第四章 数据分析和挖掘

- 1、聚类分析
  - 数据异常对聚类结果影响
    - K-mean
      - 基于点与点距离的相似度来计算最佳类归属
      - K-mean需注意的数据异常 可选择DBSCAN等其他算法代替
        - 数据的异常值
        - 数据的异常量纲
  - K-mean在超大数据量时的缺点
    - 算法消耗时间与样本量成线性增长
      - 可改用MiniBatchKMeans 分批处理
  - 用处
    - 图像压缩
    - 图像分割
    - 图像理解
    - 异常检测
    - 数据离散化
  - 高位聚类
    - 存在的问题
      - 高维度下，基于距离的相似度计算效率极低
      - 高维度下，大量属性特征使得所有维上存在簇的可能性非常低
      - 由于稀疏性和近邻特性，基于距离的相似度为0，导致高维空间中难存在数据簇
    - 解决办法
      - 降维
        - 特征选择法
        - 维度转换法
      - 子空间聚类
        - 选取与给定簇密切相关的维，然后在对应的子空间聚类
  - 聚类算法的选择因素
    - 是否高维
      - 高维可用谱聚类
    - 数据量的大小
    - 数据集有噪点/离群点
      - DBSCAN
    - 分类精确度的追求
  - 聚类效果评估
    - 聚类效果评估指标

        - inerias 样本距离最近的聚类中心的总和
            - 越小越好，表明分布越集中
        - ajusted_rand_s 调整后的兰德系数
            - 取值范围[-1,1]，负数代表结果不好，越接近1越是意味着聚类结果与真实情况吻合
        - mutual_info_s 互信息
            - 结果非负值
        - ajusted_mutual_info_s 调整后的互信息
            - 聚类集相同，AMI为1
            - 随机分区平均预期AMI约为0，也可能为负
        - homogeneity 同质化得分
            - 如果所有聚类都包含属于单个类的成员的数据点，则聚类结果将满足同质性
            - 取值范围[0,1]，越大则与实际情况越吻合
        - completeness_s 完整性得分
            - 如果作为给定类的成员的所有数据点是相同集群的元素，则聚类结果满足完整性
            - 取值范围[0,1]，越大则与实际情况越吻合
        - v_maesure_s 同质化和完整性之间的谐波平均值
            - v = 2 * (完整性 * 均匀性)/(完整性 + 均匀性)
            - 取值范围[0,1]，越大则与实际情况越吻合
        - silhouette_s 轮廓系数
            - 使用平均群内距离和每个样本的平均最近簇距离来计算
            - 非监督式评估指标之一
            - 负值表示样本已经被分配到错误集群中
        - calinski_harabaz_s 群内离散与簇间离散的比值
            - 非监督式评估指标之一
    - sklearn聚类效果评估方法
        - score方法
            - model.score(x_test, y_test)
        - 交叉验证模型
        - metrics库
    - 考虑因素
        - 对于没有任何聚类真实结果的指标，由于无法使用真实数据作对比，只能使用聚类距离指标做评估
        - 对于有分类真实结果可做对照的，则可用真实标签与预测标签的相似、重复、完整性等度量计算
        - 业务类评估，包括不同类别间的特征是否有显著差异，类内部是否具有能代表类别的显著性特征，不同类别内的样本量分别是否相对均匀等
- 2、回归分析
    - 概述
        - 优点：数据模型和结果便于理解

- 缺点：只能分析少量变量之间的关系，无法处理海量变量，尤其是变量共同因素对因变量的影响程度
- 常见算法
  - 线性回归
  - 二项式回归
  - 对数回归
  - 指数回归
  - 核SVM
  - 岭回归
  - Lasso

- 回归分析的系数
  - 概述
    - 回归系数
      - 回归方程中表示自变量x对因变量y影响大小的参数
      - 绝对值高低只能说明自变量与因变量之间的联系程度和变化量的比例
      - 量级越大的自变量，系数越高
    - 判定系数
      - 自变量对因变量的方差解释程度的值 R2R^{2} R2
      - 所有参与模型中自变量对因变量联合影响程度，而非某个自变量的影响程度
      - 没有高低统一标准，值越高越能代表自变量对因变量的解释作用越大
      - 公式：回归平方和与总离差平方和之比
    - 相关系数（解析系数）
      - 衡量变量间的相关程度或密切程度的值
      - 本质是相关性的判断
  - 相互关系
    - 只有一个自变量时，判定系数等于相关系数的平方
    - 回归系数>0，相关系数取值在(0,1)，说明二者正相关
    - 回归系数<0，相关系数取值在(-1,0)，说明二者负相关
- 自变量需要注意的问题
  - 是否产生了新的对因变量影响更大自变量
  - 原有自变量是否仍然控制在训练模型时的范围内
- 选择算法的考虑因素
  - 自变量的个数
    - 高维度可使用正则化回归方法，例如Lass，Ridge和ElasticNet；或者使用逐步回归挑选影响显著的自变量
  - 共线性的强弱
    - 较强可选岭回归
  - 数据集的噪音
    - 主成分分析
  - 模型可解释性
    - 线性回归，指数回归，对数回归，二项或多项式回归

- 集成或组合回归方法

- 相关知识点
  - cross_val_score
    - sklearn.model_selection中的交叉检验工具，可对特定的算法模型进行交叉检验
    - 常用参数
      - X, y 交叉检验的数据集X和目标y，其中X至少两维
      - cv 交叉检验的模式
        - 空值，默认三折交叉验证
        - 整数，按照制定的数量做交叉验证
      - scoring 评估算法模型的计算方法
        - 分类
        - 回归
        - 聚类
  - 指标评估
    - explain_variance_score 解析回归模型的方差得分
      - 取值范围【0.1】
      - 越接近1自变量越能解释因变量的方差变化
      - 越小说明效果越差
    - mean_absolute_error 平均绝对误差
      - 评估预测结果和真实结果的接近程度
      - 越小说明拟合效果越好
    - mean_squared_error 均方差
      - 拟合数据和原始数据对应样本点的误差的平方和的均值
      - 越小说明拟合效果越好
    - r2_score 判定系数
      - 取值范围【0.1】
      - 越小说明效果越差
      - 越接近1自变量越能解释因变量的方差变化
  - matplotlib 基本样式
    - tittle
    - legend
    - xlabel，ylabel
    - xtickets，ytickets
    - xlim，ylim 坐标轴范围，例如 plt.xlim((1,10))
    - text 增加文字到图形中
    - suptitle 图像居中主标题
    - axhline，axvline 水平/垂直增加一条线，例如 plt.axhline(y=0,xmin=0,xmax=1)
    - 小技巧
      - 关闭坐标轴刻度 plt.xticks([]); plt.yticks([])
      - 关闭坐标轴 plt.axis('off')

- 3、分类分析
  - 概述
    - 通过对已知类别训练集的计算和分析，从中发现类别规则并预测新数据的类别
    - 常用分类方法
      - 朴素贝叶斯
      - 逻辑回归
      - 决策树
      - 随机森林
      - 支持向量机
  - 分类过拟合
    - 由于过度学习训练集特征，使得训练集的准确度非常高，但将模型应用于新的数据集时准确率却很差
    - 解决办法
      - 使用更多的数据
        - 增加数据集和新数据集特征的相识度
      - 降维
      - 使用正则化方法
        - 通过定义不同特征的参数来保证每个特征有一定的效用，不会使某个特征特别重要
      - 使用组合方法
        - 例如随机森林，adaboost
  - 分类分析潜在作用
    - 提炼应用规则（变量间关系）
    - 提取变量特征（权重信息）
    - 处理缺失值
  - 聚类和分类的区别
    - 学习方式不同
      - 聚类 非监督学习算法
      - 分类 监督学习算法
    - 对原数据集要求不同
      - 分类需要标签作为监督学习的标准
    - 应用场景不同
      - 聚类 数据探索分析，数据降维，数据压缩等探索性分析和处理
      - 分类 预测性分析和使用
    - 解读结果不同
      - 聚类 将不同数据集按照各自的典型特征分成不同类别，不同人对其解读可能不同
      - 分类 一个固定值，不存在不同解读的情况
    - 模型评估指标不同
      - 聚类 没有“准确”与否，以及如何准确的相关度量，更多是基于距离的度量
      - 分类 有准确率、混淆矩阵、提升率等明显的好坏评估指标

- 分类算法的选择因素
  - 文本
    - 朴素贝叶斯，如垃圾邮件识别
  - 训练集的大小
    - 比较小，例如偶朴素贝叶斯，支持向量机（不容易过拟合）
    - 比较大，不管哪种方法，不会有显著的差异
  - 侧重的方向
    - 模型时间和易用性
      - 不推荐支持向量机，人工神经网络
    - 模型准确性
      - 选用支持向量机，随机森林
    - 预测结果的概率信息
      - 逻辑回归
  - 可能有离群点或数据不可分
    - 决策树
- 相关知识点
  - 切分数据集
    - train_test_split
      - 切分为训练集和测试集
    - split
      - 切分为训练集、测试机和验证集
  - model_tree.tree 属性
    - children_left 子级左侧分类节点
    - children_right 子级右侧分类节点
    - feature 子节点上用来做分裂的特征
    - threshold 子节点上对应特征的分裂阈值
    - values 子节点中包含正例和负例的样本数量
  - 混淆矩阵
    - 真正 TP 本身正例，分类成正例
    - 真负 TN 本身正例，分类成负例
    - 假真 FP 本身负例，分类成正例
    - 假负 FN 本身负例，分类成负例
  - sklearn.metrics 评价指标
    - auc_s AUC，ROC曲线下的面积
      - 取值范围一般【0.5,1】
      - AUC越大，分类效果越好
    - accuracy_s 准确率
      - 公式为 $\frac{(TP+TN)}{(TP+FN+FP+TN)}$ ，取值范围【0,1】
      - 数值越大，分类结果越准确
    - precision_s 精确度

        - 公式为 P = (TP+TN)(TP+FN+FP+TN) \frac{(TP+TN)}{(TP+FN+FP+TN)} (TP+FP)TP ，取值范围【0,1】
        - 数值越大，分类结果越准确
    - recall_s 召回率
        - 公式为 R = (TP+TN)(TP+FN+FP+TN) \frac{(TP+TN)}{(TP+FN+FP+TN)} (TP+FN)TP ，取值范围【0,1】
        - 数值越大，分类结果越准确
    - f1_s F1得分
        - 公式为 (TP+TN)(TP+FN+FP+TN) \frac{(TP+TN)}{(TP+FN+FP+TN)} (R+P)2*P*R ，取值范围【0,1】
        - 数值越大，分类结果越准确
- 4、关联分析
    - 概述
        - 定义
            - 通过寻找最能够解析数据变量之间关系的规则，来找出大量多元数据集中有用的关联规则
        - 常用算法
            - Apriori
            - FP-Growth
            - PrefixSpan
                - 基于前缀树的序列关联算法
            - SPADE
                - 基于垂直存储结构和格理论连接操作的序列关联算法
            - AprioriAll
                - 基于哈希树的序列关联算法
            - AprioriSome
                - 对AprioriAll的改进
        - 指标
            - Support 支持度
                - 项集{X,Y}在总项集里出现的概率
                - Support(X→Y)Support(X \rightarrow Y)Support(X→Y) = P(X,Y)P(I) \frac{P(X,Y)}{P(I)} P(I)P(X,Y) = P(X∪Y)P(I) \frac{P(X \cup Y)}{P(I)} P(I)P(X∪Y) = num(X∪Y)num(I) \frac{num(X \cup Y)}{num(I)} num(I)num(X∪Y) -num(I)表示总事务集的个数 -num(X∪Y)表示含有{X,Y}的事务集的个数（个数也叫次数）
            - Confidence 置信度
                - 在先决条件X发生的情况下，由关联规则”X→Y“推出Y的概率。即在含有X的项集中，含有Y的可能性
                - Confidence(X→Y) = P(Y|X) = P(X,Y) / P(X) = P(XUY) / P(X)
            - Lift 提升度
                - 含有X的条件下，同时含有Y的概率，与Y总体发生的概率之比
                - Lift(X→Y) = P(Y|X) / P(Y)

        - 结论
            - 支持度和置信度高，则该规则的频率高
            - 提升度低，则规则前后项互斥
    - 关联分析场景
        - 相同维度的关联分析
            - 网站页面浏览关联分析
            - 广告流量关联分析
            - 用户关键字搜索关联分析
        - 跨维度的关联分析
            - 不同场景下的关联分析
                - 发生的事件处于不同的时间下，但通常都在一个约束时间范围内
            - 相同场景下的事件关联
                - 发生的事件在一个场景下，但属于不同的时间点
    - 落地应用
        - 打包组合
        - 故意分离
            - 作用
                - 利用用户主动查找的时机来产生更多价值或完成特定转化目标
            - 条件
                - 关联规则必是强规则和有效
                - 发生关联的前后项之间有非常强的完成动机
                - 不能过多降低用户体验
        - 序列模式
            - 客户购买行为预测
            - web访问模式预测
            - 流量来源预测
            - 关键词搜索预测
- 5、异常检测分析
    - 概述
        - 定义
            - 数据集中的异常数据，即孤立点，离群点，噪音
        - 原因
            - 业务操作影响
            - 数据采集问题
            - 数据同步问题
        - 常用方法
            - 基于统计的异常检测方法
                - 泊松分布
                - 正态分布
            - 基于距离的异常检测方法

        - K-mean
      - 基于密度的异常检测方法
        - LOF
      - 基于偏移的异常检测方法
      - 基于时间序列的异常检测方法
    - 应用领域
      - 异常订单检测
      - 风控（风险客户预警）
      - 黄牛识别
      - 贷款风险识别
      - 欺诈检测
      - 技术入侵
    - 分类（根据原数据集不同）
      - 离群点
      - 新奇点
        - 识别新的或未知数据模式和规律的检测方法
        - 前提
          - 已知训练数据集是纯净的
  - 高维异常检测思路
    - 扩展现有的离群点检测模式
    - 发现子空间中的离群点
    - 对高维数据进行建模
  - 相关知识点
    - OneClassSVM
      - 基本原理
        - 在给定的一组样本中，检测数据的边界以便于区分新的数据点是否属于该类
      - 基于密度的异常检测方法，无监督学习
- 6、时间序列分析
  - 概述
    - 定义
      - 研究数据随时间变化趋势而变化
    - 常用算法
      - 移动平均
      - 指数平滑
      - 差分自回归移动平均
    - 常用场景
      - 经济预测
      - 股市预测
      - 天气预测
  - 与时间序列关系

- 自变量
  - 时间序列通常用于没有自变量可用的条件下进行预测性分析
  - 但趋势一定会存在随时间变化而变化的隐形规律
- 复杂的商业环境
  - 很多因素无法通过时间规律反映

- 时间序列预测的模式
  - 整合模式
    - 天
  - 横向模式
    - 小时
  - 纵向模式
    - 分钟

- 时间序列数据的平稳性
  - 定义
    - 数据有没有随着时间呈现出明显的趋势和规律，或者是相对均匀且随机地分布在均值附近
  - 常用判断方法
    - 观察法
      - 观察时间序列图
    - 自相关和偏相关法
      - 观察自相关和偏相关的系数
    - ADF检验
      - 通过ADF检验得到显著性水平
  - 常用解决方法
    - 对数法
      - 对数处理减少数据的波动，只针对时间序列大于0的数据集
    - 差分法
      - 一般经过一阶或二阶查分后得到平稳数据
    - 平滑法
      - 移动平均法
      - 指数平均法
    - 分解法
      - 长期趋势
      - 季节趋势
      - 随机成分

- 白噪声检验
  - 定义
    - 随机性检验
    - 用于检验时间序列的各项数值之间是否具有任何相关关系
  - 方法

        - 图像法
            - 正常时间序列是围绕均值随机性上下分布的
        - Ljung Box法
            - 判断时间序列是否存在滞后性
    - 白噪声检验和数据平稳性检验是协同进行的，所以平稳性检验通过了，白噪声一般也会通过
- 指标的分析平衡性
- 关键知识点
- 相关知识补充
    - Doc string
        - 函数注释，通过help(func_name)查看
    - lambda
        - 接受任意多个参数并且默认返回单个表达式
        - 一次性函数
        - 默认包含return

- 7、路径、漏斗、归因和热力图分析
    - 漏斗分析
        - 定义
            - 查看特定目标的完成和流失情况
        - 应用场景
            - 分析站内流程
            - 单页面多步骤分析
    - 路径分析
        - 定义
            - 基于页面产生
            - 基于目标路径、事件路径等数据主体产生
        - 应用场景
            - 活动主会场如何导流
            - 用户是否按照“预期”流程行动
            - 购买“手机用户”的浏览习惯是怎样的
            - 渠道A集中访问了某条路径，是否是“恶意流量”
        - 用户行为模式挖掘
            - 应用于站外广告渠道分析
            - 应用于户关键之搜索分析
    - 归因分析
        - 定义
            - 用于评估多个参与转化的主体如何分配贡献大小
        - 主要存在于线上
            - 原因
                - 线上转化行为的归属模式性
            - 难点

        - 用户点击了多个商品陈列位置并购买了其中的商品
        - 用户点击了多个商品陈列位置但购买了其他商品
    - 热力图分析
        - 定义
            - 分析单个页面的点击分布热力图
            - 反映用户对于页面内容的喜好程度
    - 漏斗分析和路径分析的区别
        - 分析目标不同
            - 漏斗 测量有特定目标的场景
            - 路径 侧重分析开放性的流程
        - 不同环节的关系不同
            - 漏斗 只能看到从上一级节点到下一级节点的转化关系以及到外部的流失节点
            - 路径 可展示某个节点跟其他所有节点之间的关系而不限定于是否转化或流失
        - 衡量主体之间的逻辑不同
            - 漏斗 单向的上下游关系
            - 路径 不存在明显的前后序列或转化关系
        - 表示结果不同
            - 漏斗 完成率、流失率等
            - 路径 使用时间的比例来评估，例如页面浏览量占比、访问量占比等
        - 应用主体不同
            - 漏斗 分析不同类型数据之间的转化关系
            - 路径 只用来分析相同维度下的主体，很少会做交叉类的分析
- 8、其他数据分析和挖掘的忠告
    - 数据质量的验证
        - 理解数据来源、数据统计和收集逻辑，数据入库处理逻辑
        - 理解数据在数据仓库中存放细节，包括字段类型、小数点位数、取值范围、规则约束等
        - 明确数据的取值逻辑，尤其是过程中是否对数据有转换或重新定义
        - 第一时间对数据做数据审查，包括数据有效性验证，取值范围，空值和异常值验证，是否与原始数据原则一致等
    - 数据的落地性
        - 了解业务操作流程
        - 了解目前业务的困难点和紧迫点
        - 了解业务的实际能力与权限
    - 数据事实 /= \ne /= 数据结论
        - 数据事实
            - 将数据陈列，不涉及好、坏、优、劣的定性
        - 数据结论
            - 将数据事实结合业务目标和实际情况定性为好、坏、优、劣等
    - 数据结论不要单一指标
        - 单一指标通常无法全面衡量某一业务的整体效果

- 不要预测价值立场
  - 数据分析步骤
    - 第一步定性结果
    - 第二部查找原因