机器学习基础作业 5

2025 年 4 月 22 日

问题 1. 给出回归问题的提升树算法框架.

Algorithm 1 提升树

Input: 训练数据集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, 弱学习算法回归树 $\mathcal{T}$, 损失函数 $L(y, f(x))$
Output: 强学习器 $f(x)$
初始化 $f_0(x) = 0$
初始化残差 $r_i = y_i - f_0(x_i)$, $i = 1, 2, \ldots, n$
for $m = 1$ **to** M **do**
    训练回归树 $\mathcal{T}_m$ 来拟合残差 r_i
    计算树的叶子节点 j 的值 $w_j = \frac{1}{|R_j|} \sum_{i \in R_j} r_i$, 其中 R_j 是第 j 个叶子节点的样本集合
    更新模型: $f_m(x) = f_{m-1}(x) + w_j$, 如果 x 在第 j 个叶子节点中.
    更新残差: $r_i = r_i - w_j$, 如果 x_i 在第 j 个叶子节点中
end for
Return: $f(x) = f_M(x)$

问题 2. 给出随机森林的算法框架, 并对随机森林和 *Bagging* 方法从行比较.

Algorithm 2 随机森林

Input: 训练数据集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, 弱学习算法回归树 $\mathcal{T}$, 树的数量 M
Output: 强学习器 $f(x)$
初始化 $f(x) = 0$
for $m = 1$ **to** M **do**
    从 D 中有放回地抽取 n 个样本, 构成训练集 D_m
    从所有 d 个特征中抽取 k 个, 随后训练回归树 $\mathcal{T}_m$ 在 D_m 上
    将 $\mathcal{T}_m$ 加入到随机森林中
end for
Return: $f(x) = \arg\max_y \sum_{m=1}^{M} I(f_m(x) = y)$, 其中 I 是指示函数

与原本的 Bagging 方法相比, 随机森林在 Bagging 的基础上增加了特征选择的随机性. 随机选取的目的是为了增加基学习器之间的差异性, 使得集成学习器的泛化性能更好.

问题 3. 基于 *DBSCAN* 的概念定义, 若 x 为核心对象, 由 x 密度可达的所有样本构成的集合为 X. 试证明: X 满足连接性与最大性.

证明. 连接性显然, 因为 X 中任意两个点 x_i, x_j 都由 x 密度可达, 因而它们密度连接.

对于最大性, 如果 $x_i \in X$, 且 x_j 由 x_i 密度可达, 由于 x 由 x_i 密度可达, 则 x_j 也由 x 密度可达. 从而由 X 的定义知 $x_j \in X$. □