　　　　软性　按概率分布或似然值. 计算类簇属于

1° 基于连通性聚类（层次聚类）：根据对象之间距离

(i) 单链接聚类　$\min\{d(x,y): x\in A, y\in B\}$

(ii) 全链接聚类　$\max\{d(x,y): x\in A, y\in B\}$

(iii) 平均链接聚类　$\frac{1}{|A|\cdot|B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$

2° 基于质心聚类

先设定聚类个数，寻找聚类中心，将数据分配到聚类中心，使得该类簇中的数据与聚类中心的平方距离最小.

输入数据：$T=\{x_1, x_2, \dots x_m\}$

设 $k=K$. $C=\{c_1, \dots c_K\}$ K个类簇

设 μ_k 是第k个类簇中的均值，则目标函数为

$$f=\underset{C}{\arg\min}\sum_{k=1}^{K}\sum_{x\in c_k}\|x-\mu_k\|^2$$

算法流程：（迭代次数 t）

(i) $t=0$. 随机初始化 K 个类簇中心.

repeat

(ii) 将输入数据分配到每个类簇，使得平方误差距离最小.

$$\forall k, 1\le k\le K: \ c_k^{(t)}=\{x \mid \|x-\mu_k^{(t)}\|^2\le\|x-\mu_i^{(t)}\|^2\}$$

(iii) 更新：计算新的均值（质心）

$$\mu_k^{(t+1)}=\frac{1}{|c_k^{(t)}|}\sum_{x_i\in c_k^{(t)}} x_i$$

until 质心不再变化.

(iv) 算法终止.

3° 基于分布聚类：寻找属于同一概率分布最大的数据对象 (期望最大化)

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

$$r_{ik} \triangleq p(z_i=k|x_i,\theta) = \frac{p(z_i=k|\theta)p(x_i|z_i=k,\theta)}{\sum_{k=1}^{K} p(z_i=k'|\theta)p(x_i|z_i=k',\theta)}$$

$$z_i^* = \arg\max_k r_{ik} = \arg\max_k \log p(x_i|z_i=k,\theta) + \log p(z_i=k|\theta)$$

4° 基于密度聚类：根据已有数据划分

$$\text{局部密度} \quad \rho_i = \sum_j \chi(d_{ij} - d_c) \quad \chi(x) = \begin{cases} 1 & x<0 \\ 0 & x\geq 0 \end{cases}$$

其中 $d_{ij} = dist(x_i, x_j)$，d_c 为截断距离

δ_i 是寻找数据点与其他数据点的最小距离 $\delta_i = \min_{j:\rho_j>\rho_i} (d_{ij})$

10.7

(不同点)

	原理	数据	训练数据	求解方法	速度
分类	将数据分为已知类别	已标注类别	√	寻找延拓 (决策函数)	$\hat{f} = \arg\min_{f\in H} L(f)$
聚类	寻找内在联系划分类簇	未知类簇	×	硬性 / 软性	迭代直到类簇确定

相同点：分类和聚类都对数据进行划分的操作、训练后的结果

10.8 排序

① 定义：基于某种准则将收集进行排序.

② 原理：设测试空间：χ.

未知分布：$D = \chi \times \chi$

期望空间：$Y = \{\chi \times \chi, =\}$ 排序结果 $P(x_i \succ x_j) \in D: x_i = x_j$

目标函数：$f: \chi \times \chi \to Y$

训练样本：$S = \{(x_i, x_j, =) | (x_i, x_j) \in D, i \neq j \text{ and } i,j \in [1,m]\}$

$$\hat{f} = \underset{h\in H}{\operatorname{argmin}}\, R(h) = \underset{h\in H}{\operatorname{argmin}} \Pr(x_i, x_j)\sim D\left[f(x_i,x_j)\neq 0 \wedge f(x_i,x_j)(h(x_j)-h(x_i))\le 0\right]$$

③ 类型

1°. 基于次序 (评分函数)

2°. 基于依赖 (可学习)

3°. 基于相关性 : 样本皮尔逊 $r_{x,y} = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

$$\bar{x} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \bar{y} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})}$$

斯体 pearson. $\rho_{X,Y} = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$

$$r_s = \rho_{rgx,rgy} \frac{\operatorname{cov}(rgx, rgy)}{\sigma_{rgx}\sigma_{rgy}}$$

10.9 降维 ① 将高维数据映射到低维空间，并保留所有数据的基本特征

② 原理 : 输入空间 $x^k \subseteq R^N$ $(N>3)$

特征空间 $y^L \subseteq R^N$.

高维数据集 $S = \{x_1, \dots, x_m\} \subseteq x^{k\times m}$

映射集. 至: $x^k \to y^L$

最终降维后的数据集: $Y = \{y_i \mid \forall i \in [1,m], y_i \in \phi(x_i)\} \in y^{L\times m}$

③ 线性与非线性.

1° { (i) PCA . LCA . CCA . LDA 寻找投影矩阵 (线性方法)
(ii) 多维缩放空间

2° 非线性

(i) 多尺度

(ii) 核函数

(iii) 流形学习