数据预处理

- 方法
- 1.结构化与非结构化数据
  - 1.1 结构化数据
  - 1.2 半结构化数据
  - 1.3 非结构化数据
- 2.数据清洗（data cleaning）
  - 2.1 分析数据
  - 2.2 去除唯一属性
  - 2.3 缺失值处理
    - 缺失值的处理方法
      - 2.3.1 删除含有缺失值的特征
      - 2.3.2 补全缺失值
      - 2.3.3 不处理
  - 2.4 重复值处理
  - 2.5 异常值处理
    - 2.5.1 判断异常值的方法
      - 1.画图法
      - 2.标准差
      - 3.箱型图
      - 4.模型法
      - 5.聚类法
    - 2.5.2 处理异常值的方法
  - 2.6 噪音处理
    - 2.6.1 分箱法
    - 2.6.2 回归法
  - 2.7 数据转换（数据标准化）
  - 2.8 处理类别不平衡问题
- 3.特征工程（Feature Engineering）
  - 3.1 数据降维
    - 3.1.1 PAC的目的
    - 3.1.2 PAC优缺点
  - 3.2 特征选取
    - 与业务相关的特征选择方法
    - 与业务无关的特征选择方法
      - 1.过滤法
      - 2.包装法
      - 3.嵌入法
      - 4.遗传算法
  - 3.3 特征提取

数据预处理

方法

1.结构化与非结构化数据

1.1 结构化数据

高度组织和整齐格式的数据；

结构化数据也被称为定量数据，是能够使用数据或统一的结构加以表示的信息，如数字，符号。符号。在项目中，保存和管理这些的数据一般为关系数据库，当使用结构化查询语言或 `SQL` 时，计算机程序很容易搜索这些术语。结构化数据具有的明确的关系使得这些数据运用起来十分方便。

数据以行为单位，一行数据就是一个实体的信息，每一行数据的属性是相同的，结构化数据的存储和排列是很有规律的，这对 `CURD` 操作很有帮助。

1.2 半结构化数据

半结构化数据是结构化数据的一种形式，它并不符合关系型数据库或其他数据表的形式关联起来的数据模型结构，但包含相关标记，用来分隔语义元素以及对记录和字段进行分层。也被称为自描述的结构

半结构化数据，属于同一类实体可以有不同的属性，即使他们被组合在一起，这些属性的顺序并不重要

例如，声音，图像文件，XML，HTML文档， `JSON`

属性的顺序是不重要的，不同的半结构化数据的属性的个数是不一样的

半结构化数据有好的扩展性

1.3 非结构化数据

存储在非关系数据库中，并使用`NoSQL`进行查询,可能是文本或则非文本的，内容可能有多个维度，非结构化数据字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据。

文本文件、电子邮件、社交媒体评论、网站(`Bilibili,YouTube`),移动数据（短信，位置）、通讯（即时消息，电话录音）、媒体文件等

2.数据清洗（data cleaning）

数据清洗就是对原始数据通过丢弃、填充、替换、去重等操作，实现去除异常、纠正错误、补足缺失

在数据清洗过程中，首先需要分析数据，通过分析结果处理缺失值、异常值和重复值。最后处理类别不均衡问题

2.1 分析数据

对数据进行描述性统计分析，数据的基本情况，通过作图了解数据质量，有无异常（离群点），有无噪音等

2.2 去除唯一属性

唯一属性通常是一些id属性，

2.3 缺失值处理

缺失值的处理方法

2.3.1 删除含有缺失值的特征

直接删除带有缺失值的数据

1. 删除行记录（整行删除） 适合缺失值数量较小，并且是随机出现的，删除它们对整体数据影响不大的情况

2. 删除列字段（整列删除）缺失值如果占了95%以上，可以直接去掉这个维度的数据了

2.3.2 补全缺失值

1. 统计法
    a. 对于数值型的数据，正态分布使用均值、偏态数据使用中位数等方法补足；
    b. 对于分类型数据，使用类别众数最多的值的补足。
2. 插补法
    a. 随机法： 从总体中随机抽取某个样本代替缺失样本；
    b. 最近法：寻找与该样本最接近的样本，使用其该属性数值来补全
3. 高维映射
    a. 将属性映射到高维空间，采用独热码编码（one-hot）技术。将包含K个离散取值范围的属性值扩展为K+1个属性值，若该值属性值缺失，则扩展后的第K+1个属性值为1
    b. 优点：精确，保留了所有的信息，也未添加任何额外信息
    c. 缺点：大大增加数据的维度，计算量大大提升，且只有在样本量非常大的时候效果最好
4. 模型法
    a. 可以用回归、使用贝叶斯形式化方法的基于推理的工具或者决策树归纳确定。例如，利用数据集中其他数据的属性，可以构造一个判定树，来预测缺失值的值

2.3.3 不处理

模型对于缺失值有容忍度或灵活的处理方法

常见的能够自动处理缺失值的模型包括：`KNN`、决策树和随机森林、神经网络和朴素贝叶斯等

2.4 重复值处理

对数据集直接去重

```
df.drop_duplicates()
```

2.5 异常值处理

2.5.1 判断异常值的方法

1.画图法

优点：直观

缺点：数据量大时速度慢

2.标准差

如果数据服从正态分布，在3σ原则下，异常值为一组测定值中与平均值的偏差超过3倍标准差的值。

P(μ-1σ≤X≤μ+1σ)≈0.682

P(μ-2σ≤X≤μ+2σ)≈0.954

P(μ-3σ≤X≤μ+3σ)≈0.997

3.箱型图

`QL` 为下四分位数， `QU` 为上四分位数，`IQR=QU-QL`

当一个数值大于 `QU+1.5IQR` 或者小于 `QL-1.5IQR` 时，被称为异常值

四份位数具有鲁棒性，异常值不会对四分位数产生影响，因此箱型识别异常值比较客观，在识别异常值时有一定的优越性

4.模型法

5.聚类法

2.5.2 处理异常值的方法

1. 删除异常值---明显看出是异常且数量较少可以直接删除
2. 不处理---如果算法对异常值不敏感则可以不处理，但如果算法对异常值敏感，则最好不要用，基于距离计算的一些算法，包括 `kmeans` ， `knn` 之类的
3. 平均值代替---损失信息小，简单高效
4. 视为缺失值---可以按照处理缺失值的方法来处理

2.6 噪音处理

噪音通常是数据集中的错误值

处理方法：

2.6.1 分箱法

分箱方法通过考察数据的"近邻"(即，周围的值)来光滑有序数据值。这些有序的值被分布到一些"桶"或箱中。由于分箱方法考察的是近邻的值，因此它进行局部光滑。

- 用箱均值光滑：箱中每一个值被箱中的平均值替换
- 用箱中位数平滑：箱中的每一个值被箱中的中位数替换
- 用箱边界平滑：箱中的最大和最小值同样被视为边界。箱中的每一个值被最近的边界值替换

2.6.2 回归法

可以用一个函数拟合数据来光滑数据。线性回归涉及找出拟合两个属性(或变量)的“最佳”直线，使得一个属性能够预测另一个。多线程回归是线性回归的扩展，它涉及多于两个属性，并且数据拟合到一个多维面，找出合适数据的数学方程式，能够帮助消除噪声。

2.7 数据转换（数据标准化）

数据标准化是将样本的属性缩放到某个指定的范围

数据标准化的原因：

某些算法需要样本具有零均值和单位方差；

需要取消样本的不同属性具有不同量级时的影响：

1. 数量级的差异将导致量级较大的属性占据主导地位
2. 数量级差异将导致迭代收敛速度减慢
3. 依赖于样本距离的算法对数据的数量级非常敏感

归一化后求优过程范围变小，寻优过程变的平缓，更容易正确收敛到最优解

注意：在对测试集做特征缩放时要使用的参数μ,σ，而不能根据测试集另算一组均值方差做特征缩放。

数据标准化的方法：

- min-max标准化（归一化），把最大值归为1，最小值归为0

$$x' = (x - xmin)/(xmax - xmin)$$

- z-score标准化（规范化）将数据均值化为0，方差化为1

$$x' = (x - \mu)/\sigma$$

2.8 处理类别不平衡问题

类别不平衡：是指在分类任务中存在某个或者某些类别的样本数量远多于其他类别的样本数量的情况。

1.对大类数据采取欠采样

2.对小类数据采取过采样

3.特征工程（Feature Engineering）

特征工程是对原始数据进行一系列工程处理，将其提炼为特征，作为输入供算法和模型使用

3.1 数据降维

在数据处理中，遇到特征维度比样本数量多的情况，如果用到实际中，效果不一定好。一是冗余的特征会带来一些噪音，影响计算结果，二是因为无关的特征会加大计算量，耗费时间和资源。通常会对数据重新转换一下，再跑模型。数据转换的目的不仅是降维，可以消除特征之间的特征性，并发现一些潜在的特征变量。

3.1.1 PAC的目的

PAC是一种尽可能减少信息损失的情况下找到某种方式降低数据的维度的方法

注意：

如果变量之间的方差很大，或者变量的量纲不统一，必须先标准化再进行分析

分析通常会得到协方差矩阵和相关矩阵。这些矩阵可以通过原始数据计算出来，协方差矩阵包含平方和与向量积的和。相关矩阵与协方差矩阵类似，但是第一个变量，也就是第一列，是标准化后的数据

3.1.2 PAC优缺点

优点：

1. 以方差衡量信息的无监督学习，不受样本标签限制
2. 各主成分之间正交，可消除原始数据成分间的相互影响
3. 可减少指标选择的工作量

缺点：

1. 主成分解释其含义往往具有一定的模糊性，不如原始样本完整
2. 贡献率小的主成分往往可能包含有样本差异的重要信息
3. 特征值矩阵的正交向量空间是否唯一问题

3.2 特征选取

特征选择，多个特征中挑选出对结果预测最有用的特征。因为原始的特征中可能会有冗余和噪声

特征选择和降维的区别：

前者指是去除原本特征里和结果预测关系不大的，后者做的特征的计算组合构成新特征

与业务相关的特征选择方法

特征选择的第一步是对业务进行更好的理解及咨询，了解哪些因素（特征）会对目标有影响，较大影响和较小影响的都要。这些特征作为第一候选集

与业务无关的特征选择方法

1.过滤法

方差：定一个阈值，方差小于阈值的特征舍弃掉

相关系数：设定一个阈值，选择相关系数较大的部分特征

假设检验：

sample：卡方检验。卡方检验可以检验某个特征分布和输出值分布之间的相关性，在 `sklearn` 中，可以使用chi2这个类来做卡方检验得到所有特征的卡方值与显著性水平P临界值，可以设定卡方阈值，选择卡方值较大的部分特征

互信息：

即从信息熵的角度分析各个特征和输出值之间的关系评分。互信息值越大，说明该特征和输出值之间的相关性越大，越需要保留。在sklearn中，可以使用mutual_info_classif(分类)和mutual_info_regression(回归)来计算各个输入特征和输出值之间的互信息

2.包装法

最常用的包装法是递归消除特征法（RFE）。递归消除特征法使用一个机器学习模型来进行多轮训练，每轮训练后，消除贡献小的特征，再基于新特征集进行下一轮训练。

应用在逻辑回归的过程：用全量特征跑一个模型；根据线性模型的系数（体现相关性），删掉5-10%的弱特性，观察准确率/auc的变化，逐步进行，直至准确率/auc出现大的下滑停止

3.嵌入法

最常用的使用是使用L1正则和L2正则化来选择特征

正则化惩罚项越大，模型的系数就会越小。当正则化惩罚项大到一定的程度的时候，部分特征系数会变成0，当正则化惩罚继续增大到一定程度时，所有的特征系数都会趋于0，会发现一部分特征系数会更容易变成0，这部分系数就是可以筛掉的

一般可以得到特征系数coef或者特征重要度的算法才可以作为嵌入法的基学习器

4.遗传算法

遗传算法常用于监督式特征提取与最优化，sample ： 神经网络的最佳权重

优点：

1. 在穷举搜索不可行的情况下，对高维数据集使用遗传算法会相当有效
2. 当用到的算法需要预处理数据却没有内置的特征选取机制（如最近邻分类算法），又必须保留最原始的特征。

缺点：

实施复杂，不简洁

3.3 特征提取

1. 若干项特征加和
2. 若干项特征之差
3. 若干项特征乘积
4. 若干项特征除商