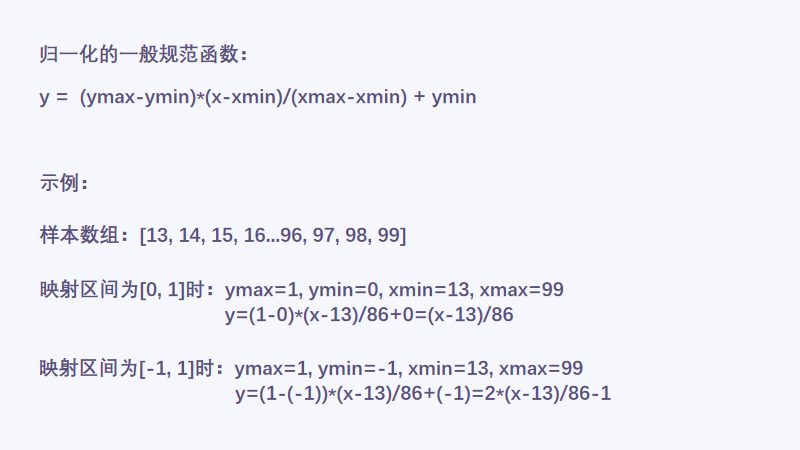
**归一化法**

**一、方法说明**

1. **概念含义**

归一化法，通常指把数据**压缩映射**到指定的区间内进行处理，常见的**映射区间**有[0,1] 和 [-1,1]。

2. **数学原理**



**归一化**本质上是一种**线性变换**，通过函数将一组数据进行**等比例缩放**处理，**映射**到指定的**区间**，由于数据被归一化到了**同一数量级**，使得数据之间的**可比性**问题得到解决。

|  |
| --- |
| **提示**：ymax 为指定区间的上限，ymin 为指定区间的下限。x 为需要被归一化的某一数据，xmax 为该数据所在数组的最大值，xmin 为该数据所在数组的最小值。 |

由于线性变换的特性，归一化处理不会改变数据之间的相对关系，例如大小关系、排序关系等，数据经过归一化处理后，大的仍然为大，小的仍然为小，排在前面的数据仍然排在前面，排在后面的数据仍然排在后面。

|  |
| --- |
| **优势**：由于无论差距多大的数据，均可以通过归一化将它们统一落在一个指定的小区间内进行比较，因此提高了数据之间的可比性，适合进行综合对比评价。 |

3. **主要作用**



**1）消除量纲**

由于归一化可以把数据都统一到一个相同的数值区间内，因此可以去除不同属性特征之间量纲的影响，以便不同属性间的特征能够进行比较和加权处理。

|  |
| --- |
| **提示**：例如我们把一个标的的**成交量**与**换手率**作为选股因子时，成交量代表的是股票成交的股数，单位通常为**股**，而换手率代表股票转手买卖的频率，单位为**%**，这两者是不同属性间的特征，无法直接进行比较。而无论是50万股，还是5%，在归一化后都变成了指定区间内的小数，从而消除了量纲的影响。 |

**2）避免数值问题**

不同的数据数量级相差过大时，计算大数的变化会掩盖掉小数的变化。

|  |
| --- |
| **提示**：例如**市值**和**换手率**，市值的数量级为**亿级**，而换手率的数量级为**小于 1 的数字**如果直接拿这两者的数据作为参考依据，数量级大的属性可能会被作为主要参考因素，而数量级小的属性可能会被忽略不计。而不管是大数还是小数在归一化后都变成了指定区间内的小数，从而避免了数值差距带来的问题。 |

4. **操作步骤**



**1）步骤一：确定样本**

确定需要被归一化的数组样本，其中包括样本的最小值，最大值。

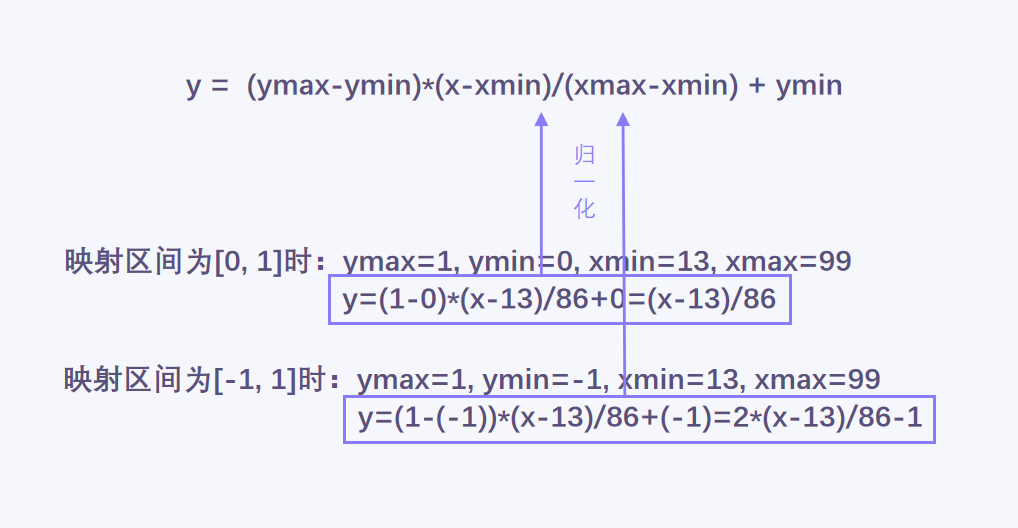


**2）步骤二：确定区间**

根据归一化的目的，确定映射的区间，通常是 [-1, 1] 和 [0, 1] 这两个区间。

**3）步骤三：归一化数据**

按照公式，对数据进行归一化，将数据变成 [-1, 1] 或 [0, 1] 之间的数。



* 例如：
* 当映射区间为 [0, 1] 时，测量数据 20 归一化后的值为：
* y=(1-0)\*(20-13)/86+0=0.081
* 当映射区间为 [-1, 1] 时，测量数据 20 归一化后的值为：
* y=(1-(-1))\*(20-13)/86+(-1)=-0.837

**4）步骤四：比较数据**

对归一化后的数据进行比较。

* 这里我们增加另一组样本数据用以进行比较：[0.1, 0.2, 0.3...0.7, 0.8, 0.9]
* 当映射区间为 **[0, 1]** 时，测量数据 **0.4** 归一化后的值为：
* y=(1-0)\*(0.4-0.1)/0.8+0=**0.375**
* 上面样本中测量数据 **20** 归一化到区间 **[0, 1]** 后的值为 **0.081**。
* **0.375>0.081** ，所以在**同一水平线上**比较时，样本二中的 **0.4** 是要**大于**样本一中的 **20** 的。
* 在实际的场景中比较数据时，我们会赋予其更加具体的含义，比如判断成交量是放量还是缩量，或者分析因子之间的相关性。

**二、方法应用**

5. **交易经验**

**交易背景**

股市中有大量关于**量价关系**的谚语，例如**放量上涨，缩量下跌**，即成交量上涨的时候，价格也会随之上涨；成交量下跌的时候，价格也会随之下跌。

虽然这个说法过于绝对，对于个股来说，甚至有可能会发生相反的情况。但成交量仍然是一个重要的指标。知道每个价格上的成交量，可以画出筹码分布，判断持仓成本的高低辅助决策，它也是我们进行选股时的一个重要参考因素。

因此有说法认为在其他参考因素相差不大的情况下，当日成交量表现较好的股票比成交量表现较差的股票，在未来几日会拥有更好的收益率。

**交易想法**

|  |
| --- |
| **当标的池里有两个候选股且待选择其中一只的时候，买入当日成交量表现较好的那只。** |

6. **问题描述**

|  |
| --- |
| **如何定义成交量的表现？表现较好的标准是什么？** |

7. **量化思路**

**难点**

对于两只不同的股票，我们**难以将它们的成交量数据直接进行比较**，作为表现较好或较差的标准。

**原因**

成交量是放量还是缩量，表现是好还是差，**需要结合股票自身属性和近期行情**，**不能将不同股票的当日成交量直接进行比较**。

* 例如：
* a股票成交量通常维持在1000万股以上，而b股票成交量通常维持在100万股以上。假设股票的表现和成交量是成正相关的，那么同是当日成交量500万股，对于a股来说就是成交量表现不好，但对于b股来说就变成了表现极好。**由于两只股票自身的属性存在差别，无法直接进行比较。**
* 对于同一只股来说，2000万股的成交量在平常行情里是放量，但是放在牛市里却是缩量。**由于成交量的表现还需结合行情，因此同一个成交量值在不同的行情里意义完全不一样**。

**思路**

|  |
| --- |
| 由于股票属性和近期行情不同，不同股票的当日成交量无法直接进行比较，那么我们能否先把两只股票的当日成交量表现放在同一水平线上，然后再进行比较呢？ |

|  |
| --- |
| **从比较两只股票的当日成交量数据，转变为把两只股票的当日成交量表现放到同一个水平线上做对比。** |

**归一化**能够把**不同数量级**的数据统一到一个固定的小数值区间内，具有**消除量纲**和**避免数值问题**的作用。而成交量的表现同时需要结合近期行情，因此我们可以把该股票过去一个月的成交量数据作为数组样本；又因成交量中的**放量**和**缩量**属于极性相反的两种情况，因此可选择将当日成交量的数据归一化到[-1, 1] 区间内。

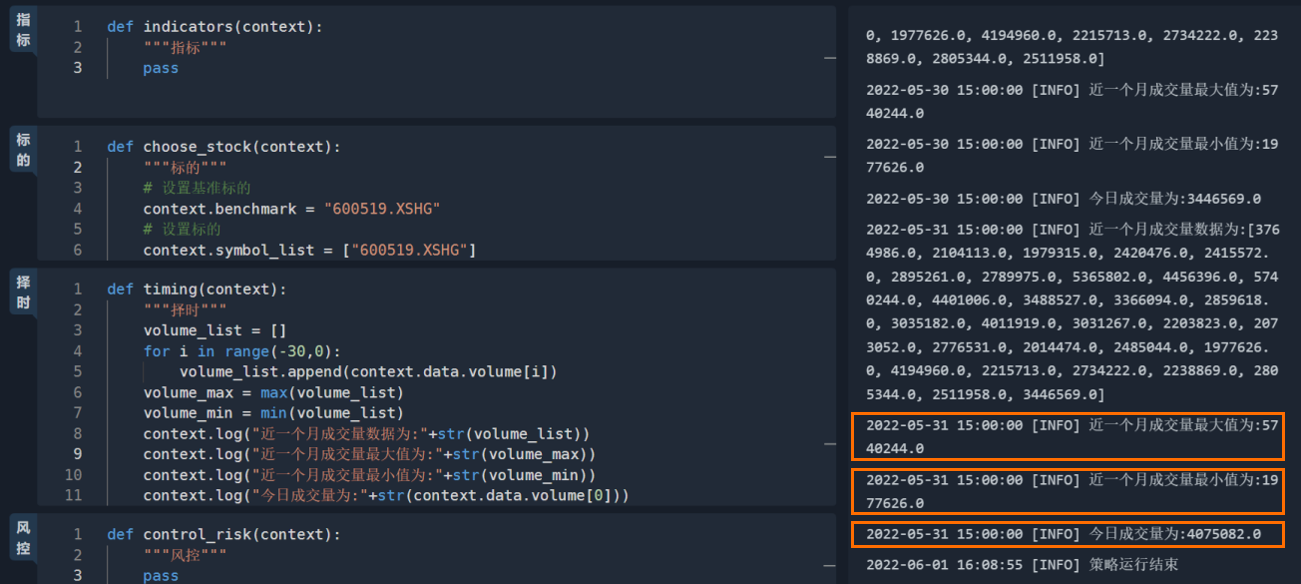
这样一来，在对两只股票的当日成交量完成归一化、能够直接对成交量表现进行比较的同时，也能通过分界线 0 来判断当日成交量在近期是属于放量还是缩量。

8. **场景应用**

|  |
| --- |
| **提示：**以下示例以当前交易日 2022-05-31 为例，选取的两只标的分别为贵州茅台(600519.XSHG)和广汇汽车(600297.XSHG)。 |

第一步：确定样本。确定需要被归一化的数组样本，这里取**贵州茅台 (600519.XSHG)** 过去一个月的成交量数据作为数组样本，即数组最小值为过去一个月的**成交量最小值**，最大值为**成交量最大值**。

这一步可以直接用 HiTrader 中的 volume 属性获取得到。可以看到贵州茅台 (600519.XSHG) 过去一个月的成交量最大值为 **5740244** 股，最小值为 **1977626** 股。



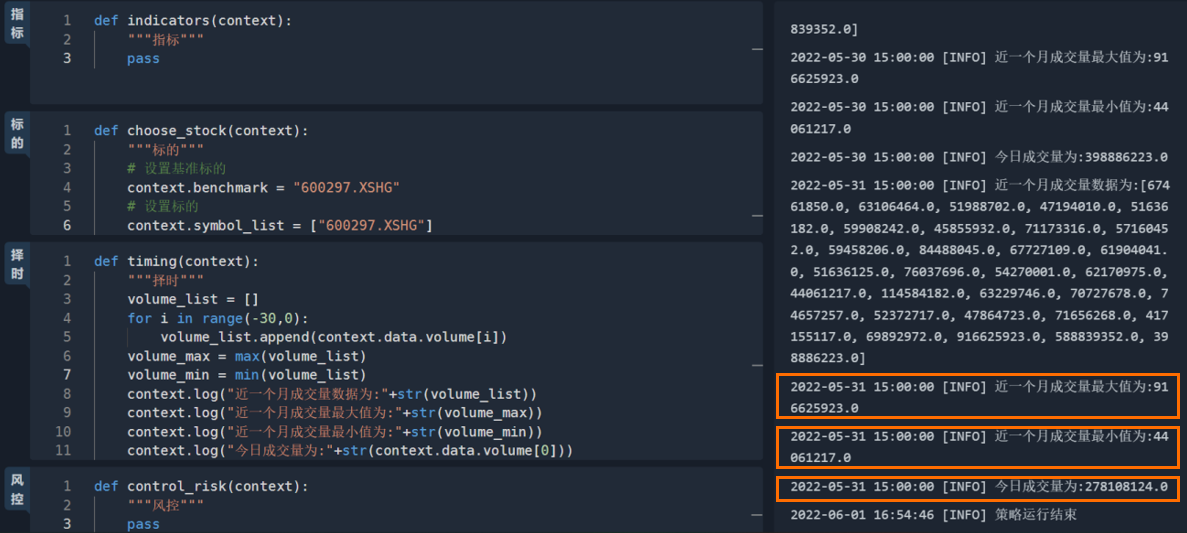
第二步：确定区间。由于我们希望可以借助分界线 0 来判断成交量是**放量**还是**缩量**，所以确定映射区间为 [-1, 1]。

第三步：归一化数据。先将贵州茅台 (600519.XSHG) 当日 (2022-05-31) 成交量数据归一化映射到区间[-1, 1]，得到 -1~1 之间的小数。由上图我们得知当日成交量为 **4075082** 股，将以上数据代入公式即可得到归一化之后的值为 **0.115**。



接着我们对另一只标的**广汇汽车 (600297.XSHG)** 重复以上三个步骤：

* 确定样本，取广汇汽车 (600297.XSHG) 过去一个月的成交量数据作为数组样本，最大值为**916625923** 股，最小值为 **44061217** 股；



* 确定区间，和上面一样，将映射区间设置为 [-1, 1]；
* 归一化数据，由上图我们可知当日成交量为 **278108124** 股，将以上输入代入公式即可得到归一化之后的值为 **-0.464**。



第四步：比较数据。贵州茅台 (600519.XSHG) 当日 (2022-05-31) 成交量为 **4075082** 股，而广汇汽车 (600297.XSHG) 当日 (2022-05-31) 成交量为 **278108124** 股，直接比较数据的话，我们很容易得出 **278108124>4075082** 进而选择买入成交量更大的广汇汽车 (600297.XSHG) 。

但是当把它们归一化到同一数量级进行比较，在消除了数值差距影响之后，可以看到**0.115>-0.464，**且以 0 为分界线**，0.115 代表成交量是放量**而 **-0.464 代表成交量是缩量，**所以最终我们选择买入**当日成交量表现更好的贵州茅台 (600519.XSHG)。**

**三、总结延伸**

9. **总结**

|  |
| --- |
| 方法的核心是——**把不同样本的同一特征、或者同一样本的不同特征，放到同一个水平线上做对比。** |

我们可以抓住这个核心，从而对其进行变式，以灵活适应实际情形。

10. **延伸**

除了上述的方法应用之外。

* 1）可以考虑使用**不同的数组样本**。
* 例如取过去 100 天的成交量数据作为数组样本。
* 2）可以考虑**归一化到不同的区间**，满足不同的数据比较需求。
* 例如归一化到 [0, 1]。
* 3）可以考虑使用该方法对同一只股票的**不同指标数据**进行归一化。
* 例如评定一只股票的好坏时需要参考多个指标，将不同的指标数据归一化到同一个区间，方便对这些参考因素进行加权处理，对股票进行评定。

|  |
| --- |
| **备注**：   * 当数组样本存在**异常值、极端值**的时候，单纯使用归一化无法避免异常值、极端值的影响，需要先对数组样本进行**去除异常值**的处理。 * 例如：数组样本中的数据集中在 20~63，只有一个异常数据为 1000，如果直接按照数据最大值最小值的方式确定样本的上下限，将会极大地影响归一化后的数据结果，因此需要根据实际情况，去除数组样本的异常值极端值。 |