**时间序列Arima模型理论及在Data Studio中的应用**

目 录

[1 Arima模型理论 1](#_Toc418088504)

[1.1平稳序列建模 1](#_Toc418088505)

[MA滑动平均过程与AR自回归过程 2](#_Toc418088506)

[ARMA自回归滑动平均混合模型 2](#_Toc418088507)

[1.2 非平稳时间序列的ARIMA过程 3](#_Toc418088508)

[2 ARIMA在Data Studio中的应用 5](#_Toc418088509)

# 1 Arima模型理论

Arima模型是以加权的方法对白噪声的组合来建立模型的，并以模型和实际数据的残差服从均值较小的正态分布为目标。

## 1.1平稳序列建模

所谓的平稳性是指某一时间序列是由同一个随机过程生成的，即时间序列x(t)(t=1,2,3,…..)的每一个数值都服从同一个随机分布，该随机分布生成的事件序列满足以下条件：

均值与时间t无关，任何k阶滞后序列的均值都相同

方差与时间t无关，任何k阶滞后序列的方差都相同

X(t)与X(t-k)的协方差只与滞后阶数k有关，与时间t无关。

自相关系数只与滞后阶数k有关，与t无关。

从直观上看，所谓的平稳过程可以通过查看acf自相关图来确定，如果acf成指数级衰减，则表示平稳（如下图）。



平稳序列

上图是一个有均值为10，标准差为1的正态分布生成的时间序列。acf自相关图用来展示X(t)时间序列与X(t-1), X(t-2),.., X(t-k)各阶之后时间序列之间的相关系数。如果k阶滞后序列X(t-k)与原始序列X(t)的相关系数不在[-0.2,0.2]之间，则称X(t-k)与X(t)不具有相关性。如果一个时间序列的任何一阶滞后序列与原始序列都不具有相关性，则该时间序列不具有自相关性。



对于一个平稳性的时间序列，如果其均值为0，且不具有自相关性，这样的平稳性时间序列为白噪声序列。

### MA滑动平均过程与AR自回归过程

q阶滑动平均过程（MA(q)）是把若干白噪声e做加权得到的，其公式如下：

$$Y\_{t}=e\_{t}+θ\_{1}e\_{t-1}+…+θ\_{q}e\_{t-q}$$

p阶自回归过程（AR(p)）是使用序列本身做为变量，并对其加权得到，公式如下：

$$Y\_{t}=φ\_{1}Y\_{t-1}+φ\_{1}Y\_{t-1}+…+φ\_{1}Y\_{t-1}+e\_{t}$$

### ARMA自回归滑动平均混合模型

在实际数据分析中很少有指标数据可以单纯用AR或者MA来建立模型，通过把平均混合模型MA和自回归模型AR进行叠加，就组成了自回归滑动平均混合模型ARMA（p，q），其公式如下：

$$Y\_{t}=e\_{t}+θ\_{1}e\_{t-1}+…+θ\_{q}e\_{t-q}+φ\_{1}Y\_{t-1}+φ\_{1}Y\_{t-1}+…+φ\_{1}Y\_{t-1}+e\_{t}$$

下图为模拟一个ARMA（1,1）生成的序列



在AR和MA中可以通过观察acf与pacf的拖尾和截尾来得到AR和MA的阶数，而对于ARMA的acf与pacf图同时表现出拖尾的特征，为了确定参数p、q可以使用混合自相关图



观察得到p、q的两种可能的取值，具体取值通过比较分别建模后的质量确定。

## 1.2 非平稳时间序列的ARIMA过程

对于不能用ARMA等模型建模的非平稳时间序列，要进行平稳化处理，在ARMA（p，q）的基础上引入了差分阶数d的概念。建立ARIRMA（p,d,q）模型。

常用的平稳化处理手段有取对数和差分。对非平稳的时间序列data如果d阶差分diff（data，d）为平稳序列，则可以建立ARIMA（p，d，q）模型，一般情况下d最多可以取到3，避免过差分。如果仅仅差分处理不能平稳化，则可以先取对数，再差分diff（log（data），d），大部分非平稳序列都能经过取对数和差分处理转化为平稳序列。再根据ARMA建模方法确定其他参数。

下图是1986年1月至2015年3月至今每桶原油的月度价格。可以见看出按每年有较强周期性，并且逐年上升，平稳模型对于这样的时间序列并不适用，选择使用ARIMA模型建模是可行的。



通过R语言的decompose函数吧时间序列分解为长期趋势和周期性变化。

从trend可以看出，确实存在一个向上的趋势。



使用shapiro.test检验diff（trend，d）直到diff（trend，d）或者diff（log（trend，d））为正态分布。确定d后使用混合自相关图eacf确定可能的p和q，建立模型。

# 2 ARIMA在Data Studio中的应用

Data Sutdio中使用了ARIMA模型，需要的参数定义如下

<parameter-string key=*"column.date"* description=*"日期列"* optional=*"false"* expert=*"false"* name=*"日期列"* />

 <parameter-string key=*"column.target"* description=*"目标列"* optional=*"false"* expert=*"false"* name=*"目标列"* />

 <parameter-int key=*"column.targetindex"* description=*"目标列序号"* optional=*"false"* expert=*"false"* name=*"目标列序号"* />

 <parameter-int key=*"trainpercent"* description=*"训练数据集百分比"* optional=*"false"* expert=*"false"* name=*"训练数据集百分比"*/>

 <parameter-category key=*"frequency"* description=*"时间周期，比如 年，季度，月"* default=*"1"* name=*"时间周期"*>

 <value>1</value>

 <value>12</value>

 <value>4</value>

 <value>7</value>

 </parameter-category>

 <parameter-int key=*"startyear"* description=*"起始年"* optional=*"false"* expert=*"false"* name=*"起始年"* />

 <parameter-int key=*"startquarter"* description=*"起始季度"* optional=*"false"* expert=*"false"* name=*"起始季度"* />

 <parameter-int key=*"startmonth"* description=*"起始月"* optional=*"false"* expert=*"false"* name=*"起始月"* />

 <parameter-int key=*"startindex"* description=*"起始星期"* optional=*"false"* expert=*"false"* name=*"起始星期"*/>

 <parameter-int key=*"startday"* description=*"起始天"* optional=*"false"* expert=*"false"* name=*"起始天"*/>

 <parameter-boolean key=*"auto"* description=*"使用默认的参数"* default=*"false"* optional=*"true"* expert=*"false"* name=*"使用默认的参数"*/>

 <parameter-int key=*"p"* description=*"自回归阶数"* optional=*"true"* expert=*"false"* name=*"移动平均阶数(p)"*/>

 <parameter-int key=*"d"* description=*"差分阶数"* optional=*"true"* expert=*"false"* name=*"差分阶数(d)"*/>

 <parameter-int key=*"q"* description=*"移动平均阶数"* optional=*"true"* expert=*"false"* name=*"移动平均阶数(q)"*/>

 <parameter-int key=*"P"* description=*"季节自回归阶数"* optional=*"true"* expert=*"false"* name=*"季节自回归阶数P)"*/>

 <parameter-int key=*"D"* description=*"季节差分阶数目"* optional=*"true"* expert=*"false"* name=*"季节差分阶数(D)"*/>

 <parameter-int key=*"Q"* description=*"季节移动平均阶数"* optional=*"true"* expert=*"false"* name=*"季节移动平均阶数((Q)"*/>

* Column.data参数选择数据中的一列，用来指示数据的时间信息
* Column.target指定用于分析的数据列
* Trainpercent 设置数据集中作为训练数据的百分比
* Frequency 周期，供选择的有1,4,7,12；分别代表年，季度，周，月
* 参数 p ，d ，q分别代表ARIMA（p,d,q）de 自回归阶数，差分阶数和滑动平均阶数。
* 参数P，D，Q代表季节性趋势的p，d，q
* XML中的起始年月星期等信息是通过数据集中选择的日期列中最小的日期计算得到的。

在Data Studio中对应设置如下