程序说明

程序流程简单说明：

1. 特征提取：特征分析，然后提取，主要是 6 类特征：借书数量、教室消费次数、图书馆刷卡天数、早起晚归时间、自习天数、成绩，共 110 个左右。
2. 预处理：归一化、标准化、缺失值填充
3. 两类模型：
   1) 回归模型：Linear Regression，LR
   2) 各种分类模型 SVM, LR, Lasso
4. 模型组合：Bagging，AdaBoosting
5. 误差测量：Spearman 相关系数

程序流程的简单解释：

1. 特征提取：学习时间离第 3 学期的考试越近，越能提升成绩，而且需要提前预测第 3 学期的成绩排名，所以需要把提取的关于学习的特征以学期和具体时间段分组；消费数据只取教室消费是因为教室消费与成绩关系最大，而且经测试后取教室消费和全部消费效果是一样的，为了简化计算，取教室消费，取自习天数等天数特征而不取具体自习时间也是基于同样的考量。更多特征见具体程序。
2. 预处理：为了对付那些标准差相当小的的特征并且保留下稀疏矩阵中的 0 值，也为了减少大方差数据的影响，保证概率或系数之和等于 1，将特征等归一化、标准化。特征有缺失值是因为学生没有具体记录，所以以 0 填充即可。
3. 模型选择：选择 Linear Regression 主要是因为第 3 学期的成绩大致可以用前两个学期的成绩预测；由于提取了 110 个特征，所以模型中必然有很多接近 0 的参数，因为 Lasso 模型惩罚函数是 L1 正则化函数；SVM 用于两分类，在用 SVM 过程中，核函数选择 Sigmoid 函数时效果不如 RBF 核函数，所以后面也没有用神经网络。
4. 模型组合：Bagging 和 AdaBoosting 的效果比单个回归或分类器效果更好。
5. 误差测量：学习成绩预测关心的是学生的大致排名，所以预测排名与实际排名仅仅差几个名次是允许的，相差很大则不被接受。因此此次排名问题最适合用 Spearman 相关性系数衡量实际排名和预测排名的相关性。

样本选择的创新点：

共提供了某个学院 538 个学生样本，样本不多，尤其很大模型需要大样本，分为训练集和测试集后可用样本更少。基于此，把两个学生的特征放到一起，如果前一个学生排名数字大于后一个学生标签赋值为 1，否则赋值为 0，每个学生 55 个特征，两个学生合在一起才有了 110 个特征。每个学生都可以与 537 个学生放一起组成一个样本，这样样本数就由 538 变成了 538 * 537 = 288906，从中取适当的样本训练即可。

步骤：

一、首先，对数据表进行改名，排序，修改规则如下：

- 训练集的“借书.txt” >> Book Train
- 预测集的“借书.txt” >> Book Predict
- 训练集的“图书馆门禁.txt” >> Library Train
- 预测集的“图书馆门禁.txt” >> Library Predict
- 训练集的“消费.txt” >> Card Train
- 预测集的“消费.txt” >> Card Predict
- 训练集的“成绩.txt” >> Score Train
- 预测集的“成绩.txt” >> Score Predict

二、几个主要的模型：

- Basic Rule 文件夹：
  - Basic Rule.py：第一学期成绩*0.35+第二学期成绩*0.65 作为第三学期成绩的预测值
  - data 文件夹：存放用于提取特征的数据表
  - result 文件夹：存放结果
  - 35 and 65 of rank.txt：保存的结果

- Lasso Rank 文件夹：
  - Extract All Train Feature From Train For Lasso.py：提取 Lasso Rank 模型的训练集特征，110 个特征
  - Extract All Predict Feature For Lasso.py：提取 Lasso Rank 模型的预测集特征，110 个特征
  - Lasso Rank.py：Lasso Rank 模型训练以及预测。首先对数据进行归一化处理，然后利用 Lasso 进行训练，然后进行预测。
  - data 文件夹：存放用于提取特征的数据表
  - result 文件夹：存放结果
  - 中间文件文件夹：存放一些中间变量，用于观察变量结构和具体信息。
  - Lasso Rank(0018).txt：保存的结果

- SVM Rank 文件夹：
  - Extract All Train Feature From Train For SVM.py：提取 SVM Rank 模型的训练集特征，110 个特征
  - Extract All Predict Feature For SVM.py：提取 SVM Rank 模型的预测集特征，110 个特征
  - SVM Rank.py：SVM Rank 模型训练以及预测。首先对数据进行归一化处理，然后利用 rbf 核函数的 SVM 进行训练，然后进行预测。
  - data 文件夹：存放用于提取特征的数据表
  - result 文件夹：存放结果

- SVM Rank(rbf).txt：保存的结果

- Bagging1.py： Lasso Rank 与 Basic Rule 的 Bagging，前者权重为 0.9，后者权重为 0.1
  - bagging_(Lasso Rank)0.9_Basic Rule0.1.txt：模型融合的结果

- Bagging2.py： Lasso Rank 与 Basic Rule 的 Bagging，前者权重为 0.5，后者权重为 0.5
  - bagging_(Lasso Rank)0.5_Basic Rule0.5.txt：模型融合的结果

- Bagging3.py： Lasso Rank 与 SVM Rank 的 Bagging，前者权重为 0.5，后者权重为 0.5
  - bagging_(Lasso Rank)0.5_SVM Rank0.5.txt：模型融合的结果

- AdaBoosting.py：10 个模型进行 AdaBoosting，这 10 个模型分别为 Lr, Lr, Lr, Svm, Gbrt, Gbrt, Gbrt, Gbrt, Lr, Lr，具体参数见代码
  - AdaBoosting.txt：模型融合的结果

- Bagging_Final.py： AdaBoosting 结果与 Bagging 结果的 Bagging，前者权重为 0.8，后者权重为 0.2
  - Final Result.txt：模型融合的结果

三、最终的成绩

- 最终的成绩.txt 文件