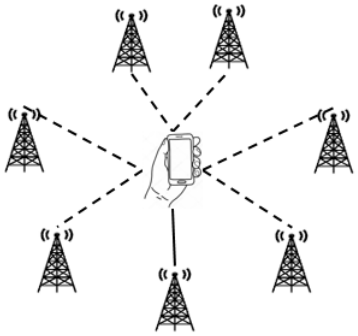
软件学院 数据分析/挖掘课程编程作业

1. [手机信号数据集] 本次作业利用手机与基站连接信号强度的测量报告数据(measurement report: MR)来预测手机所在 GPS 经纬度位置，通过利用 tensorflow 或 Keras 软件包训练 CNN/RNN 模型，预测 MR 对应得经纬度位置，绘制误差概率分布图，确定中位误差；并将预测后的经纬度位置投射至百度地图进行可视化显示。要求如下：
   a) [CNN 回归模型：6 分]将每个 MR 样本看成是一个 2 维度矩阵，其中矩阵列为 MR 样本中所含邻近基站、行为 MR 所含临近基站信号特征，训练 MR 样本 2 维矩阵与其对应经纬度位置的 CNN 回归模型，预测 MR 测试样本的经纬度位置；
   b) [CNN 多分类模型 4 分]若将 MR 所在区域划分为 M*N 栅格，使得每个 MR 样本对应经纬度位置落于某个特定栅格，训练 MR 样本 2 维矩阵与其对应经纬度位置所在栅格的 CNN 多分类模型，预测 MR 测试样本在多栅格的概率，其中概率最大栅格的中心点可以作为 MR 测试样本位置；
   c) [LSTM 回归模型 6 分]若考虑将 MR 样本数据集按照 IMSI 进行分组，然后对每个分组的 MR 样本按照 timestamp 进行逆向排序（第一个样本所含 timestamp 最早），以生成该 IMSI 对应的 MR 样本序列数据；同时将 MR 样本序列数据所含每个 MR 样本的 GPS 位置连接起来形成轨迹信息，训练 MR 样本序列数据与对应 GPS 轨迹的 RNN/LSTM 模型，预测测试 MR 样本序列的 GPS 轨迹，利用预测的 GPS 轨迹获取测试 MR 样本序列中每个 MR 样本对应的 GPS 坐标；
   d) [LSTM 多分类模型 4 分]利用 c）组织 MR 样本序列数据，若考虑 GPS 位置对应栅格、构建 GPS 位置栅格序列，训练 MR 样本序列数据与对应 GPS 位置栅格序列的 RNN/LSTM 模型，预测测试 MR 样本序列的 GPS 位置栅格序列，利用 GPS 位置栅格输出每个 MR 样本对应的 GPS 坐标；
   e) [CNN/LSTM 多分类混合模型 10 分]按照上述步骤 d)组织 MR 样本序列数据和 GPS 位置栅格序列，首先通过 b)对 MR 序列中的 MR 样本和对应 GPS 位置构建 CNN 多分类模型，输出 MR 样本在多栅格的概率向量，针对 MR 序列对应的概率向量序列，通过 d）LSTM 回归模型训练 MR 序列的概率向量序列与 GPS 位置栅格序列，预测测试 MR 样本序列的 GPS 栅格序列，输出 MR 样本对应的 GPS 坐标。
   f) [加分题：Autoencoder/LSTM 多分类混合模型 10 分] 利用 autoencoder 对 d) MR 样本序列数据 encoder 处理、联合训练编码序列与 GPS 位置栅格序列对应关系，预测测试 MR 样本序列的 GPS 位置栅格序列。
   g) [文档+答辩：10 分] 上述所有内容整理写 word 文档，然后在第 17 周答辩。



2. **数据说明：** 一台可以正常通信的手机，每时每刻都保持着和附近基站的连接，这些连接的信息组成了测量报告数据，也就是 MR 数据。MR 数据中主要包括：相连基站的 ID 以及对应的信号测量信息。它们和手机的上下文信息共同组成了 MR 数据。一条典型的 MR 数据格式如下：

	IMEI	IMSI	MRTime	
RNCID_1	CellID_1	Dbm_1	AsuLevel_1	SignalLevel_1
…	…	…	…	…
RNCID_6	CellID_6	Dbm_6	AsuLevel_6	SignalLevel_6
Longitude	Latitude	Altitude	Accuracy	Speed

其中， IMEI 和 IMSI 共同组成手机的唯一 ID，MRTime 为当前时间戳；RNCID 和 CellID 共同组成基站的全球唯一 ID，对应的基站 GPS 坐标可以在工参表中查询得到，Dbm，AsuLevel 和 SignalLevel 是手机和基站间的信号测量数据；Longitude 和 Latitude 是 GPS 位置标签，手机在报告位置同时，也会给出定位的理论误差 Accuracy 以及理论高度 Altitude；Speed 是当前的运动速度，单位 m/s。注意，数据中存在缺失值，以-999 和-1 的形式存在。

GPS 标签的位置范围是：左下角(31.28175691, 121.20120485)，右上角(31.29339344, 121.21831882)，可从数据中统计得到。若转化成 UTM 坐标，则：左下角(328770, 3462224)，右上角(330420, 3463487)。可根据计算需要适当扩大范围。

3. 提交说明

train2g.csv 中给出训练数据，test2g.csv 测试数据不带标签，gongcan.csv 给出基站 GPS 信息。

对于每一小问，输出 pred.csv 文件，每一行对应 test2g.csv 中每一条 MR 数据的预测 GPS 位置(逗号分隔，可参考样例输出)。

·GroupID_TeamLeaderID:

a

src/

pred.csv

b

src/

pred.csv

…