工业大数据技术与应用实验大纲

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

1）画出 1045、5045、8000 这三个小区的流量值随时间的分布图；

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：自己去计算这 100 个小区在原始数据中的位置（索引）；

3）画图 5045 小区的时间自相关图；

4）画出 5045 这个小区与周边 10*10 个小区的皮尔逊相关系数（具体见课堂 ppt 中对空间相关性的说明和讲解）。

内容二：非关系型数据库存储

1）对 5045 小区的流量数据，以滑动窗口大小为 4，预测大小为 1，进行数据准备，并且将准备好的数据存储到一种非关系型数据库中（推荐 HBase 或 MongoDB）；

2）对所有小区的一种业务数据，以滑动窗口大小为 4，预测大小为 1，进行数据准备，并且存储到一种非关系型数据库中（推荐 HBase 或 MongoDB）；

内容三：预测模型构建

1）将数据根据时间索引划分成训练集（70%）和测试集（30%），比如有 10 天的数据，那么就选择前 7 天的数据作为训练集，后 3 天的数据作为测试集；

2）对 5045 小区的测试集进行预测，可以先不运行在 Hadoop 上，而是以 scikit-learn 工具包进行机器学习；

3）利用 Hadoop Spark 的 MLlib 包对数据进行训练并预测。

内容四：预测结果展示及分析

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）画出这 100*100 个小区，任意一个时刻，真实值（100*100 的一个矩阵）以及预测值（100*100 的一个矩阵）的对比；

3）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。