机器学习课程实验九

2022 年 11 月 20 日　苏博南　202000460020

1　使用 CART 算法构建决策树

考虑一个数据集，包含 m 个样本，每个样本都有 n 个特征，即可表示为一个 $m \times n$ 的矩阵。那么对于某一个特征 i，我们把 m 个样本中该特征的取值排个序：

$$x_i^{(1)} \leq x_i^{(2)} \leq ... \leq x_i^{(m)} \tag{1}$$

然后就可以构造 $m-1$ 个形式如下的 predicate：

$$P(i,j) = x_i \leq \frac{x_i^{(j)} + x_i^{(j+1)}}{2}, j = 1, 2, ..., m-1 \tag{2}$$

然后给定 i, j，每个 predicate 都可以把数据集划分为如下两部分：

$$\begin{aligned} D_1 &= \{x^{(t)} \mid x_i^{(t)} \leq \frac{x_i^{(j)} + x_i^{(j+1)}}{2}, t = 1, ..., m\} \\ D_2 &= \{x^{(t)} \mid x_i^{(t)} > \frac{x_i^{(j)} + x_i^{(j+1)}}{2}, t = 1, ..., m\} \end{aligned} \tag{3}$$

故对整个数据集，我们可以构造 $n \times (m-1)$ 个 predicate，也就有这么多种方法可以把数据集一分为二。那么我们要做的就是找到一个 predicate，使得划分后的**基尼系数**加权和最小。

我们可以定义对一个数据集，它的**基尼系数**为：

$$Gini(D) = 1 - \sum_{i=1}^{K} (\frac{|D_i|}{|D|})^2 \tag{4}$$

其中 K 为数据集的类别数（二分类问题就是 2），然后 $|D_i|$ 为类别为 i 的样本数，$|D|$ 为总样本数。那么对一个 predicate 和其对应的划分，也就可以得到该划分的基尼系数：

$$Gini(P) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{5}$$

那么对于 $n \times (m-1)$ 个 predicate，我们选择基尼系数最小的把数据集一分为二。如此重复，直至被划分后的子数据集内全是同一个类，那么划分结束。构造出了一个二叉的决策树。

2　算法结果

为了节省时间和便于展示，我只选择了 100 个样本点进行构建决策树，并在 10 个其他样本点上测试，准确率为 90%，可以画出决策树：

```
alcohol ?< 10.60005
|
+- pH ?< 3.45005
|  |
|  +- volatile acidity ?< 0.18005
|  |  |
|  |  +- residual sugar ?< 1.40005
|  |  |  |
|  |  |  +-  ?< 1.0
|  |  |  |
|  |  |  `-  ?< 0.0
|  |  |
|  |  `-  ?< 0.0
|  |
|  `- citric acid ?< 0.34005
|     |
|     +-  ?< 0.0
|     |
|     `-  ?< 1.0
|
`- pH ?< 3.19005
   |
   +- chlorides ?< 3.805e-2
   |  |
   |  +- citric acid ?< 0.38005
   |  |  |
   |  |  +-  ?< 1.0
   |  |  |
   |  |  `-  ?< 0.0
   |  |
   |  `-  ?< 0.0
   |
   `- residual sugar ?< 1.00005
      |
      +-  ?< 0.0
      |
      `- total sulfur dioxide ?< 148.00005
         |
         +-  ?< 1.0
         |
         `-  ?< 0.0

[0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,1]
0.9
```