Python 机器学习算法实战

【课程目标】

本课程为高级课程，专注于机器学习算法，原理，以及算法实现及优化。

通过本课程的学习，达到如下目的：

1、熟悉常见的机器学习的算法。

2、掌握机器学习的算法原理，以及数据推导。

3、学会使用 Python 来实现机器学习算法，以及优化算法。

4、掌握 scikit-learn 扩展库来实现机器学习算法。

【授课时间】

3-5 天时间

【授课对象】

IT 系统部、大数据系统开发部、大数据建模等 IT 技术人员。

【学员要求】

本课程只讲算法实现，不涉及完整的数据建模和模型使用，所以要求学员之前

已经掌握数据建模基础，熟悉建模过程。

1、 每个学员自备一台便携机(必须)。

2、 要求有 Python 开发基础，事先安装 Python 3.9 版本以上。

3、 要求有基本的数据分析和数据挖掘的知识。

注：讲师现场提供开源的安装程序、扩展库，以及现场分析的数据源。

【授课方式】

机器学习任务 + 算法原理 + 数学推导 + Python 实现

从任务出发，了解算法原理，以及数学推导过程，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： 机器学习基础

1、机器学习简介

2、机器学习的种类

- 监督学习/无监督学习/半监督学习/强化学习
- 批量学习和在线学习

- 基于实例与基于模型

3、机器学习的主要战挑

- 数据量不足
- 数据质量差
- 无关特征
- 过拟合/拟合不足

4、机器学习任务

- 监督：分类、回归
- 无监督：聚类、降维、关联规则

5、机器学习基本过程

6、机器学习常用库

第二部分： 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

第三部分： 特征工程处理

1、 数据预处理的主要任务

- 数据集成：多个数据集的合并
- 数据清洗：异常值的处理
- 数据处理：数据筛选、数据精简、数据平衡
- 变量处理：变量变换、变量派生、变量精简
- 数据归约：实现降维，避免维灾难

2、 数据集成

- 数据追加（添加数据）
- 变量合并（添加变量）

3、 数据清洗（异常数据处理）

- 取值范围限定
- 重复值处理
- 无效值/错误值处理
- 缺失值处理
- 离群值/极端值处理

4、 特征工程

- 变量变换：原变量取值更新，比如标准化
- 变量派生：根据旧变量生成新的变量
- 类型转换：数值型与类别型相互转换
- 特征选择：选择合适的自变量来建模
- 变量合并：多个变量合并，减少变量个数

5、 变量变换

- 为什么要做变量变换？
- 函数转换：中心化、对数变换、平方根变换…
- 标准化转换：min-max、mean、max absolution、Z-score…
- 正则化转换：将数据缩放到单位范式(L1/L2 变换)
- 正态化转换：将变量转换成正态分布(Box-Cox、Yeo-Johnson)

6、 类型转换

- 数字化：将字符串转换成数字
- 离散化：将数值型转换成类别型
- 哑变量化：将类别型转换成数值型

7、 特征选择

- 特征选择模式：Filter/Wrapper/Embedded

- Filter 特征选择：选择重要变量，剔除不重要的变量
  - 从变量本身考虑：方差阈值法
  - 从输入变量与目标变量的相关性考虑

8、 变量合并

- 因子分析（FA）
  - 因子分析的原理
  - 因子个数如何选择
  - 如何解读因子含义
- 主成分分析（PCA）

案例：提取影响电信客户流失的主成分分析

第四部分： 回归算法实现

1、建模的本质，其实是一个最优化问题

2、回归模型的基础

3、基本概念：损失函数

4、线性回归常用算法

- 普通最小二乘法 OLS

- 梯度下降算法
- 牛顿法/拟牛顿法

5、最小二乘法

a）数学推导

b）OLS 存在的问题

6、过拟合解决方法：正则化

- 岭回归（Ridge）
- 套索回归 Lasso
- ElasticNet 回归
- 各种算法的适用场景

7、超大规模数据集的回归模型：迭代算法

- 梯度概念
- 梯度下降/上升算法
- 批量梯度 BGD/随机梯度 SGD/小批量梯度 MBGD
- 学习率的影响
- 早期停止法

8、梯度算法的关键问题

9、牛顿法/拟牛顿法

- 泰勒公式(Taylor)
- 牛顿法(Newton)
- 拟牛顿法(Quasi-Newton)的优化
  - DFP/BFGS/L-BFGS

10、 算法比较

第五部分： 逻辑回归算法

1、逻辑回归基础

2、LR 的常用算法

- 最大似然估计法
- 梯度算法
- 牛顿法

3、最大似然估计法

- 似然函数/损失函数
- 数学推导

4、模型优化

- 迭代样本的随机选择

- 变化的学习率

5、逻辑回归+正则项

6、求解算法与惩罚项的关系

7、多元逻辑回归处理

- ovo
- ovr
- 优缺点比较

8、逻辑回归建模实战

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）

第六部分： 决策树算法

1、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

- ID3、C4.5、C5.0
- CART

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

第七部分： 神经网络算法

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

8、学习率的设置

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

第八部分： 线性判别算法

1、判别分析简介

2、判别分析算法

- 中心和方差
- 类间散席 Sb
- 类内散席 Sw

3、特征值和特征向量

4、多分类 LDA 算法

5、算法实战

案例：MBA 学生录取判别分析

案例：上市公司类别评估

第九部分： 最近邻算法（KNN）

1、KNN 的基本原理

2、K 近邻的关键问题

- 距离公式
- 投票机制

3、KNN 算法实现

- Brute（蛮力计算）
- Kd_tree（KD 树）
- Ball_tre（球树）

4、算法比较

第十部分：贝叶斯算法（NBN）

1、贝叶斯简介

2、贝叶斯分类原理

- 先验概率和后验概率
- 条件概率和类概率

3、常见贝叶斯网络

4、计算类别属性的条件概率

5、估计连续属性的条件概率

6、预测分类概率（计算概率）

7、拉普拉斯修正

案例：评估银行用户拖欠货款的概率

第十一部分： 支持向量机算法（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

6、SMO 算法

第十二部分： 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

第十三部分： 聚类分析（客户细分）实战

1、聚类基本原理

2、K 均值聚类算法

- K 均值算法

3、距离计算公式

- 闵可夫斯基距离(Minkowski Distance)
- 曼哈顿距离(Manhattan Distance)
- 欧氏距离(Euclidean Distance)
- 切比雪夫距离(Chebyshev Distance)
- 余弦距离(Cosine)
- Pearson 相似距离
- 马哈拉诺比斯距离（Mahalanobis）
- 汉明距离(Hamming distance)

- 杰卡德相似系数(Jaccard similarity coefficient)
- 相对熵（K-L 距离）

4、K 均值算法的关键问题

- 初始中心的选取方式
- 最优 K 值的选取

5、聚类算法的评价方法

- Elbow method（手肘法）
- Calinski-Harabasz Index（CH 准则法）
- Silhouette Coefficient（轮廓系数法）
- Gap Statistic（间隔统计量法）
- Canopy 算法

6、算法实战

案例：使用 SKLearn 实现 K 均值聚类

第十四部分： 关联规则算法

1、关联规则基本原理

2、常用关联规则算法

- Apriori 算法
  - 发现频繁集
  - 生成关联规则
- FP-Growth 算法
  - 构建 FP 树
  - 提取规则

3、算法实战

案例：使用 apriori 库实现关联分析

案例：中医证型关联规则挖掘

第十五部分： 协同过滤算法

1、协同过滤基本原理

2、协同过滤的两各类型

- 基于用户的协同过滤 UserCF
- 基于物品的协同过滤 ItemCF

3、相似度评估常用公式

4、UserCF 算法实现

- 计算用户间的兴趣相似度

- 筛选前 K 个相似用户
- 合并相似用户购买过的物品集
- 剔除该用户已经购买过的产品，得到候选物品集
- 计算该用户对物品的喜欢程度，物品集排序
- 优先推荐前 N 个物品

5、ItemCF 算法实现

- 计算物品间的相似度
- 筛选前 K 个喜欢的物品
- 合并与前 K 个物品相似的前 L 个物品集
- 剔除该用户已经购买过的物品，得到候选物品集
- 计算该用户到候选物品的喜爱程度，物品排序
- 优先推荐前 N 个物品

6、关于冷启动问题

7、协同过滤算法比较

结束：课程总结与问题答疑。