智能容量规划报告

作者：李哲

概述：报告分为三个部分，第一部分主要介绍数据的处理方法与数据描述，第二部分介绍建模过程中使用的三个模型，第三部分介绍模型的选择与最终的预测结果

第一部分：数据处理

一、数据概述

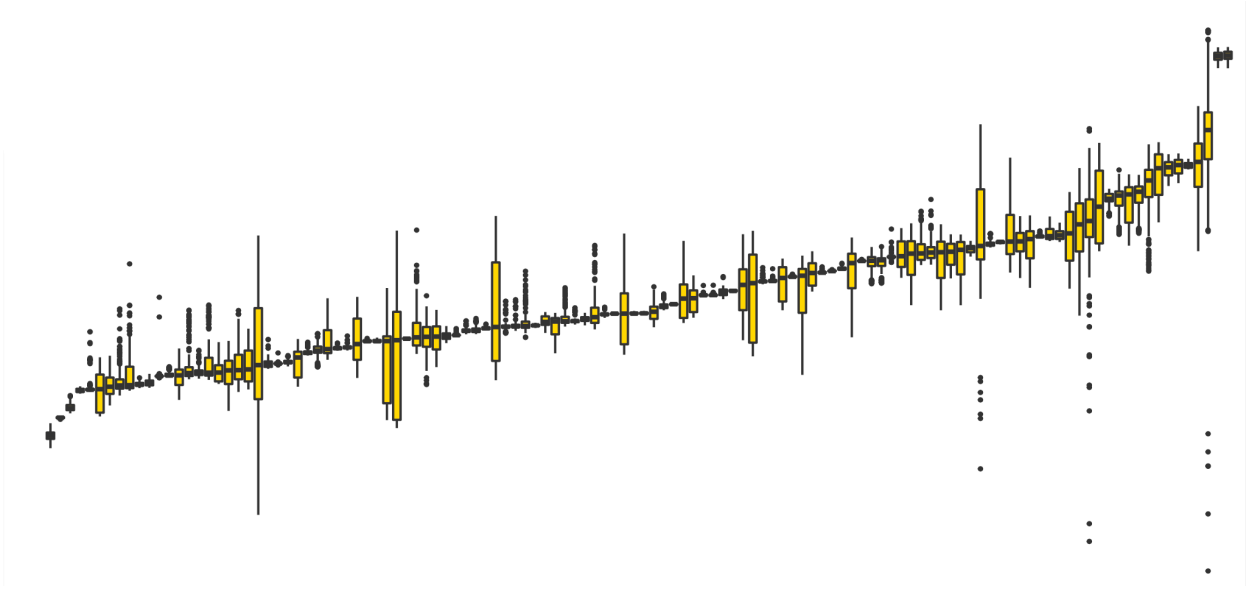
数据为120个服务器7天（7×24小时）的使用率数据。其中，有5行数据出现大量缺失的情况，需要对缺失数据进行补全操作

二、缺失值补全

通过观察含有缺失值的5行数据特点发现，出局具有周期性缺失的特点，即每天（每24个点）的相同时间段缺失，采用**“线性预测+白噪声”**的方法补全数据

三、数据描述

补全数据后，作出120个服务器利用率箱线图如下图所示：



从上图中可以看出，120个服务器的数据差异较大，大部分服务器的数据变化范围较小，少数服务器的数据变化范围较大，部分服务器有离群值较多的特点。

第二部分：模型介绍

对数据的建模共使用了三种模型：

- 周期因子模型（ `R` 实现）
- $ARIMA$模型（ `R` 实现）
- $LSTM$模型（ `Python` 实现）

训练模型过程是将前6天数据作为训练集，第7天数据作为测试集，计算各模型的$MAPE$，选取最佳模型

一、周期因子模型

- 第一步：将数据转化为6x24的数据框，一行代表一天的数据
- 第二步：计算每行的均值，并将这行的24个数据除以该均值
- 第三步：按列选取中位数，得到周期因子
- 第四步：计算前6天所有数据的均值，得到base
- 第五步：利用base乘以周期因子得到第七天的预测数据
- 第六步：计算MAPE

二、ARIMA模型

- 第一步：对每一个服务器数据找出最优ARIMA模型中的**自回归阶** `p` ，**差分阶** `d` 以及**移动平均阶** `q`
- 第二步：利用第一步得到的参数建立训练模型
- 第三步：预测第7天数据
- 第四步：计算MAPE

三、LSTM模型

采用滑动窗口的LSTM方法预测，即输入24个点，预测下一个点

- 第一步：训练集数据处理，使之满足滑动窗口要求
- 第二步：设置参数，训练模型（**具体参数设置见代码**）
- 第三步：采用滑动窗口的LSTM方法预测第7天数据
- 第四步：计算MAPE

第三部分：模型选择与预测

通过比较三种模型的$MAPE$并结合测试集的预测情况发现**周期因子模型**为最佳模型，因而选择该模型进行预测，将168个数据放入模型，得到72个预测点，具体预测数据见**"generate_data"**