大作业 1：Cluster 量化选股策略

作业背景

请仔细阅读研究报告《国海证券：新量化分类选股:Cluster 量化选股策略》，然后参照研报中给的 Cluster 动量和反转策略，进行回测。

（1）Cluster趋势：在采样期通过对收益率序列进行Cluster，实现对股票的分类，依据动量因子选出在采样期平均表现最好的类，在下一个时期持有，持有期到期后，再重新采样持有，如此滚动选股。

（2）Cluster反射：在采样期通过对收益率序列进行Cluster，实现对股票的分类，依据反转因子选出在采样期平均表现最差的类，在下一个时期持有，持有期到期后，再重新采样持有，如此滚动选股。

策略的具体说明

什么是采样期和持有期？

采样期就是我们筛选股票的一个历史区间，而持有期就是我们持有股票的时间段。比如，我在 3 月底根据 1、2、3 三个月的所有股票的历史数据计算出了一些指标，然后根据计算的结果选出了部分股票，这个时候采样期就是 3 个月（1 月 2 月和 3 月），接着从 4 月初开始就持有这些选出的股票，一直持有到 5 月底，对应的持有期就是 2 个月（4 月和 5 月）。等到 5 月末的时候，我再根据过去 3 个月（3 月 4 月和 5 月）的历史数据重新计算指标，筛选股票，选出后从 6 月初开始持有，一直到 7 月底。依次循环……

上面这个例子就是一个采样期为 3 个月，持有期为 2 个月的策略。

采样期内我们要要做什么？

在得到采样期内所有股票的收益率后，首先要根据股票的收益率进行聚类。什么是聚类呢？大概意思就是根据所有股票在采样期的收益率序列，把收益率序列长得像的归为一类。而怎么定义长得像呢？就是根据两个收益率序列之间的距离来定义，两个收益率序列的距离越近，他们就越有可能是一个类的。根据收益率序列之间的距离最终可以把所有股票分成若干个类，每一类里面的股票

都是收益率序列比较接近的。这就是聚类，分类的个数就叫聚类个数。研报中聚类用到的算法是 Kmeans 方法，具体的算法内容请查看研报。（作业中可以使用 sklearn 这个包里面的 KMeans 函数：http://scikit-learn.org/stable/modules/clustering.html#k-means）。

在得到每只股票的类别之后，统计每一类下所有股票的平均累计收益率，选取平均累计收益率最高的类里的股票（动量）或者平均累计收益率最低的类里的股票（反转），得到动量或者反转的股票代码后，采样期要做的事情就结束了。

接下来就是持有期，持有期就比较简单了，就是根据采样期选出的动量/反转股票代码，计算这些股票在持有期每周的平均收益率。在得到所有持有期的收益率之后，就可以计算资金曲线，从而计算出年化收益、最大回撤和超额年化收益等指标。

回测样本选取全部 A 股，数据选取的是从 06 年初到最近的周数据，剔除采样期内有周数据缺失的股票。参数为聚类的个数、采样期和持有期。

默认聚类个数为 30 个。从 5 周到 40 周，步长为 5，分别遍历采样期和持有期，统计每组参数下的策略的年化收益、最大回撤和年化超额收益(基准为上证指数)。最后输出两个 csv 文件，分别是 output_动量.csv 和 output_反转.csv，内容如下图所示：

	A	B	C	D
1	params	return	max_drawc	excess_retu
2	5_5	0.388287	-0.71445	0.304812
3	5_10	0.3316	-0.70631	0.248125
4	5_15	0.250715	-0.72005	0.167241
5	5_20	0.235716	-0.70197	0.152241
6	5_25	0.294916	-0.8336	0.224739
7	5_30	0.202631	-0.71705	0.132454
8	5_35	0.297114	-0.64412	0.21364
9	5_40	0.272909	-0.68754	0.209024
10	10_5	0.360663	-0.74127	0.275396
11	10_10	0.192652	-0.7809	0.107385
12	10_15	0.287979	-0.76078	0.202713
13	10_20	0.242134	-0.75114	0.156867
14	10_25	0.251115	-0.72984	0.179273
15	10_30	0.176367	-0.80048	0.091101
16	10_35	0.297238	-0.70799	0.233808
17	10_40	0.231794	-0.73429	0.146527
18	15_5	0.419162	-0.70443	0.344065
19	15_10	0.316752	-0.72985	0.241655
20	15_15	0.279274	-0.76151	0.204177
21	15_20	0.249815	-0.72867	0.181765
22	15_25	0.287754	-0.6655	0.212657

作业中可能可以使用到的操作：

1. 数据的导入、导出

2. append

3. grouby

4. resample

5. KMeans

6. concat

7. pivot

8. sort_values

9. expanding

10. prod