

特征选择与稀疏学习

1. 子集搜索与评价

data preprocessing - 特征选择 原因: 维度灾难. 去除相关特征会降低学习难度
如何评价候选特征子集的好坏? ① subset search. 每次增加特征(子集)为前向搜索; 每次去掉一个即为后向. 同时增减为双向搜索

② 子集评价 (subset evaluation) 例如信息增益

常见的特征选择方法: 过滤式 (filter), 包裹式 (wrapper), 嵌入式 (embedding)

2. 过滤式选择 filter

Relief: 相关统计量来度量特征的重要性. 例如指定一个阈值 τ . 然后选择比 τ 大的相关统计量分量所对应的特征; 也可指定欲选取特征个数 k . 再选择相关统计量分量数 $< k$ 个特征.
Relief 关键在子集统计量. 为二分类问题设计. Relief-F (多分类)

3. 包裹式选择

直接地根据特征使用机器学习性能作为特征子集的评价标准
LVW (Las Vegas Wrapper): 随机采样. 每次特征评价都需要训练学习器
若特征子集过多, 可能会并不出来

4. 嵌入式选择与 L1 正则化

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

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

L_1 比 L_2 更为获得 sparse 解. w 含有更多的非零分量

L_1 正则化可用近端梯度下降 (proximal gradient descent) 求解

$$\min_x f(x) + \lambda \|x\|_1 \quad \|\nabla f(x) - \nabla g(x)\|_2 \leq 2\lambda \|x\|_2$$

$$f(x) = \frac{1}{2} \|x - (x_k - \frac{1}{L} \nabla f(x_k))\|_2^2 + \text{const}$$

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k) \quad x_{k+1} = \arg \min_x \frac{1}{2} \|x - z\|_2^2 + \lambda \|x\|_1$$

$$\Rightarrow x_{k+1}^i = \begin{cases} z^i - \lambda/L & \lambda/L < z^i \\ 0 & |z^i| \leq \lambda/L \\ z^i + \lambda/L & z^i < -\lambda/L \end{cases}$$

5. 稀疏表示与字典学习

字典学习: 为普遍稠密的样本找到合适的字典, 将样本转化为合适的稀疏表示形式. 从而简化模型训练

采用字典B. LASSO求解. 再求 α : 稀疏表示B.

使用SVD对B进行奇异值分解. 得到稀疏性. 用户可设置记忆量 k

压缩感知 (compressed sensing): 根据部分信息来恢复全部信息

人脸识别. 十秒可过流