基本机器学习模型复习

CSDN学院

机器学习模型的一般框架

- 目标函数
- 优化方法
- 模型评估/超参数调优

目标函数

- 目标函数通常包含两部分
  - 损失函数
  - 正则项

损失函数

- 由概率导出的损失函数
  - 最大似然 等价于小损失（负log似然损失/logloss/neg_logloss）
  - 回归：L2损失、L1损失
  - 分类：logistic损失（交叉熵损失）、softmax损失

- 非概率导出的损失函数
  - 回归：Huber损失/ϵ不敏感损失
  - 分类：合页损失（Hingloss）/指数损失
  - 树模型：Gini指标

正则项

- L0正则：稀疏解
- L1正则：稀疏解
- L2正则：收缩参数
- 组合正则

优化方法

- 解析解
- 梯度下降
- 随机梯度下降
- 坐标下降/SMO
- 牛顿法/拟牛顿法

- 树模型：优化与建树过程融合

模型

- 线性模型
  - 线性回归、Logistic回归、线性SVM
- 核方法
  - 原问题→对偶问题→核技巧（样本之间的点积用核函数代替）
  - 核函数的含义
  - 核函数的参数会影响模型性能（模型超参数）
- 树模型：非线性模型
- 基于树的集成模型
  - 随机森林
  - GBDT

模型评估/超参数调优

- 优化过程：根据训练集求得最佳模型参数
  - 给定模型超参数的情况下
- 模型的超参数（模型复杂度的参数）调优
  - 模型复杂度会严重影响模型性能
  - 通常通过验证集/交叉验证 寻找最佳模型超参数

线性模型超参数

- 线性模型超参数的超参数
  - 正则惩罚项（L1、L2、…）
  - 正则系数（λ）

核化线性模型超参数

- 核化线性模型超参数的超参数
  - 正则惩罚项
  - 正则参数
  - 核函数参数（通常用RBF核，核函数宽度为核函数参数）

树模型超参数

- 树的深度/叶子结点的数目（L0正则）
- 叶子结点最小样本数目
- 叶子结点样本最小权重和
- 待分裂节点的最小样本数目
- 待分裂节点的最小不纯度
- 分裂带来的最小不纯度降低量

随机森林超参数

- 树模型的超参数
  - 树的深度/叶子结点的数目（L0正则）
  - 叶子结点最小样本数目
  - 叶子结点样本最小权重和
  - 待分裂节点的最小样本数目
  - 待分裂节点的最小不纯度
  - 分裂带来的最小不纯度降低量
- 树的数目
- 行采样比例
- 列采样比例

GBDT超参数

- 树模型的超参数
  - 树的深度/叶子结点的数目（L0正则）
  - 叶子结点最小样本数目
  - 叶子结点样本最小权重和
  - 待分裂节点的最小样本数目
  - 待分裂节点的最小不纯度
  - 分裂带来的最小不纯度降低量
- 树的数目 vs. 学习率
- 行采样比例
- 列采样比例
- 正则参数

模型评估

- 通常采用交叉验证
  - 设置超参数搜索范围
  - 初始化一个GridSearchCV实例
  - 调用GridSearchCV的fit函数

CSDN
不 止 于 代 码

THANK YOU

AI100