DataWhale零基础入门金融风控

贷款违约预测--模调参&模型融合

本次直播的内容

- 单模型建模调参
- 多模型融合上分
- 问题答疑

分享人：小一

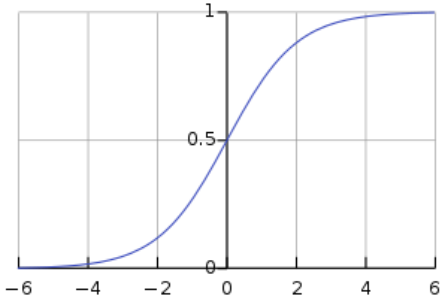
数据分析工程师

金融风控爱好者

单模型建模（1/3）

DataWhale

逻辑回归模型



假设数据服从伯努利分布，
通过极大化似然函数的方法，
运用梯度下降来求解参数，来达到将数据二分类的目的。

➢ 优点

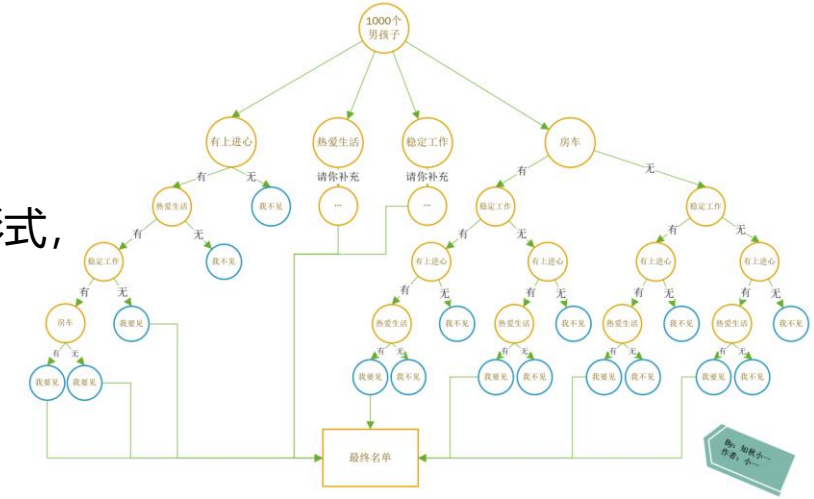
- 简单易理解，可解释性强
- 训练速度较快，计算量和特征的数目相关
- 内存资源占用小，只存储维度的特征值

➢ 缺点

- **需要预处理缺失值和异常值**
- **不能解决非线性问题**
- **准确率并不是很高**

树模型

通常是二叉树的形式，
少有多叉树存在



➢ 优点

- 简单直观，生成的决策树可以可视化展示
- **数据不需要预处理，不需要归一化、处理缺失数据**
- 既可以处理离散值，也可以处理连续值

➢ 缺点

- 决策树算法非常容易过拟合，导致泛化能力不强（可进行适当的剪枝）
- 采用的是贪心算法，容易得到局部最优解

单模型建模（2/3）

DataWhale

集成模型

集成方法主要包括 Bagging 和 Boosting，都是将已有的分类或回归算法通过一定方式组合起来，形成一个更加强大的分类

- 基于 Bagging 的集成模型：
  - 随机森林
- 基于 Boosting 的集成模型：
  - Adaboost
  - GBDT
  - XGBoost
  - LightGBM

	bagging	boosting
➢ **样本选择上**	从原始集中有放回的选取	每一轮的训练集不变，只是训练集中每个样本在分类器中的权重发生变化
➢ **样例权重上**	使用均匀取样，所以每个样本的权重相等	根据错误率不断调整样本的权值，错误率越大则权重越大
➢ **预测函数上**	所有预测函数的权重相等	每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重
➢ **并行计算上**	各个预测函数可以并行生成	各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。

单模型建模（3/3）

DataWhale

数据集划分

- **留出法**

  将数据集D划分为两个互斥的集合，其中一个作训练集，另一个为测试集

- **交叉验证法**

  将数据集D分为k份，其中k-1份作为训练集，剩余的一份作为测试集

- **自助法**

  从数据集D中取一个样本作为训练集中的元素，然后把该样本放回，重复该行为m次，未取到的样本为测试集

调参优化

- **贪心调参**

  依次调整，每一个参数都是局部最优

- **网格搜索**

  常使用GridSearchCV进行全局参数调整

- **贝叶斯调参**
  - 定义优化函数(rf_cv)
  - 建立模型
  - 定义待优化的参数
  - 得到优化结果

多模型融合

DataWhale

平均&投票

- **平均**
  - 简单平均
  - 加权平均
- **投票**

模型1 A-99%；B-1%
模型2 A-49%；B-51%
模型3 A-40%；B-60%
模型4 A-90%；B-10%
模型5 A-30%；B-70%

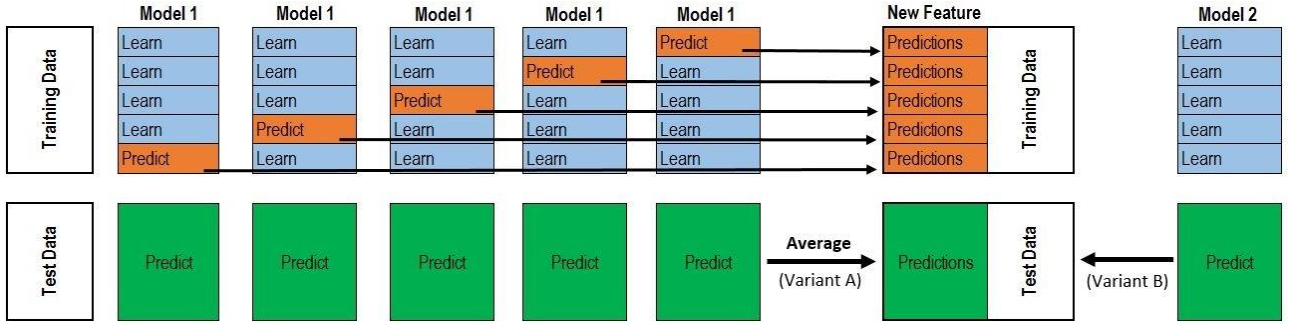
→ A-两票；B-三票；最终结果为B；Hard Voting

模型1 A-99%；B-1%
模型2 A-49%；B-51%
模型3 A-40%；B-60%
模型4 A-90%；B-10%
模型5 A-30%；B-70%

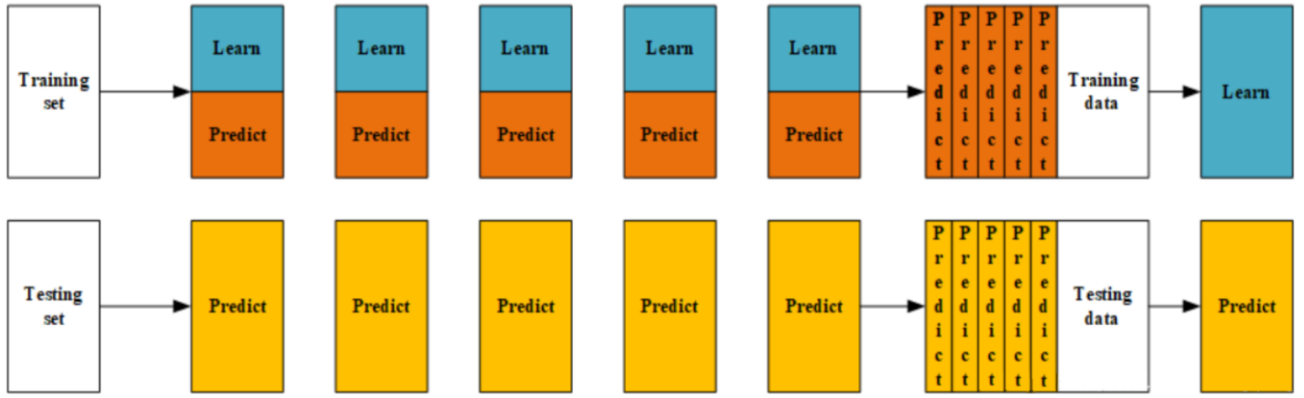
→ A - (0.99 + 0.49 + 0.4 + 0.9 + 0.3) / 5 = 0.616
→ B - (0.01 + 0.51 + 0.6 + 0.1 + 0.7) / 5 = 0.384
最终结果为A

stacking & blending

- **Stacking融合**



- **Blending融合**



贷款违约预测–建模调参&模型融合

Datawhale

一个专注于AI领域的开源组织

了解更多竞赛干货分享