Task04 并行集成的树模型

参考

- 课程链接: DataWhale/machine-learning-toy-code/Part B
- 视频链接: https://pan.baidu.com/s/1IydoCbA22tOkfjz_HjOk9Q 提取码: pgx9

练习

- 【练习】r2_score和均方误差的区别是什么？它具有什么优势?

  解答:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{MSE}{Var}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

  r2_score得到的是一个比值, 消除了不同样本集的量纲, 同时给定取值上界. 对于同一个模型可以直观比较在不同数据上的效果, 也可以直观比较不同模型在统一数据上的效果;

  MSE并没有消除量纲

- 【练习】假设使用闵氏距离来度量两个嵌入向量之间的距离，此时对叶子节点的编号顺序会对距离的度量结果有影响吗?

  解答: 不会, 闵式距离只是对应位置上的值相减, 位置的顺序变化不会带来结果的变化。

知识回顾

1. 什么是随机森林的oob得分?

   解答: 随机森林每一个基学习器使用了重复抽样得到的数据集进行训练, 对于没抽样到的样本称为oob(out of bag)样本, 每个基学习器训练完毕后使用oob样本进行预测, 并评分得到oob得分

2. 随机森林是如何集成多个决策树模型的?

   解答: 回归问题使用各学习器的均值, 分类问题使用投票策略或概率聚合策略

3. 请叙述孤立森林的算法原理和流程。

   解答:

   算法原理: 孤立森林是异常检测算法, 异常点更容易出现在相对稀疏的区域, 更容易通过更少的次数与其他点隔开

   流程: 见教程

代码实现

- **随机森林** 见这里
- **孤立森林** 见这里