Homework5

司可 经济系 15320171151903

决策树的原理

决策树是机器学期算法的一个热点。决策树，顾名思义，就像一棵树一样。它是用一棵树的结构一样，将数据进行分类。直观理解，就是树的一个节点代表一个数据，然后利用它的不同特征值的取值，建立不同的分支，在每个分支中重复建立节点和分支的过程，这样就可以形成一棵树的形状。

详细来讲，决策树方法就是将预测空间划分为不同的区域，并在每个区域内拟合模型。预测空间的划分方法是对预测变量 $x_1, \ldots, x_p$ 进行连续的二叉分解，每次分割的时候，都将当前的空间一分为二，这样使得每一个叶子节点都是在空间中的一个不相交的区域。在进行决策的时候，会根据输入样本的特征值，选取一个变量 $x_j, j = 1, \ldots, p$，根据 $x_j \leq c$ 和 $x_j > c$ 对预测空间进行划分，然后对每一部分进行处理，最后使得样本落入所有区域中的一个区域。这样构造的决策树由内部节点和终端节点(叶子)组成。每个内部节点都是一个预测空间, 而叶子则是最终的预测。树的大小是它的叶子的数量。

决策树可以应用于回归和分类问题。根据处理数据类型的不同，决策树又分为两类：分类决策树与回归决策树，前者可用于处理离散型数据，后者可用于处理连续型数据。分类树分析是指预测结果是数据所属的类别。回归树分析是指预测的结果可以被认为是一个真实的数字(例如房价，或病人住院时间)。

分类与回归树的英文是 Classification and regression tree，缩写是 CART。CART 算法是一种递归划分算法，通过连续的二叉分割来划分预测器空间。在每一步，它选择分裂来实现 E_{in} 的最大下降。它由树的生成、树的剪枝构成。

通过以上我们明白了，构建决策树，就是根据数据的不同取值，建立树的分支，以及在每个分支中重复建立下层节点和分支。这一过程的关键就在于如何选择划分点以及如何决定叶节点的输出值。我们的目标是使 E_{in} 最小化。在回归设置中，使用 L2 误差，这意味着最小化：

$$\mathrm{RSS} = \sum_{m=1}^{M} \sum_{x_i \in R_m} (y_i - \bar{y}_m)^2$$

$$\bar{y}_m = \frac{1}{n_m} \sum_{x_i \in R_m} y_i$$

其中 $R_1, \ldots, R_M$ 表示划分的 M 个区域。n_m 是 R_m 中的观测值。

给定一个区域 $R \subseteq R^p$, 最优分割就是选择 x_j 和分割点 s, 这两个次区域

$R_{left} = \{x \in R: xj < s\}$， $R_{right} = \{x \in R: xj \geq s\}$对 y 的响应是均匀的。对于回归树，这意味着选择(j, s)来最小化：

$$\sum_{x_i \in R_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{x_i \in R_{\text{right}}} \left(y_i - \bar{y}_{\text{right}}\right)^2$$

该过程生成一个大型树 T_0。然后我们通过选择 α 和子树 T 来修剪树枝。对于,每个 $\alpha \geq 0$ $T \subset T0$，最小化：

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \bar{y}_m)^2 + \alpha |T|$$

其中|T|是 T 的大小，α 是一个调优参数,控制子树之间的权衡的复杂性及其所适合的训练集。最优的 α 可以使用交叉验证选择。

总结起来，建立一个回归树的步骤是：1、使用递归二进制分裂在训练集上生成一棵树，只有当每个终端节点的观测值小于某个最小值时才停止。2.修剪树枝，通过最优的 α 来获得最好的子树。3.使用 K-fold 交叉验证选择 α。也就是说，将训练集划分为 K 组。对于每个 k = 1，…K:

(a)对除第 k 次以外的所有训练集重复步骤 1 和步骤 2。

(b)评估除去第 k 组的均方误差，表示为 α 的函数。

平均每个 α 值的结果,选择能使平均误差最小化的 α。

4. 返回步骤 2 对应的子树选择 α 的价值。

分类树是一种对数据进行分类的树状结构。在使用分类树进行分类时，根据数据的某一特征值，将数据分配到其子结点。这时，每一个子结点对应着该特征的一个取值。如此递归地对数据进行分配，直到结束将所有数据分类完全。

与构建回归树一样，我们先找到使 E_{in} 最小化的 $R_1, \ldots, R_M$，令 $Q_m = E_{in}(R_m)$，然后在之后的每个步骤中，用 CART 算法选择能使 $n_{left}Q_{left} + n_{right}tQ_{right}$ 最小化的最优分割。其中 Q_m 也称为节点杂质测量，常用的节点杂质测量方法有误分率、基尼系数和交叉熵。

使用决策树有许多优点，比如它是高度可译的，能自动检测非线性关系，自动建模。然而也有一些缺陷，比如预测性能相对较差，通常方差较大，且不稳定，观测数据的微小变化常常导致完全不同的树。这是由于它们的层次性，一旦进行了分割，它下面的所有分割也会更改。还有难以捕捉简单的关系，它需要大量的参数(分割)来捕获简单的线性关系和加法关系。换句话说，决策树方法的缺点就是线性模型的优点，反之亦然。