# 指标关联性分析

## 目标：

开发出一套通用的工具，可以针对一组样本集合和一组特征集合，提取出在该样本集合下相关联的特征组/对，并且输出每组/对关联性的强弱。

## 特征之间的关联性：

### 关联性定义：

样本集合下的特征关联性，即在样本集合中，两组或多组特征的值在以下几个方面相似，且均存在正相关、负相关，以及对应关联性强弱的结果：

1. **排序名次上**：基于不同的特征的值进行样本排序，样本在排序结果的名次上一致性高的特征组/对。容易解读。
2. **相关性上**：线性相关、指数相关等。容易解读。
3. **时间序列上**：在同一个样本的不同特征之间，给定时间窗口下（这个时间窗口不一定要一样，甚至不一定要一样长），不同特征的取值上是否存在相似性。考虑周期性。
   1. 在等长窗口上，可以直接用相似度来量化，方法有很多。容易解读。
   2. 不等长窗口上用动态规划来做。不容易解读，到时候再看怎么描述。
4. **分布上**：将样本集合在一项特征上的取值看做一个变量的分布，看不同变量分布之间的散度（评估方法有很多）。散度用于量化两个随机变量之间的[独立性](http://www.cnblogs.com/zhangchaoyang/articles/2642032.html)和[相关性](http://www.cnblogs.com/zhangchaoyang/articles/2631907.html)。不容易解读，到时候再看怎么描述好。
5. **出现频次上**：将不同特征的值进行分箱操作后，得到样本集合在特征上的分类，分析不同特征组的分类在样本集合上同时出现的频次，同时出现次数多的即相似。（典型的关联规则挖掘，算法复杂度高）。容易解读。

### 交付内容：

1. 一个工具（程序）：
   1. 特征间关联性挖掘：输入为一组样本在一组特征上的数据，输出为若干组关联特征组和每组相应的关联性强弱量化（值域在-1~1之间）。
   2. 时间序列关联性挖掘：输入为一个样本在一组连续版本上的一组特征数据，输出为若干组关联特征和每组关联特征对应的时间窗口，以及关联性强弱量化（值域在-1~1之间）。
   3. 每种关联性定义下的输出结果都要保留。

### 注意事项：

1. 以上定义1、2、3.1项作为6月交付的目标；
2. 以上定义3.2、4、5作为7月的交付目标；
3. 关联分析中，极可能存在找到大量的关联性强的特征组，但是其中的绝大多数对于应用场景来讲是没有价值的，所以需要过滤步骤。过滤步骤的解决思路后面补。
4. 针对不同类型关联性的解读文字，能够直接用到产品中，后面补。