工业大数据技术与应用实验内容

主题：对大规模电信系统数据进行分析及预测

数据：见 QQ 群（两个月的数据，10 分钟时间尺度，五种业务）

内容一：探索性分析（用到的工具/包为 matplotlib 或者 seaborn 或者 MATLAB）

1）画出 1045、5045、8000 这三个小区的流量值随时间的分布图，任意画出一周的流量分布即可，无需画出所有两个月的长度；

2）画出这 100*100 个小区中，横轴 45:55，纵轴 45:55，这 10*10 个小区的流量值随时间的分布图（注：任意画出一个时刻的即可，即：只给出一个热度图，自己去计算这 100 个小区在原始数据中的位置（索引））。

内容二：数据库存储（用到的工具为 MySQL 或 HBase 或本地 CSV 文件）

1）对 5045 小区的 Interent 流量数据，以滑动窗口大小为 4，预测大小为 1，进行数据准备（构建输入特征和预测目标），并且将准备好的数据存储到数据库 MySQL 或非关系型数据库 HBase 或写入到本地 CSV 文件中。

内容三：预测模型构建（用到的工具为 scikit-learn）

1）对 5045 小区的 Internet 流量数据进行预测（内容二构建的数据集中，前 70%为训练数据集，后 30%为测试数据集），以 scikit-learn 工具包进行机器学习模型的构建，模型可选择逻辑回归、Lasso、支持向量机、决策树、多层感知机等。

内容四：预测结果展示及分析（用到的工具为 matplotlib 或者 MATLAB）

1）画出 5045 小区的预测值跟真实值的对比，得出量化结果，结果以 MSE 和 MAE 表示；

2）简单分析结果，比如：预测的 MSE 是多少，在哪些时刻预测的好，在哪些区域预测的好等。