2 特征选择

统计学院 柳在唯

1. 特征选择问题

- 选取对训练数据具有分类能力的特征
- 准则：信息增益 或 信息增益比
- 总体的最优特征的选择是递归进行的

e.g.

ID	年龄	有工作	有个人住房	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

2. 信息增益

- 熵 Entropy
- 条件熵 Conditional Entropy

- 信息增益 Information Gain
- 信息增益比 Information Gain Ratio

熵

- 表示随机变量不确定性的度量
- 在离散情形下

$$H(X) = -\sum_{i=1}^{n} p_i log p_i$$

- 一些说明
- E.g. Bernoulli分布

条件熵

- 在已知随机变量X的条件下随机变量Y的不确定性
- 在离散情形下

$$H(Y|X) = \sum_{j=1}^{n} p_j H(Y|X = x_j)$$

- 经验熵与经验条件熵

信息增益

- 得知特征X的信息而使得类Y的信息的不确定性减少的程度

$$g(D,A) = H(D) - H(D|A)$$

- 互信息
- 信息增益比

$$g_R(D,A) = \frac{g(D,A)}{H(D)}$$

举例总结