

## 特征选择与降维学习

1. 为什么要特征选择（数据预处理的重要过程？

(1) 属性过多，造成维数灾难题

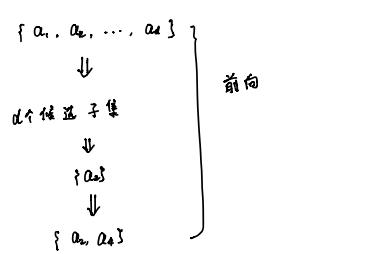
(2) 留下关键信息

目标：不丢失重要特征，去除无关特征，保留有关特征。

冗余：有时不起作用，有时有用

2. 通用的特征选择方法：迭代产生候选子集，评价好坏

(1) 子集搜索



也有后向不断删除，结合后有双向搜索

(2) 子集评价：求 IG

$$\text{Gain}(A) = \text{Ent}(D) - \sum_{i=1}^{|Y|} \frac{|D_i|}{|D|} \text{Ent}(D_i)$$

$$\text{Ent}(D) = -\sum_{k=1}^{|Y|} p_k \log p_k$$

过滤式：不考虑分类器，直接选特征

包裹式：结合分类器，可随机搜索

嵌入式：引入正则项，在优化过程，获得稀疏解，更快  
进行特征选择

### 1. 过滤式选择

设计相关统计量度量特征重要性。

(1) 定义阈值 (2) 挑选 k 个特征。

$$\text{二分类: } \delta^j = \sum_i -\text{diff}(x_i^j, x_{i, \text{mid}}^j)^2 + \text{diff}(x_i^j, x_{i, \text{low}}^j)^2$$

若  $\text{diff}(x_i^j, x_{i, \text{mid}}) < \text{diff}(x_i^j, x_{i, \text{low}})$ , 属性 j 有效

$$\text{多分类: } \delta^j = \sum_i -\text{diff}(x_i^j, x_{i, \text{mid}}^j)^2 + \sum_{l \neq j} c_l p_l \times \text{diff}(x_i^j, x_{i, l, \text{low}}^j)^2$$

$p_l$  为第 l 类在数据集 D 中的比例。

小结：时间开销小，准确率不如包裹式

### 2. 嵌入式与 L\_1 正则化

$$L_1 \text{ 正则: } \min_w \sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|_1$$

$$L_2 \text{ 正则: } \min_w \sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$$

L\_1 正则更容易得到稀疏解。

求解 L\_1 正则:

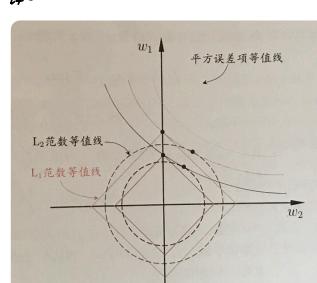


图 11.2 L<sub>1</sub> 正则化比 L<sub>2</sub> 正则化更易于得到稀疏解

$$\begin{aligned} \min_x f(x) + \lambda \|x\|_1 &\approx \left( f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\lambda}{2} \|x - x_k\|^2 \right)_{\min} \\ &= \left( \frac{\lambda}{2} \|x - (x_k - \frac{1}{\lambda} \nabla f(x_k))\|^2 + \text{const.} \right)_{\min} \\ \Rightarrow x_{k+1} &= x_k - \frac{1}{\lambda} \nabla f(x_k) \Rightarrow z = ? \end{aligned}$$

代入  $\min_z \frac{\lambda}{2} \|z - x_k\|^2 + \lambda \|z\|_1$

(A, x, z)

```

输入: 数据集 D;
特征集 A;
学习算法 L;
停止条件控制参数 T.
过程:
1: E = ∞;
2: d = |A|;
3: A* = A;
4: t = 0;
5: while t < T do
6:   随机产生特征子集 A';
7:   d' = |A'|;
8:   E' = CrossValidation(L(D, A'));
9:   if (E' < E) ∨ ((E' = E) ∧ (d' < d)) then
10:    t = 0;
11:    E = E';
12:    d = d';
13:    A* = A';
14:  else
15:    t = t + 1
16: end if
17: end while
输出: 特征子集 A*

```

图 11.1 LVW 算法描述

$$x_{k+1}^i = \begin{cases} z^i - \frac{\lambda}{2}, & z^i > \lambda/2 \\ 0, & |z^i| \leq \lambda/2 \\ z^i + \frac{\lambda}{2}, & z^i < -\lambda/2 \end{cases}$$

### 3. 稀疏表示与字典学习

稠密矩阵  $\Rightarrow$  适当稀疏

数据集  $\{x_1, x_2, \dots, x_m\}$

$$\min_{B, d_i} \|x_i - B a_i\|_2^2 + \lambda \sum_{i=1}^k \|d_i\|_1$$

最小  
稀疏

$k$ : 词汇量

$B$ : 字典

$d_i$ : 对应稀疏表示

(1) 固定  $B$ :  $\min_{d_i} \|x_i - B a_i\|_2^2 + \lambda \|d_i\|_1$ , 使用正则梯度下降求解

(2) 固定  $a_i$ :  $\min_B \|x_i - B a_i\|_F^2$ ,  $x \rightarrow d \times n$

$$= \min_{b_i} \|x_i - \sum_{j=1}^k b_j a_j^T\|_F^2$$

K-SVD: 增删更新

$$= \min_{b_i} \left\| \left( x_i - \sum_{j \neq i} b_j a_j^T \right) - b_i a_i^T \right\|_F^2$$

此时  $E_i$  为零向量

$$= \min_{b_i} \|E_i - b_i a_i^T\|_F^2$$

④ SVD, 求最大奇异值对应正交向量

为保持第 k 步得到的稀疏性, 只保留非零  $a_i^T$ , 其他也保留非零  $a_i^T$

### 4. 压缩感知

$$y = Ax \rightarrow A 完全无关冗余$$

↓ 测量矩阵

$$y = \Phi S = AS \rightarrow A, S 有稀疏性 可以还原$$

① RIP: 限字典稀疏性

$$(1 - \delta_k) \|s\|_2^2 \leq \|A_S s\|_2^2 \leq (1 + \delta_k) \|s\|_2^2$$

$$s.t. \quad \delta_k \in (0, 1)$$

$$\begin{cases} \min_s \|s\|_1 \\ s.t. \quad y = As \end{cases} \Rightarrow \begin{cases} \min_s \|s\|_1 \\ s.t. \quad y = As \end{cases} \text{ 使用正则梯度下降}$$

实例: 部分数据  $\rightarrow$  先整数据.

$$\min_x \text{rank}(x)$$

$$\text{s.t. } (X)_{ij} = (A)_{ij}, \quad (i, j) \in \Omega \quad \Omega \text{ 非元素}$$

$$\|x\|_F = \sqrt{\sum_{j=1}^{m \times n} b_j(x)}$$

$x$  的奇异值

$$\begin{cases} \min_x \|x\|_F \\ s.t. \quad (X)_{ij} = (A)_{ij}, \quad (i, j) \in \Omega \end{cases}$$

驻点规则求解

若  $\text{rank}(A) = r$ ,  $n \ll m$  仅需构造  $O(C \times r \log^2 m)$  个元素就能完美恢复  $A$ .