# 软件学院 数据分析/挖掘课程作业之3

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，其中通过MR信号强度数据提取相关特征信息，训练sk-learn 的分类器，预测GPS经纬度位置; 通过绘制误差概率分布图，确定中位误差。要求如下：
   1. [5分]利用sk-learn 提供高斯朴素贝叶斯分类器Gaussian Naive Bayes (GaussianNB)、K近邻分类器KNeighborsClassifier、决策树分类器DecisionTreeClassifier等3个分类器及集成学习方法AdaBoostClassifier等6个分类器以及随机森林分类器(可参考[SKLearn RandomForestClassifier](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html))共计7个分类器来训练MR数据“信号强度特征”与和“对应GPS经纬度所在栅格”的分类模型、预测MR测试数据未知GPS经纬度位置所在栅格、最后计算预测的栅格中心位置作为该MR记录的位置。随机选取80%的数据记录作为训练集，余下20%作为测试集合，计算预测位置和证实位置的误差（采用欧式距离），按照计算误差从小到大进行排序。重复10次训练集/测试集的选择和误差计算，求平均误差，绘制平均误差概率分布图，其中x-轴为排序编号，y-轴为对应的平均误差，中位误差为x-轴50%，排序编号的y轴对应平均误差。对每个分类器在测试集上的结果，计算precision，recall和f-measurement来评价分类器的好坏。本次作业提供2G GSM和4G LET网络的数据集，要求对比和讨论这两个不同数据集合的定位结果，讨论和比较上述7个训练模型的精度。其中MR信号强度数据特征信息可参考如下工作
      1. 参考文献1：https://dl.acm.org/citation.cfm?doid=2983323.2983345：Fangzhou Zhu, Chen Luo, Mingxuan Yuan, Yijian Zhu, Zhengqing Zhang, Tao Gu, Ke Deng, Weixiong Rao, Jia Zeng: City-Scale Localization with Telco Big Data. CIKM 2016: 439-448
      2. 参考文献2：Yukun Huang, Weixiong Rao, Fangzhou Zhu, Ning Liu, Mingxuan Yuan, Jia Zeng, Hua Yang: Experimental Study of Telco Localization Methods. MDM 2017: 299-306， https://ieeexplore.ieee.org/document/7962466/
   2. [加分题：5分] 若考虑到MR数据中移动设备的时间空间约束（如移动速度），那么对**同一个移动设备**时间邻近的MR记录对应的位置存在一定的空间距离约束，利用这个特性设计新的算法来优化上述定位算法。
   3. [5分]按照每个MR记录对应的主基站对MR记录进行分组，使得每组MR记录都有相同的主基站，总的分组个数即为主基站个数。假定某主基站的经纬度坐标为<x0, y0>，该分组中的某MR记录对应的经纬度坐标为<x, y>，则计算该MR记录与主基站的**相对位置**为：<x‘ y‘>=< x-x0, y-y0>。在完成每个分组的MR记录相对位置计算之后，针对每个分组构建一个对应的MR定位模型，不过该模型是以MR记录与主基站的相对位置作为标签。使用处理好的训练集用于训练模型，测试数据集用于测试统计，通过上述**随机森林**模型预测测试数据的相对位置<x‘’，y‘‘>，然后计算还原为原始位置：<x‘’+ x0，y‘‘+ y0>；比较和讨论a)和c）方法之间的优劣。
   4. [5分]假定存在K个主基站，也即存在K个对应的MR记录分组和MR定位模型；根据c)中计算每个分组定位模型的中位误差进行排序，取得中位误差最小的前top-k分组（记为topk+）和中位误差最大的后top-k分组（记为topk-），其中k=，尝试利用topk+分组中的MR数据融入到topk-分组中，重新处理c）的定位步骤，比较和讨论d)和c）方法之间的优劣
   5. [加分题：5分] 若尝试构建MR分组之间的相似性，为记为topk-分组选择k=个最相似的其他MR分组，将这些相似MR分组记录融入到topk-分组，重复重新处理d）的定位步骤，比较和讨论d)和e）方法之间的优劣

**Tips：划分栅格：**

1. 利用所有MR数据的GPS labels，确定一个整体的位置范围，并将该范围转换成一个大的矩形（所有数据的位置标签均落入该矩形）。该矩形的坐标范围为:

左下角坐标[121.2086601, 31.28380832] 右上角坐标 [121.2143275, 31.29248568]

再将该矩形划分成若干个等大的小正方形栅格（边长为20m左右较为适合）。分类器训练时，将所有MR数据的GPS labels转换为栅格ID labels，并划分训练集和测试集进行训练。测试时，分类器输出对每一条MR数据预测的栅格ID。

1. 计算误差时，将分类器预测的栅格ID转换成该栅格的中心点坐标，并用该中新点坐标代表预测出的位置。最后用真实的GPS label进行误差计算。
2. 计算precision，recall和f-measurement时，直接用预测得到的栅格ID和由真实GPS转换得到的栅格ID标签进行比较计算。
3. [交易数据集挖掘] 利用**trade\_new.csv**的trade文件,
   1. [6分]利用训练数据中每个用户前K-1个月的交易记录构建交易特征，第k个月用户购买行为构建该月用户购买行为类别，通过机器学习算法**训练用户前K-1个月的交易特征**和第**k个月的购买行为标签**购买相应的机器学习模型，测试用户在第k+1个月的购买行为。trade\_new.csv里是2016.2 ~ 2016.7月的数据，最后要求**利用2016.5~2016.7（前三个月）的数据提取特征，预测用户在2016.8月的购买行为（数据未给出）**，具体要求见后续说明。用户交易特征构建，可以参考文献：
      1. Guimei Liu, Tam T. Nguyen, Gang Zhao, Wei Zha, Jianbo Yang, Jianneng Cao, Min Wu, Peilin Zhao, Wei Chen: Repeat Buyer Prediction for E-Commerce. 155-164 https://dl.acm.org/citation.cfm?doid=2939672.2939674
   2. [方法1：4分]首先，利用上述特征，设计用户在第k+1个月存在购买商品IDs的多分类问题回归目标值；以利用题目1（a）的分类器构建用户交易特征与用户购买商品IDs之间的多分类模型，测试和讨论该模型的准确性。
   3. [方法2：总计12分]包括如下4个步骤：
      1. [2分]首先预测用户是否会在第k+1个月存在购买可能，那么购买行为类别标签为YES和NO，即在训练过程中，某用户在第k个月存在购买记录，其对应购买行为类别标签为YES，否则为NO。通过利用题目1（a）的分类器进行建模和测试，并比较各个分类器的优缺点；
      2. [2分]其次，针对用户在第k+1个月存在购买行为， 以用户购买商品品牌(brand)作为分类器（多）类别标签，构建用户交易特征与用户购买商品品牌（多）类别标签之间的机器学习模型，重复上述a）预测；
      3. [2分]第三，针对用户在第k+1个月存在购买行为，以用户购买商品品类(category)作为分类器（多）类别标签，构建用户交易特征与用户购买商品品类（多）类别标签之间的机器学习模型，重复上述a）预测
      4. [2分]第四，针对用户在第k+1个月存在购买行为，以用户购买商品总体耗费金额作为回归问题的目标值；构建用户交易特征与用户购买金额的机器学习回归模型，重复上述a）预测。注意：以上步骤机器学习模型均为分类器模型，均将其变为对应的回归模型，如将随机森林分类器(可参考[SKLearn RandomForestClassifier](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html))转为为随机森林回归([SKLearn RandomForestRegressor](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html" \l "sklearn.ensemble.RandomForestRegressor))。
      5. [加分：6分]若针对c（i-iv）的结果，考虑如下优化问题：以用户购买金额为约束条件，以用户购买可能性、用户购买品牌和用户购买品类，求解用户购买商品编号和数量的最优解；讨论和分析该方法和b）之间的优缺点。

提交日期：2018/06/03日 23：59PM,提交内容发送至tongjidam18@163.com，提交内容包括：

1. 每个作业提交内容以学号+hw3.zip作为命名方法；其中包括每个小题的子目录，命名方式分别为对应小题的序号，每个子目录包括对应目的代码和word报告。其中报告包括1）代码运行结果屏幕拷贝；2）讨论分析部分；3）性能比较图表

## 第2题的说明:

1. 对数据中bndno字段的处理：bndno为空的使用pandas.DataFrame.fillna(-1)填充为-1
2. 对于b和c (i ~ iv)这五个小题，输出文件，基本文件名格式为：  
   [学号]\_2[小题编号]\_[sklearn模型名称].txt  
   例1：  
   学号1451228，b)小题使用高斯朴素贝叶斯分类器Gaussian Naive Bayes进行分类，该分类器在sklearn中的包路径为sklearn.naive\_bayes.GaussianNB，则输出文件名为：  
   1451228\_2b\_ GaussianNB.txt  
   例2:  
   学号1451228，c)-iv小题使用随机森林回归，其sklearn包路径为sklearn.ensemble.Rando mForestRegressor，则输出文件名为：  
   1451228\_2civ\_ RandomForestRegressor.txt
3. b) 预测用户在第k+1个月会购买哪些商品，对于测试集中每一对<用户, 商品>对，输出其在下个月是否购买，Yes表示购买，No表示不购买，每一行对应一对<用户, 商品>对及其标签，三个字段以逗号分隔: [用户], [商品], [标签]  
   例：  
   781924, 10113009, Yes  
   781924, 11302032, No
4. c)-i 预测用户是否会在第k+1个月购买：要求对给定测试集，对于测试集里面的每一个用户，输出Yes或者No，将结果输出到txt文件，输出格式为每行对应一个用户和其标签，用逗号分隔：[用户], [标签]  
   例：  
   2900003115009, Yes  
   1590140305107, No  
     
   c)-ii 对于给定测试集，对于测试集里面每一个<用户, 品牌>对，输出其在下个月是否购买该品牌，Yes表示购买，No表示不购买，输出每一行对应一对<用户, 品牌>对及其标签，三个字段以逗号分隔: [用户], [品牌], [标签]  
   例：  
   2900003115009, -1, Yes  
   1590140305107, 34224, No  
     
   c)-iii 同c)-ii，将品牌换成品类  
     
   c)-iv 对于给定测试集，对测试集里面的每一个用户，输出其下个月的预测消费金额，输出每一行对应一个用户和其消费金额，用户和消费金额之间以逗号分隔：  
   [用户], [消费金额]  
   如果判断用户下个月不购买，则冒号后面消费金额为0  
   例：  
     
   2900003115009, 200.5  
   1590140305107, 56.7  
   2900001436410, 0  
     
   c)-v 本题命名格式中不包含模型名称，示例：1451228\_2cv.txt  
   对于给定测试集，对测试集里每一个用户，输出其在下个月购买的商品及数量，输出每一行对应一个用户及其所有预测会购买的商品及数量，输出格式为用户和<商品, 数量>对之间以两个冒号分隔，<商品, 数量>对之间以逗号分隔，同一对中商品和数量以冒号分隔：  
   [用户]:: [商品1]:[数量1], [商品2]:[数量2], [商品3]:[数量3]  
   如果判断用户下个月不购买，则双冒号后面没有内容  
   例：  
   2900003115009:: 11054040:2, 15110032:1  
   1590140305107::