**基金策略分类、基金选股择时水平、基金优选**

**一、基金策略分类**

本文档遵循特征聚类的方法对基金进行策略分类。

策略分类的潜在含义即策略相似，且标的相似；仅有策略相似而标的不相似，或者仅标的相似而策略不相似，走出来的收益都是不相似的。

那么，我们需要充分利用基金的如下特征：**基金收益率序列；基金公司；基金经理；基金类型：股票型，偏股混合型（行业主题型，大类资产配置型，行业配置型，大类资产+行业配置型，混合均衡，选股重仓型），偏债混合；持仓前三大行业。**

方法如下：

1. 底层算法阐述：

sklearn.cluster下有若干个聚类方法，包括

[cluster.AffinityPropagation](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html#sklearn.cluster.AffinityPropagation)亲和力聚类

[cluster.AgglomerativeClustering](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering)层次聚类

[cluster.DBSCAN](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN)密度簇聚类

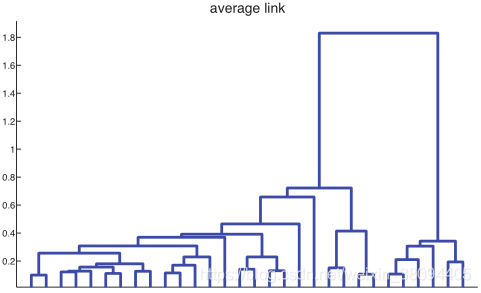
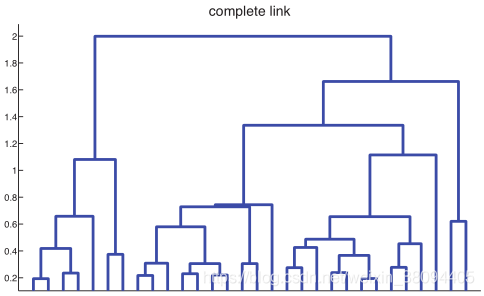
[cluster.KMeans](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans)中心聚类

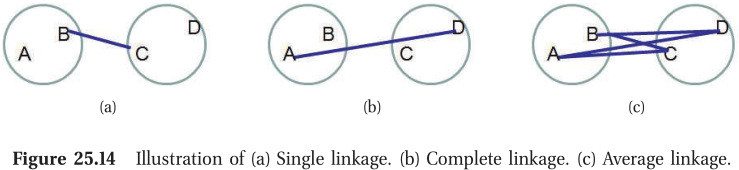
[cluster.SpectralClustering](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html#sklearn.cluster.SpectralClustering)特征向量（PCA、SVD）

算法细节不做解释，以下仅进行优缺点梳理：

**Kmeans**适合知道具体有多少分类，算法速度快，缺点是边缘点很可能分错类，且具体有多少类需要事先确定；**SpectralClustering**适合驱动力分解，例如大类资产驱动力解释；**DBSCAN**适合有明显密度区分的聚类，密度密集的区域被划分为一类；**AffinityPropagation**适合有明显质心的分类，因为每个类都需要确定一个质心，并且可以适用不确定有多少类的分类（我一开始使用这种分类方式，但是发现错误较多，原因是明显的）；

**AgglomerativeClustering**层次聚类显然十分适合基金分类模型：一，基金分类是多层次的，从子类开始，如果两个子类比较相似那就合并，不相似那就分开；二，避免其他算法的劣势，其他启发式算法虽然能够达到一个比较好的状态，但并非一定是最佳状态，该分类在给定参数下只有一个结果，因为算法原理是从底层一层一层堆叠上来的；三，调节参数具有明显意义，参数即为距离阈值，参数增加即降低内聚度，增加耦合度，反之亦然；四，基金分类本身就应按策略大类-一级策略类-二级策略类（以此类推）来分。





其次是single, complete, average的选择，经检验，三种方法差别不大，但就细节而言，single可以选出区分度最明显的分类（类似于SVM），但是其对异常值十分敏感；complete严格限制了每个类的数量和类内距离，可能会导致较多分类；average相对比较平稳，鲁棒性较为良好（有验证函数）

AgglomerativeClustering的入参是**给定的相似度矩阵**，**average的度量方式**，以及**距离阈值**。其中，相似度有众多度量方法，采用如下三种的混合体，即余弦相似度矩阵、相关矩阵、欧氏距离。后文阐述。

1. 相似矩阵的构建

相似矩阵可以用余弦相似度矩阵、相关矩阵、欧氏距离矩阵三种构建方法。

其中，余弦相似度和相关矩阵十分适合衡量日收益率的相关性（注意是日收益率），缺点是有相当一部分仅有仓位差别，而持仓风格相似的基金无法区分出来；于是还需要加上基金之间累计收益的欧氏距离作为两个特征共同判断（注意是累计收益）。但是这又有非常不好的一点就是：初始比较小的累计收益差距会导致后来较大的累计收益差距，同时，一段时间内基金之间的策略可能相似，而另一段时间内基金的策略可能差距较大；所以模型采用滚动聚类的方法。

经验证，比较适合收益率序列相似度度量的是：日收益率的相关矩阵和累计收益率的欧氏距离。原因可以解释为：日收益率不适合用欧氏距离，某一天的异常收益率会迅速拉远两只基金的距离，且欧氏距离是随维数的增加边际递减的；而相关系数也显然不适合累计收益。

如上所述，相似度矩阵拟采用日收益率的余弦相似度和累计收益的欧氏距离（余弦相似度和相关矩阵二选一，相关系数优点：更纯净的相关关系；余弦相似度优点：考虑到均值；但总体差别不大），接下来的方法是将两个特征进行融合：

一个特征并非是一个数，Word2Vec中，每个Word都可以通过该方法将Word通过one-hot向量转化成维度更低的向量。全市场的基金如果看做one-hot向量，那么全市场的基金组成一个矩阵，就是单位阵，单位阵可以理解为每只基金都不相关的相似矩阵，那么相似矩阵CS的列向量本身就可以看做每个基金的特征，CS[i][j]表示基金j在i上的投影，或者相似性；那么将余弦相似度矩阵CS和欧氏距离矩阵ED拼接起来，即pd.concat([CS, ED])，就是同时利用上余弦相似度和欧氏距离的特征矩阵。

3、另外，模型还做了一些数据处理：

**基金收益率数据的处理：**

（1）本质上，基金建仓时间最多也就前50个点。但是计算相似矩阵的时候，经检验，直到扩大到扣除前150个交易日后，相似矩阵才普遍基本稳定，所以，我们计算相似矩阵之前，通常扣除前150个交易日。

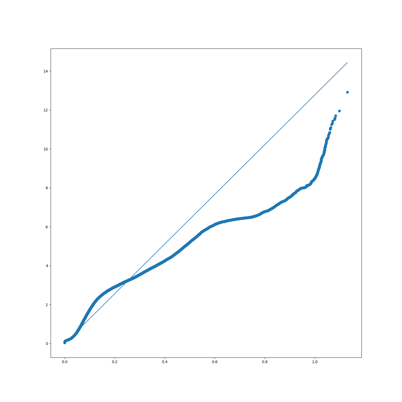
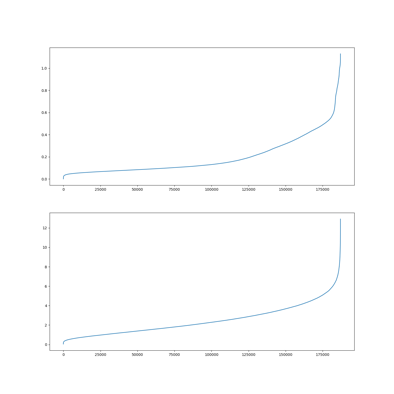
（2）在计算累计收益的欧氏距离的时候，应该要将两个累计收益率的起点拉到同一个交易日进行计算。这样做的问题是，如果有基金按照周公布收益率，公布的数值是一周的收益率。好在公募基金整体数据质量不错，经检验，这样的基金只有5只以内。

（3）在计算相似矩阵的时候，sklearn.metrics在计算之前通常都会判断时间序列是否有nan，如果有nan就报错。所以我采用如下方法：把所有缺失值在10%以内的基金挑出来，dropna过后计算相似矩阵；把缺失值在10-20%的基金挑出来，dropna过后计算相似矩阵堆叠在第一个矩阵之上；以此类推，缺失值多于80%就不计算了。这样是个近似值，但可以不用修改源码（修改源码也尝试过，但是防止在后续模型中有不一样的需求，还是不修改源码比较好）

（4）该模型实际上也可以快速检验出系统中收益率出现错误的基金（异常值），实际上这样的基金不少，约有几十只。

**相似矩阵的构造：**

（1）经验证，余弦相似度和欧氏距离的分布并不是线性均匀分布的，分布如下（左图横轴为CS顺序排列，纵轴为ED顺序排列，可以看出二者分布近似相同，右图是二者概率梯度，横轴是顺序排列的点，纵轴是数值）：

可以看出模型左侧值分布过于密集，不利于分类，那么采用平方根处理的方法能够减缓分布不均匀的问题。直观上亦可以理解，本质上余弦相似度和欧氏距离就是一个二维距离计算方法，使用的时候应该要开平方。

（2）CS和ED还需要进行标准化处理，将二者拉到同等重要的地位。

以上数据处理方法在模型后续还会持续用到。

4、聚类模型构建、其他特征融入及时间序列问题

（1）聚类模型构建

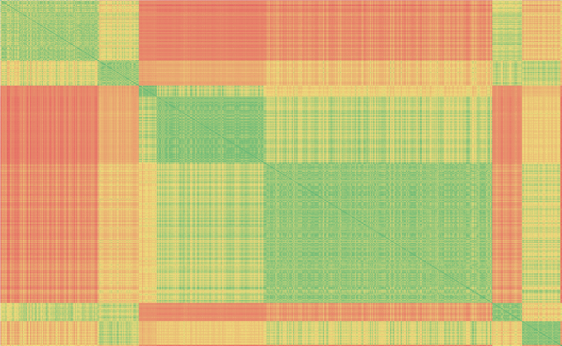
在得到CS\_ED（即pd.concat([CS, ED])）之后，就进行欧氏距离矩阵的计算，这个矩阵就定义为收益率序列矩阵SS。

可以用该矩阵SS进行初步的层次聚类，可以发现聚类效果非常好。

以2010-11-30 ~ 2012-11-30之间的数据为例，聚类图在文件夹20111130文件夹下。进一步地，对每个基金分类仍然可以聚合或者拆分，直至满足要求即可。

经过检验，在分类为10-20类的时候，能够比较完美地融合高内聚低耦合的需求，那么，对于参数的选择，我们就选择大于10个分类的最小数作为分类方法。

相似矩阵按照基金分类排序的条件筛选图如下，红色代表相似性低，绿色代表相似性高：



（2）其他特征融入

按照与上图相似的方法，继续构造基金经理特征向量，基金公司特征向量，行业特征向量，基金类型向量，作为模型的进一步修正。模型结果已在本地数据保存。

注意基金经理特征向量的构建，仍然首先计算基金的相似矩阵SS，然后按照基金经理顺序排列基金，如果基金i1到基金i2代表基金经理i所管理的基金，如果基金j1到基金j2代表基金经理j所管理的基金，那么mean(SS.loc[i1 :i2, i1 :i2])代表基金经理i在自己分项上的特征数值，mean(SS.loc[i1 :i2, j1 :j2])代表基金经理j在基金经理i分项上的特征数值，以此类推。

其他的特征以此类推。

那么第一步的结果就此完成。最终分类结果如下图：

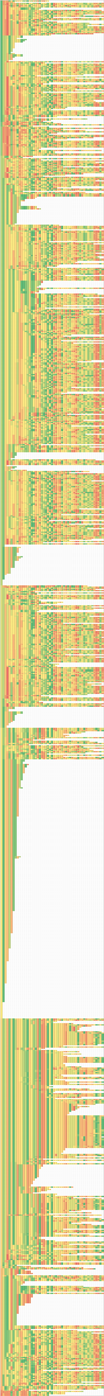


（3）时间序列问题

分类结果如上所述，第二个问题是的时间序列问题，我们希望分类尽量具有延续性。

通过时间序列的计算，2010-11-30 ~ 2015-11-30的分类图如下图，其中columns代表时间倒序，index代表基金代码。按照最新一期的分类排序。通过下图可以看出，分类比较完美，具有良好的鲁棒性。数据如12345.xlsx所示。

当然也会掺杂一些不同分类的数据，这可以正常理解，毕竟基金经理在某个时段可能执行策略A，另一段时间执行策略B，这个问题在后续仍会深入探究。



**二、基金选股择时水平**

5、基准的构建、选股择时能力判断

（1）基准的确定

那么我们可以将类别作为基准计算的基础，基准就定义为类别基金的均值，那么，**每个类别按照收益率从高到低排序即可确定其是属于第几类**。进一步地，我们在判断基金的选股择时能力的时候，可以清晰确认何种基金具有选股能力，何种基金具有择时能力。

（2）选股择时能力判断

选股能力的判断：选股能力就是基金收益相对基准收益的超额收益长期稳定为正，即超额日收益率的t值满足一定的要求（大于0.65），或者累计收益率曲线长期稳定向上，即拟合曲线斜率明显为正，这是判断基金是否具有选股能力的**唯一标准**。

择时能力的判断：

首先需要定义几个情景（不够细致，后续仍需细化）：

2011年全年，全市场普跌；

2012年白酒风格受塑化剂影响；

2013-2015，白酒风格，中小盘风格；

2016-2017，TMT风格，大盘风格；

2019.1-2019.12，医药风格明显；

2020.3-2020.7，以及2020.10-2021.2，消费风格明显；

2021.3-2020.8，新能源，芯片，光伏等科技风格；

2021-10-2022.1，传统能源风格明显。

择时能力判断，主要看两个条件，二者满足其一即可。

条件1：在上述情景中，超额收益在短期是否有显著抬升；

条件2：在上述情景中，基金是否出现换组，换到收益率更高的分类中（前文有阐述第几类是按照收益率排序所确定）

这一部分完全可以量化，但是量化工作尚未进行。

**三、优选基金**

6、优选基金

通过超额收益超过20%的所有基金进行挨个检查，我发现一个非常显著的问题：灵活配置的公募基金经理大多具有一定的择时能力，该择时能力并不是反映在市场下跌或风格下跌之后基金的小幅回撤或零回撤，相反，很多公募基金经理在风格回撤或盘整的时候，会有较多的回撤，但是公募基金经理的择时能力体现在另外一点，就是能够踩准大部分结构风格的情景，如果该灵活配置型基金经理猜准了上述的绝大多数行情，也可以视作是择时能力的一种体现。

还有一部分基金以整个公司的姿态全面进入灵活配置的相关业务。

最后，以下是我挨个检查后的优选基金、次优基金以及备选基金三级基金池，括号代表勉强能达到标准的基金。

优选股票型、混合型基金：

519002

004685

121010

001532

163412

501022

001345（大类资产配置）

121010

002593

006392

166019

005775

（001656）

（000327）

（090018）

优选债基：

001499

002351

001447

000556

001249

002057

002222

519755

519738（519738比519755更好）

003028

（630009）

（001412）

（121012）

次优股票、混合基金：

100026

002943

110013

519767

260101

002537（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

000584（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

001018（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

006567（风格偏好明显，制造业工业）

001716（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

519133

519712

163807

000431

660015

002132

481017

005535

001650（形态不错，收益过低）

519672（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

163823

002620

160215

002291

003751

001120

002770

（002717）

备选股票、混合基金：

673060

002770（有一定风格偏好）

519700（有一定风格偏好）

001487（风格转换期间回撤不小，但是几乎每一波红利都能吃得到）

001907（风格转换期间回撤不小，但是能吃到行业趋势的红利）

002340

410007

001985

001694

004604

519909

320015（大盘价值型，有规模执念）

（673020）

（002910）

（180031）

（001224）