软件学院 2020 年数据分析/挖掘课程编程作业 2（30 分）

1. [交易数据] 本次作业利用交易数据集合，开展针对商品销量的预测分析。

字段名称: uid, **sldatime**, pno, cno, cmrid, **vipno**, id, **pluno**, bcd, pluname, spec, pkunit, dptno, dptname, **bndno**, bndname, **qty**, **amt**, disamt, ismmx, mtype, mdocno, isdel

示例记录 1:16072913541329219, **2016-07-29 13:54:22**,13,8323,男[45 以上], **2900003115009**, 2, **22002240**, 200328600506004228, 红油桃（中）, , 千克, **22002**, 桃,,,0.422,5.06,0.0,0,,,0

示例记录 2:16060809581811553, 2016-06-08 09:58:40, 18,8334, 女[18-25], 2900001575201, 5, 34150006, 6926458841290, MSU 男童平脚裤 74129, 1*1, 盒, **34150**, 男童裤, 34224.0, 真想你, 1.0, 27.9, 0.0,0,,,0

字段说明:订单编号,**购买时间**,收银员编号,收银机编号,性别年龄,**会员编号**,商品单内编号,**商品编号**,条码,商品名称,包装规格,商品单位,商品类型编号,商品类型名称,品牌编号,品牌名称,购买数量,**金额**,是否打折,是否促销,促销类型,促销单号,是否更正

商品类别结构 **pluno 22002240**：商品类别结构可由商品编号构建，商品编号的前两位，前三位，前四位，前五位为商品逐渐细化的品类。例如"红油桃"的商品编号为 22002240，则其品类由粗到细为 22：蔬果课；220：水果；2200：实果类：22002：桃。

a) 数据预处理：首先要求对每个商品编号/商品类别结构/品牌编号等按照购买时间对购买数量进行汇总求和，包括按天、周、约为周期分别形成该商品编号/商品类别结构/品牌编号对应购买数量汇总的每天、每周等 3 个时间序列数据，该时间序列数据按照商品编号分组对应的时间信息进行排序。

b) 特征工程：对每个商品编号，设计如下特征信息：

   i. 商品编号、品牌编号、4 级品类结构、日期（标记为 d）、是否工作日、当日销量、前 1 日(即 d-1 日)至前 7 日(即 d-7 日)当天销量，计 16 个特征；

   ii. 该商品对应品牌前 1 日至前 7 日当日销量，计 7 个特征；

   iii. 该商品对应 4 级品类前 1 日至前 7 日当日销量，计 28 个特征

   iv. 该商品前第 2 周（即 d-8 日至 d-14 日）、前第 3 周（即 d-15 日至 d-21 日）、前第 4 周（即 d-22 日至 d-28 日）中该周的每日销量平均值、该周的某日最大值、该周的某日最小值，计 3*3=9 个特征；

   v. 该品牌前第 2 周（即 d-8 日至 d-14 日）、前第 3 周（即 d-15 日至 d-21 日）、前第 4 周（即 d-22 日至 d-28 日）中该周的销量每日平均值、该周的某日最大值、该周的某日最小值，计 3*3=9 个特征；

   vi. 该 4 级品牌前第 2 周（即 d-8 日至 d-14 日）、前第 3 周（即 d-15 日至 d-21 日）、前第 4 周（即 d-22 日至 d-28 日）中该周的销量每日平均值、该周的某日最大值、该周的某日最小值，计 3*3*4=36 个特征；

c) 未来销量预测：针对训练数据中商品每天的**当日销量**为目标特征、其他特征（即历史信息）均为属性特征，利用 SVM、随机森林、MLP 等 3 个方法进行建模，预测测试数据中某商品对应日期当日（标记为 d'）至第 6 日（d'+6）共计 7 天的每日销量，可考虑如下算法：首先完成商品 d' 当日的销量预测，然后利用该预测销量更新上述 b)的相关特征，继续预测 d'+1 当日销量。。。重复该步骤，直至完成第 6 日（d'+6）当日销量预测。

d) 性能评测

   i. 在 a）的每日时间序列数据中，对每个商品按照安排时间从早到晚的顺序排列，分别选取该商品 80%和 20%d 的时序数据作为训练和测试数据

   ii. 对比①仅使用 b.i 特征、②仅使用 b.i+b.iv 特征、③仅使用 b.i+b.ii+b.iii+b.iv 特征、④使用

b.i+b.ii+b.iii+b.iv+b.v+b.vi 特征等 4 类场景的性能对比，并加以讨论。

iii. 指标：root relative squared error (RSE)，见参考文献的公式(5).

2. 数据集下载地址,如 hw1：

   链接：https://pan.baidu.com/s/18xjDDjcZYY6yqsbecDrkOw 提取码：ng6a

3. 提交方式：

   提交日期：2020/06/20 日 23：59PM，提交内容发送至 tongjidam20@163.com，每个作业提交内容以学号+hw2.zip 作为命名方法；其中包括 3 个子目录，命名方式分别为 q1,q2 和 q3，每个子目录包括对应目的代码和 word 报告。

   q1: 数据预处理结果

   q2: 特征工程处理结果

   q3: 面向 4 类场景的 3 个机器学习模型（SVM、随机森林、MLP）预测结果对比，比较不同特征场景和不同学习模型的讨论分析部分，绘制参考文献 Table3 和 Figure 4 的多步预测结果。

4. 参考文献：Jiaming Yin, Weixiong Rao, Mingxuan Yuan, Jia Zeng, Kai Zhao, Chenxi Zhang, Jiangfeng Li, Qinpei Zhao: Experimental Study of Multivariate Time Series Forecasting Models. ACM CIKM 2019: 2833-2839, link: https://dl.acm.org/doi/10.1145/3357384.3357826 (free access).