# 方法调研（一）

我们在拿到题目后，首先要判断题目所给的数据的类型，若为优化类问题，则寻找对应的优化类算法来求解，若为数据类问题，则利用数据分析的相关方法来求解。本文为数据分析类题目的方法调研。

步骤一：先将数据进行预处理，主要是对数据中的异常值和缺失值进行处理。

对异常值，一般使用箱线图法来筛选异常值，此处的异常值指的是太大或太小的值。

对缺失值，一般会有三种方法来去除，方法一为直接丢弃含有缺失值的数据，方法二为用这个数据类的平均值填入缺失位置，方法三为使用拉格朗日插值法填入差值。通常来说，方法一并不适用于大多数情况，因为丢弃的数据可能在别的维度含有重要信息，方法二是使用最多的一种方法，而方法三比方法二更符合数据的分布规律，具有更强的鲁棒性。

步骤二：对数据之间的相关性进行分析

在数据分析类题目中，若两组数据的正\负相关性在0.8以上，统计学上理解为二者具有强相关性，而当所给的数据量非常大时，两组强相关的数据只处理其中一组可节省运算的时间。

一般使用三种相关性系数来描述数据之间的相关性程度

一为皮尔森相关性系数，适用条件为：两个变量之间是线性关系，都是连续数据。两个变量的总体是正态分布，或接近正态的单峰分布。两个变量的观测值是成对的，每对观测值之间相互独立。当两个变量的标准差都不为零时

二为斯皮尔曼等级相关系数适用条件为：只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关系数来进行研究。

三为肯德尔等级相关系数，适用条件为：计算的数据是分类变量分类变量可以理解成有类别的变量，可以分为无序的，比如性别（男、女）、血型（A、B、O、AB），以及有序的，比如肥胖等级（重度肥胖，中度肥胖、轻度肥胖、不肥胖）。

步骤三：接下来便要进行变量筛选方法的选择

相关方法的介绍及优缺点如下

|  |  |  |
| --- | --- | --- |
| 方法描述 | | 优缺点 |
| 最优子集选择 | 将所有的特征组合进行建模，然后根据AIC、BIC等准则选择最优的模型，如向前和向后逐步选择 | 适用于小数据量、简单的关系，但经常表现为高方差，因此不容易获得全模型的预测误差。 |
| 压缩系数法 | 基于惩罚项变量选择方法，主要指岭回归和Lasso回归 | 当用于强相关的自变量时，在参数估计时，会导致解不可逆，十分不稳定对特征值的分解存在局限性。 |
| 降维法 | 将变量进行变换后的新变量进行降维，主要指主成分分析回归和最小二乘法 | 对特征值的分解存在局限性 |
| 树的结构方法 | 如随机森林模型，在其预测数据过程中，可对变量进行重要性排序来完成变量筛选任务 | 可以用统一的方法处理数值型变量和分类型变量 |

步骤四：在将变量筛选出来后，我们便要对数据进行预测分析

相关回归分析方法及其优缺点如下：

|  |  |  |
| --- | --- | --- |
| 方法描述 | | 优缺点 |
| 回归算法 | 采用对误差的衡量来探索因变量和自变量之间的关系，如线性回归、逻辑回归等 | 建模迅速，适用于小数据量、简单的关系，对非线性的数据拟合不好 |
| 正则化方法 | 通常是回归算法的延伸，根据算法的复杂度对算法进行调整。如LASSO、岭回归等 | 可以防止过拟合和提高模型的泛华能力，但会造成欠拟合 |
| 决策树学习 | 根据数据的属性采用树状结构建立决策模型。如分类及回归树、C4.5、随机森林等 | 具有很高的复杂度和高度的非线性关系，模型容易解释，但存在拟合倾向，运行速度慢和内存消耗高 |
| 集成算法 | 由多个相对较弱的学习模型独立的就同样的样本进行训练，将预测结果以某种方式整合起来进行总体预测。如Boosting，Bagging，AdaBoost，梯度递进等 | 当前最先进的预测几乎都采用了集成算法。精准率高于其他单个模型预测出来的结果 |

分析/评价方法比较

|  |  |  |
| --- | --- | --- |
|  | 方法描述 | 优缺点 |
| 层次分析法 | 通过专家打分法得出判断矩阵，且矩阵需通过一致性检验，根据此矩阵计算结果 | 系统性分析，所需的定量数据较少，定性成分较多，不能为决策提供新方案 |
| 熵权法 | 通过计算每一项指标的熵，并利用熵计算指标的冗余度，据此得出指标权重 | 可反应指标的区分能力，无法考虑指标之间的横向影响，对样本依赖较大 |
| 模糊综合评价 | 通过使用隶属度数值将定性指标转化为定量指标，利用熵权法计算权重并选取算子进行综合评价 | 可用过精确的数字处理模糊的评价对象，评价结果为一个矢量，包含信息丰富，可能使指标权重的确定具有主观性 |
| 主成成分分析法 | 将部分强相关的指标整合为一个指标以实现降维 | 可减弱噪声影响，提高计算效率，对异常值较为敏感，可解释性不足 |
| 秩和比法 | 计算每个对象每个指标对应的秩与RSR值，依据RSR值对对象进行分档排序 | 适用性广，对指标的选择无特殊要求，可能会丢失部分信息 |
| 优劣解距离法 | 计算每个对象每个指标距离系统中最优与最差指标的距离进行综合评价 | 适用性强，可刻画多个指标的综合影响力度，必须有两个以上的对象才可适用 |
| 灰色分析法 | 通过计算灰色系统中各项指标值与参考值的关联度和指标权重进行评价 | 可处理灰色系统的数据，参考值的确定可能会受到主观因素的影响进而影响评价结果 |
| 数据包络法 | 基于线性规划对效率进行评价分析 | 可处理多输入多输出数据，对异常值敏感，无法分析绝对效率 |