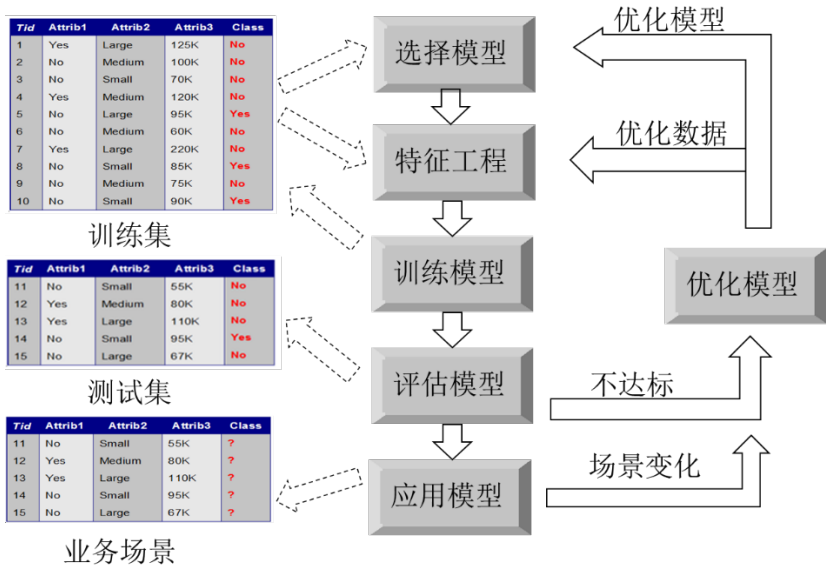
Python 数据建模（分类模型篇）

【课程目标】

本课程主要讲解如何利用 Python 进行分类数据建模。



通过本课程的学习，达到如下目的：

1、掌握数据建模的标准流程。

2、掌握各种分类预测模型的原理，以及算法实现。

3、掌握各种分类模型类的重要参数，以及应用。

4、掌握模型的评估指标、评估方法，以及过拟合评估。

5、掌握模型优化的基本方法，学会超参优化。

6、掌握集成优化思想，掌握高级的分类模型。

【授课时间】

2-5 天时间

（要根据学员的实际情况调整重点内容及时间）

【授课对象】

业务支持部、IT 系统部、大数据系统开发部、大数据分析中心、网络运维部等相关技术人员。

【学员要求】

1、 每个学员自备一台便携机(必须)。

2、 便携机中事先安装好 Python 3.9 版本及以上。

3、 安装好 Numpy,Pandas,statsmodels,sklearn,scipy 等常用库。

注：讲师现场提供分析的数据源。

【授课方式】

建模流程+ 案例演练 + 开发实践 + 可视化呈现

采用互动式教学，围绕业务问题，展开数据分析过程，全过程演练操作，让学员在分析、分享、讲授、总结、自我实践过程中获得能力提升。

【课程大纲】

第一部分： 预测建模基础

1、数据建模六步法

- 选择模型：基于业务选择恰当的数据模型
- 属性筛选：选择对目标变量有显著影响的属性来建模
- 训练模型：采用合适的算法，寻找到最合适的模型参数
- 评估模型：进行评估模型的质量，判断模型是否可用
- 优化模型：如果评估结果不理想，则需要对模型进行优化
- 应用模型：如果评估结果满足要求，则可应用模型于业务场景

2、数据挖掘常用的模型

- 数值预测模型：回归预测、时序预测等
- 分类预测模型：逻辑回归、决策树、神经网络、支持向量机等
- 市场细分：聚类、RFM、PCA 等
- 产品推荐：关联分析、协同过滤等

- 产品优化：回归、随机效用等
- 产品定价：定价策略/最优定价等

3、 属性筛选/特征选择/变量降维

- 基于变量本身特征
- 基于相关性判断
- 因子合并（PCA 等）
- IV 值筛选（评分卡使用）
- 基于信息增益判断（决策树使用）

4、训练模型及实现算法

- 模型原理
- 算法实现

5、模型评估

- 评估指标
- 评估方法
- 过拟合评估

6、模型优化

- 优化模型：选择新模型/修改模型

- 优化数据：新增显著自变量
- 优化公式：采用新的计算公式

7、模型应用

- 模型解读
- 模型部署
- 模型应用

8、好模型是优化出来的

第二部分： 分类模型评估

1、三个方面评估：指标、方法、过拟合

2、两大矩阵

- 混淆矩阵
- 代价矩阵

3、六大指标

- 正确率 Accuracy
- 查准率 Precision
- 查全率 Recall

- 特异度 Specify
- F 度量值（ F_1 / F_β ）
- 提升指标 lift

4、三条曲线

- ROC 曲线和 AUC
- PR 曲线和 BEP
- KS 曲线和 KS 值

5、多分类模型评估指标

- 宏指标：macro_P, macro_R
- 宏指标：micro_P, micro_R

6、模型评估方法

- 原始评估法
- 留出法（Hold-Out）
- 交叉验证法（k-fold cross validation）
- 自助采样法（Bootstrapping）

7、其它评估

- 过拟合评估：学习曲线

- 残差评估：白噪声评估

第三部分： 逻辑回归

问题：如何评估客户购买产品的可能性？如何预测客户行为？

如何预测客户流失？银行如何实现欠贷风险控制？

1、逻辑回归模型简介

2、逻辑回归的种类

- 二项逻辑回归
- 多项逻辑回归

3、逻辑回归方程解读

4、带分类自变量的逻辑回归

5、逻辑回归的算法实现及优化

- 迭代样本的随机选择
- 变化的学习率

6、逻辑回归+正则项

7、求解算法与惩罚项的互斥关系

8、多元逻辑回归处理

- ovo
- ovr

9、逻辑回归建模过程

案例：用 sklearn 库实现银行贷款违约预测

案例：订阅者用户的典型特征（二元逻辑回归）

案例：通信套餐的用户画像（多元逻辑回归）

第四部分： 决策树

1、分类决策树简介

演练：识别银行欠货风险，提取欠贷者的特征

2、决策树的三个关键问题

- 最优属性选择
  - 熵、基尼系数
  - 信息增益、信息增益率
- 属性最佳划分
  - 多元划分与二元划分
  - 连续变量最优划分
- 决策树修剪

- 剪枝原则
- 预剪枝与后剪枝

3、构建决策树的算法

- C5.0、CHAID、CART、QUEST
- 各种算法的比较

4、决策树的超参优化

5、决策树的解读

6、决策树建模过程

案例：商场酸奶购买用户特征提取

案例：客户流失预警与客户挽留

案例：识别拖欠银行货款者的特征，避免不良货款

案例：识别电信诈骗者嘴脸，让通信更安全

案例：电力窃漏用户自动识别

第五部分： 人工神经网络

1、神经网络简介（ANN）

2、神经元基本原理

- 加法器
- 激活函数

3、神经网络的结构

- 隐藏层数量
- 神经元个数

4、神经网络的建立步骤

5、神经网络的关键问题

6、BP 算法实现

7、MLP 多层神经网络

案例：评估银行用户拖欠货款的概率

案例：神经网络预测产品销量

第六部分： 支持向量机（SVM）

1、支持向量机简介

- 适用场景

2、支持向量机原理

- 支持向量

- 最大边界超平面

3、线性不可分处理

- 松弛系数

4、非线性 SVM 分类

5、常用核函数

- 线性核函数
- 多项式核
- 高斯 RBF 核
- 核函数的选择原则

第七部分： 模型集成优化篇

1、 模型的优化思想

2、 集成模型的框架

- Bagging
- Boosting
- Stacking

3、 集成算法的关键过程

- 弱分类器如何构建
- 组合策略：多个弱学习器如何形成强学习器

4、 Bagging 集成算法

- 数据/属性重抽样
- 决策依据：少数服从多数
- 随机森林 RandomForest

5、 Boosting 集成算法

- 基于误分数据建模
- 样本选择权重更新
- 决策依据：加权投票
- AdaBoost 模型

6、 GBDT 模型

7、 XGBoost 模型

8、 LightGBM 模型

第八部分： 案例实战

1、客户流失预测和客户挽留模型

2、银行欠贷风险预测模型

结束：课程总结与问题答疑。