# 数学建模流程



些同学临近比赛开始紧张心慌，快比赛了觉得自己什么都不会怎么办？队友和我都不会编程建模怎么办？说到底这只是一次比赛，赛前一定要调整好自己的心态，至于怎么去学，没有比直接参赛更能促进学习的方法，比赛中遇到不会的知识再正常不过了，学会去网上搜现学现用就够了。第一次参加数学建模比赛的同学，可能不太知道数学建模拿到题目之后的整个流程是怎么样的，今天就和大家来分享一下数学建模的全流程。

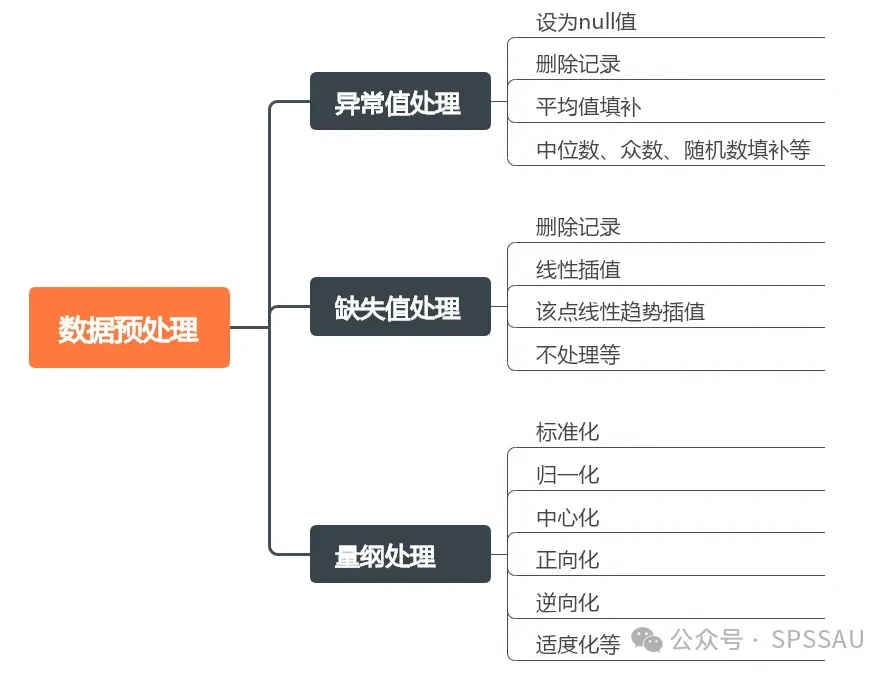
数学建模常用方法大致可分为四大部分：数据预处理、基本描述、建模、模型评价，如下图：



接下来将逐个进行介绍说明。

## 一、数据预处理

国赛的数据预处理是指进行数据的清洗和变换，预处理的质量会直接影响到后续建模和预测的准确性。数据清洗也就是我们常说的缺失值处理和异常值处理，数据变换是指对数据进行标准化、归一化等无量纲化处理，常见的数据预处理方式见下图：



1、异常值处理

异常值也称离群值，其数值明显偏离它所属样本集的其余观测值。忽视异常值的存在可能会对建模结果产生不良影响。

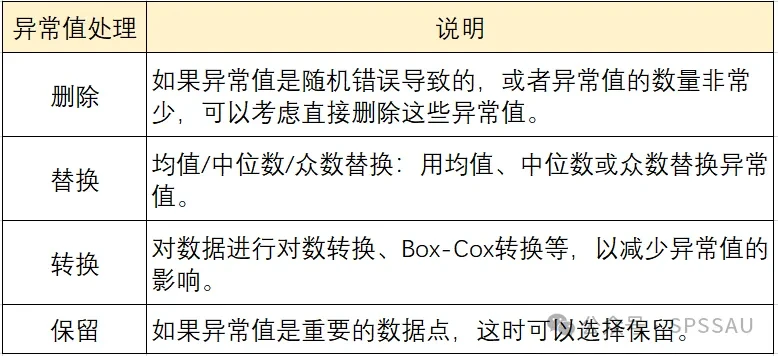
（1）识别异常值

* 可视化方法：使用箱线图、散点图等可视化工具来识别异常值。

* 统计方法：使用Z-score、IQR（四分位距）等方法来识别数据点是否为异常值。

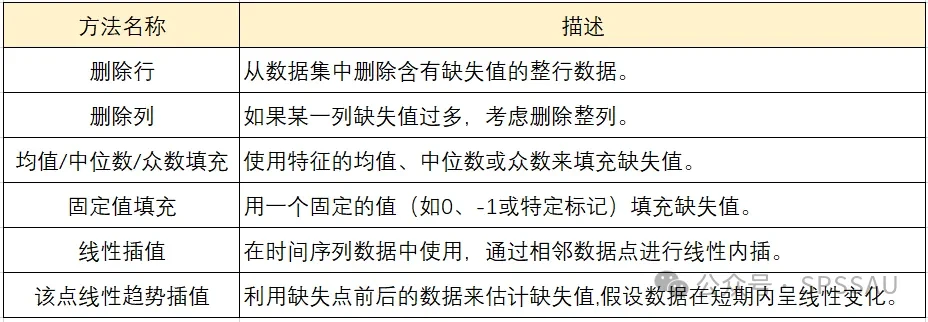
（2）处理异常值的方法

异常值处理通常有以下4种方式：

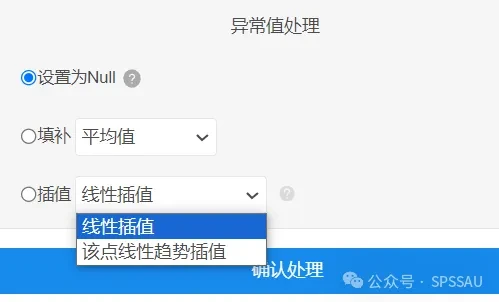


2、缺失值处理

对缺失值进行处理方法通常有删除、填充、插值3类方法，说明如下表：



对于异常值和缺失值的处理，可以使用SPSSAU【数据处理】模块的【异常值】方法进行处理，操作如下：



3、量纲处理

量纲处理是指通过数据变换来消除原始变量的量纲影响的方法，使模型更加稳定和准确。常见的量纲处理方法包括标准化、归一化、中心化、正向化、逆向化、适度化等等，说明如下：



SPSSAU提供17种数据无量纲化处理方法，在【数据处理】模块选择【生成变量】操作如下：



选择合适的无量纲化方法取决于数据的分布特征和后续建模的需求。通常需要尝试多种方法并比较效果。

## 二、描述性分析

在数学建模中，描述性统计分析是一种基本的数据探索方法，是理解数据特征的重要步骤。它帮助模型开发者理解数据的基本特征、分布情况以及变量之间的关系。

描述性统计分析步骤：

* 计算基本统计量：均值、中位数、标准差等

* 绘制可视化图表：直方图、箱线图等

* 分析数据分布：是否正态分布，是否有偏态

* 检查相关性：变量间的关系

描述性统计分析方法整理表格如下：



## 三、建模

数学建模关键的步骤就在于模型的选择与构建，根据问题的性质选择合适的模型类型，如分类模型、评价模型、预测模型等。

1、分类模型

分类模型是一种按照数学模型来分类数据的算法。它通过对给定的历史数据进行分析，根据已知的分类规则来对新的数据进行预测。常见的分类模型有聚类分析、判别分析、logistic归以及机器学习等。



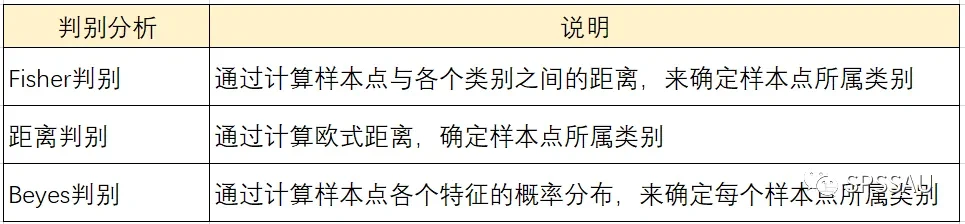
（1）聚类分析

常用的聚类分析分为K-means聚类、K-modes聚类、K-prototype聚类以及分层聚类。其中前三种聚类方法是按行聚类（R型聚类），分层聚类是按列聚类（Q型），最常用的为K-means聚类，各自适用场景说明如下表：



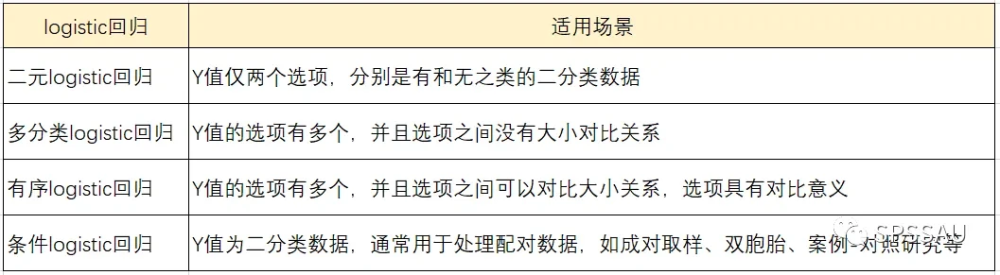
（2）判别分析

判别分析用于根据已知的分类信息对观测数据进行分类。判别分析有很多种，比如Fisher判别、距离判别、Beyes判别等，其中Fisher判别使用频率最高。



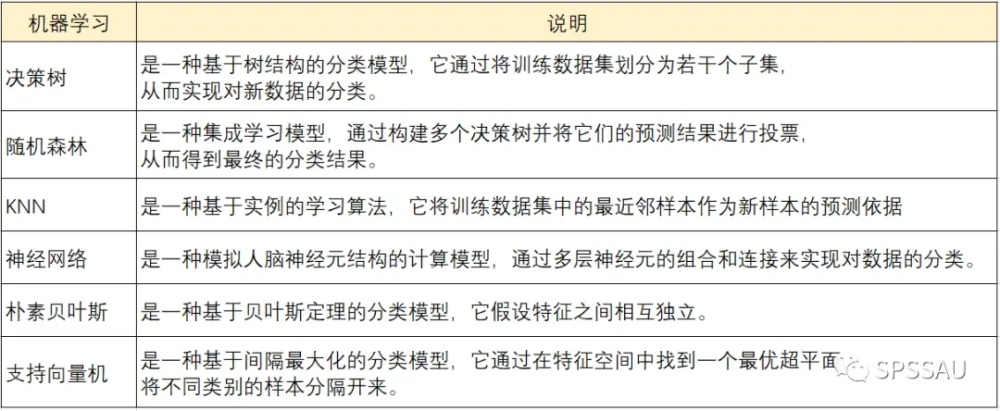
（3）logistic回归

logistic回归可以用于分类，它的核心思想是利用逻辑函数将线性回归的结果转化成一个概率值，这个概率值可以用来进行分类。logistic回归分析可细分为二元logistic回归、多分类logistic回归、有序logistic回归。各自适用场景说明如下表：



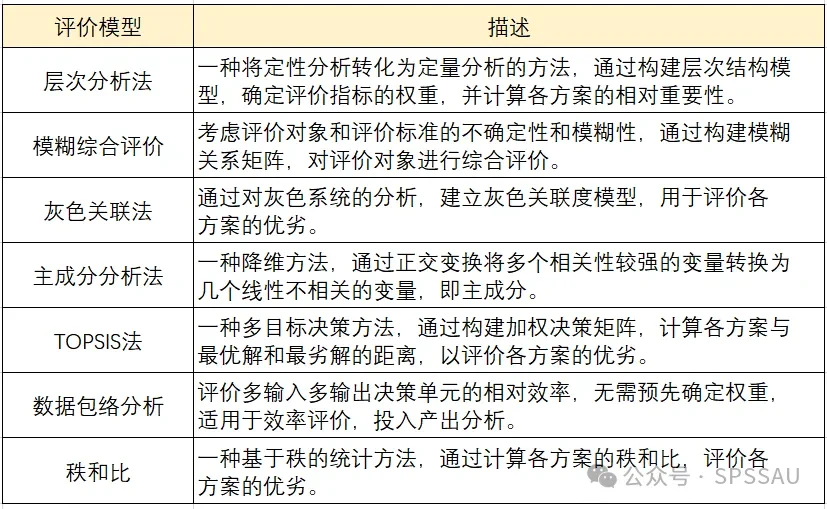
（4）机器学习

机器学习算法中常用于分类的模型包括决策树、随机森林、KNN、神经网络、朴素贝叶斯、支持向量机等等。具体说明如下：



2、评价模型

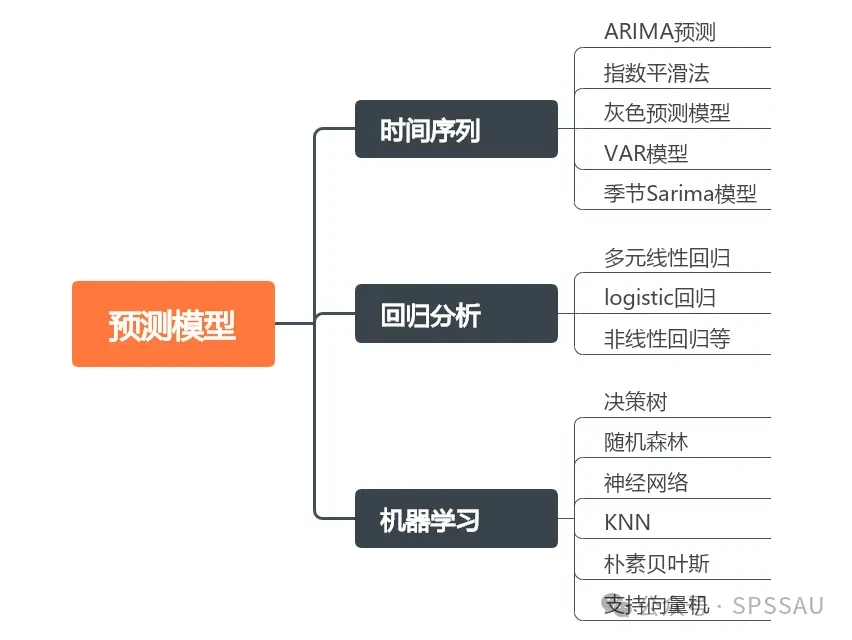
在数学建模中，评价模型通常指的是对构建的模型进行评估，以确定其性能和适用性。评价模型的目的是确保模型能够准确地反映实际问题，并且能够提供有效的决策支持。在评价模型中，常用的方法如层次分析法、模糊综合评价、灰色关联法、主成分分析法、TOPSIS法、数据包络分析、秩和比等等，如下表：



除以上单一评价模型以外，还可以使用多种方法进行组合评价

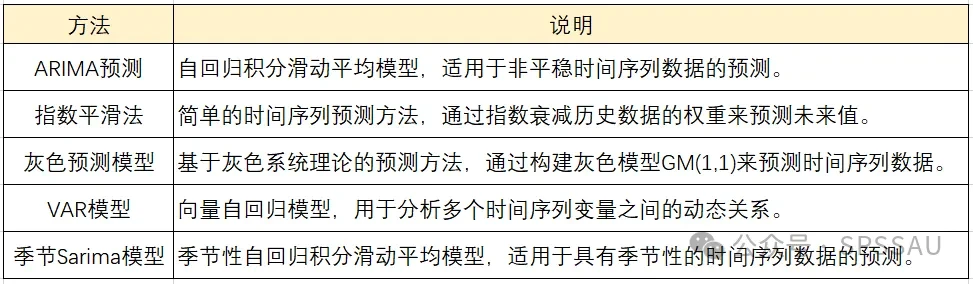
3、预测模型

数学建模预测模型是一种利用数学模型和统计方法来预测未来可能发生的结果的模型。它通常基于历史数据和已知信息，通过建立数学模型来分析问题，并预测未来的发展趋势。预测类常见的方法有时间序列类预测、回归分析进行预测、机器学习进行预测、马尔科夫预测或者其它方法组合预测等。



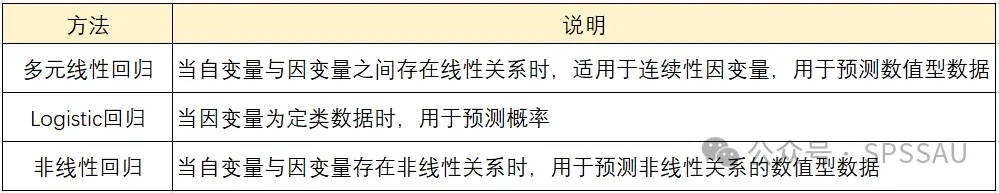
（1）时间序列预测

时间序列数据是按照时间顺序排列的数据集合，例如每天的销售量、每月的股票价格等。预测时间序列数据可以帮助我们了解未来的趋势和模式，从而做出更准确的决策。比较常用的时间序列数据预测方法有ARIMA预测、指数平滑法、灰色预测模型、VAR模型、季节Sarima模型，说明如下：



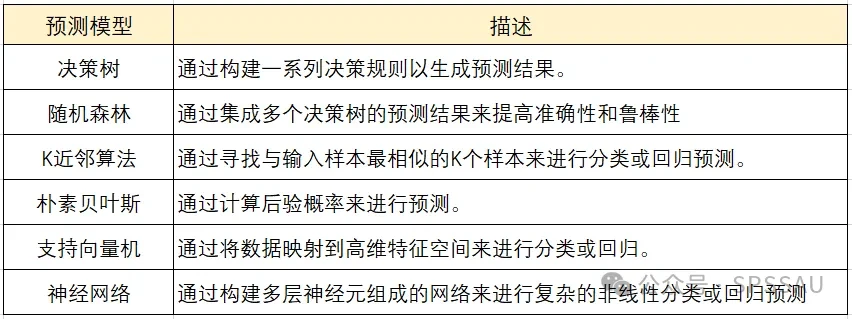
（2）回归分析预测

回归分析是一种常用的统计方法，用于建立变量间的关系模型，并通过该模型对未知数据进行预测。常用方法如多元线性回归、logistic回归、非线性回归等，说明如下：



（3）机器学习预测

机器学习用于从数据中学习模式和规律，并利用这些知识进行预测。通过训练算法来自动发现数据中的模式，并根据这些模式进行未知样本的预测。常用方法及说明如下：



## 四、模型评估

在进行模型构建与分析时，模型评价是非常关键的一步。合适的评价指标可以帮助我们准确地衡量模型的性能，从而进行优化和改进。然而，不同的模型和应用场景需要使用不同的评价指标。下图为回归模型和分类模型常用评价指标：

