时间序列预测之--ARIMA模型

什么是 ARIMA模型

ARIMA模型的全称叫做自回归移动平均模型，全称是(ARIMA, Autoregressive Integrated Moving Average Model)。也记作ARIMA(p,d,q)，是统计模型(statistic model)中最常见的一种用来进行时间序列 预测的模型。

1. ARIMA的优缺点

优点： 模型十分简单，只需要内生变量而不需要借助其他外生变量。

缺点:
1.要求时序数据是稳定的（stationary），或者是通过差分化(differencing)后是稳定的。
2.本质上只能捕捉线性关系，而不能捕捉非线性关系。
注意，采用ARIMA模型预测时序数据，必须是稳定的，如果不稳定的数据，是无法捕捉到规律的。比如股票数据用ARIMA无法预测的原因就是股票数据是非稳定的，常常受政策和新闻的影响而波动。

2. 判断是时序数据是稳定的方法。

严谨的定义： 一个时间序列的随机变量是稳定的，当且仅当它的所有统计特征都是独立于时间的（是关于时间的常量）。
判断的方法:

1. 稳定的数据是没有趋势(trend)，没有周期性(seasonality)的; 即它的均值，在时间轴上拥有常量的振幅，并且它的方差，在时间轴上是趋于同一个稳定的值的。
2. 可以使用Dickey-Fuller Test进行假设检验。（另起文章介绍）

3. ARIMA的参数与数学形式

ARIMA模型有三个参数:p,d,q。

- p--代表预测模型中采用的时序数据本身的滞后数(lags) ,也叫做AR/Auto-Regressive项
- d--代表时序数据需要进行几阶差分化，才是稳定的，也叫Integrated项。
- q--代表预测模型中采用的预测误差的滞后数(lags)，也叫做MA/Moving Average项

先解释一下差分： 假设y表示t时刻的Y的差分。

$$if\ d=0,\ y_t = Y_t$$

$$if\ d=1,\ y_t = Y_t - Y_{t-1}$$

$$if\ d=2,\ y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$$
$$= Y_t - 2Y_{t-1} + Y_{t-2}$$

ARIMA的预测模型可以表示为:

Y的预测值 = 常量c and/or 一个或多个最近时间的Y的加权和 and/or 一个或多个最近时间的预测误差。

假设p，q，d已知，
ARIMA用数学形式表示为:

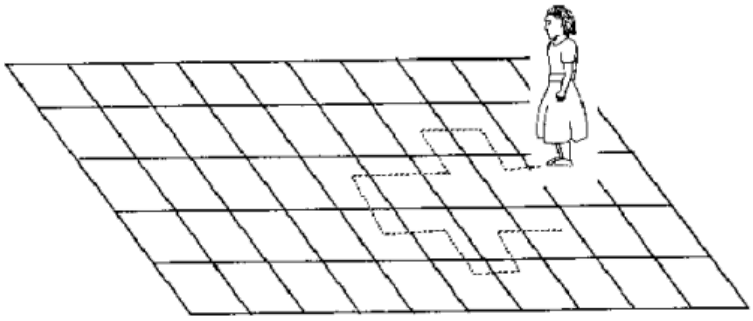
$$\widehat{y_t} = \mu + \phi_1 * y_{t-1} + \ldots + \phi_p * y_{t-p} + \theta_1 * e_{t-1} + \ldots + \theta_q * e_{t-q}$$

其中,ϕ表示AR的系数，θ表示MA的系数

4.ARIMA模型的几个特例

1.ARIMA(0,1,0) = random walk:

当d=1，p和q为0时，叫做random walk，如图所示，每一个时刻的位置，只与上一时刻的位置有关。



预测公式如下：

$$\widehat{Y}_t = \mu + Y_{t-1}$$

2. ARIMA(1,0,0) = first-order autoregressive model:

p=1, d=0,q=0。说明时序数据是稳定的和自相关的。一个时刻的Y值只与上一个时刻的Y值有关。

$$\widehat{Y}_t = \mu + \phi_1 * Y_{t-1}.$$
$$where,\ \phi \in [-1,1], 是一个斜率系数$$

3. ARIMA(1,1,0) = differenced first-order autoregressive model:

p=1,d=1,q=0. 说明时序数据在一阶差分化之后是稳定的和自回归的。即一个时刻的差分（y）只与上一个时刻的差分有关。

$$\hat{y}_t = \mu + \phi_1 * y_{t-1}$$
结合一阶差分的定义，也可以表示为：$\widehat{Y}_t - Y_{t-1} = \mu + \phi_1 * (Y_{t-1} - Y_{t-2})$
或者$\widehat{Y}_t = \mu + Y_{t-1} + \phi_1 * (Y_{t-1} - Y_{t-2})$

4. ARIMA(0,1,1) = simple exponential smoothing with growth.

p=0, d=1 ,q=1.说明数据在一阶差分后市稳定的和移动平均的。即一个时刻的估计值的差分与上一个时刻的预测误差有关。

$$\hat{y}_t = \mu + \alpha_1 * e_{t-1}$$
注意$q=1$的差分y_t与$p=1$的差分y_t的是不一样的
其中，$\hat{y}_t = \widehat{Y}_t - \widehat{Y}_{t-1},\ e_{t-1} = Y_{t-1} - \widehat{Y}_{t-1}$, 设$\theta_1 = 1 - \alpha_1$
则也可以写成：$\widehat{Y}_t = \mu + \widehat{Y}_{t-1} + \alpha_1(Y_{t-1} - \widehat{Y}_{t-1})$
$= \mu + Y_{t-1} - \theta_1 * e_{t-1}$

5. ARIMA(2,1,2)

在通过上面的例子，可以很轻松的写出它的预测模型：

$$\hat{y}_t = \mu + \phi_1 * y_{t-1} + \phi_2 * y_{t-2} - \theta_1 * e_{t-1} - \theta_2 * e_{t-2}$$
也可以写成：$\widehat{Y}_t = \mu + \phi_1 * (Y_{t-1} - Y_{t-2}) + \phi_2 * (Y_{t-2} - Y_{t-3}) - \theta_1 * (Y_{t-1} - \widehat{Y}_{t-1}) - \theta_2$

6. ARIMA(2,2,2)

$$\hat{y}_t = \mu + \phi_1 * y_{t-1} + \phi_2 * y_{t-2} - \theta_1 * e_{t-1} - \theta_2 * e_{t-2}$$
$$\widehat{Y}_t = \mu + \phi_1 * (Y_{t-1} - 2Y_{t-2} + Y_{t-3}) + \phi_2 * (Y_{t-2} - 2Y_{t-3} + Y_{t-4}) - \theta_1 * (Y_{t-1} - \widehat{Y}_{t-1})$$

7. ARIMA建模基本步骤

1. 获取被观测系统时间序列数据；
2. 对数据绘图，观测是否为平稳时间序列；对于非平稳时间序列要先进行d阶差分运算，化为平稳时间序列；
3. 经过第二步处理，已经得到平稳时间序列。要对平稳时间序列分别求得其自相关系数ACF 和偏自相关系数PACF，通过对自相关图和偏自相关图的分析，得到最佳的阶层 p 和阶数 q
4. 由以上得到的d、q、p，得到ARIMA模型。然后开始对得到的模型进行模型检验。
具体例子会在另一篇文章中给出。

分类: 数据挖掘与分析

标签: 时间序列, 数据预测, 统计模型

好文要顶 关注我 收藏该文

geek精神
关注 - 0
粉丝 - 15
+加关注

7 0

« 上一篇: Python之内建函数Map,Filter和Reduce
» 下一篇: python时序数据分析--以示例说明

posted @ 2017-05-08 20:22 geek精神 阅读(60858) 评论(4) 编辑 收藏

评论列表

#1楼 2017-05-08 22:25 心中呈和

赞!
为您的无私奉献点赞!
(看到博客中大量的文章被阅读成百上千次，但竟然无人评论！)

(呼吁大家践行 "开放、平等、协作、快速、分享"的互联网精神！)

支持(0) 反对(0)

#2楼 [楼主] 2017-05-09 10:32 geek精神

@ 心中呈和
谢谢你的鼓励

支持(0) 反对(0)

#3楼 2018-04-25 15:37 笨蛋敏

这是我看过对ARMIA分析最简单明了的文章，很赞，另外想问作者，你现在也在做时间序列预测吗，想问一下，有没有用过神经网络来做过

支持(1) 反对(0)

#4楼 2019-07-10 08:39 老笨啊

博主，你好!
对于时间序列，我一直有几点不明白之处，还请指点:
1. 对于差分，其做法就是用后面的数据减去前面的数据。个人感觉其实就是消除了长期趋势的影响----可以想象成等差序列相互扣减，剩余为0。而非等差序列的数组，扣除相应的值后，剩余的便是消除了长期趋势的影响。----不知道是否理解有误;
2. 对于移动平均法，用的是滑动窗口的方法处理数据，感觉是消除周期性的影响。---这点上没有很通俗的理解，不知道是否正确。
3. ARIMA模型，其实是将时序数据，看成是加法模型。通过差分、移动平均方法，来消除趋势因素、周期因素的影响后，查看剩余的残差部分是否是平稳的，也就是看残差是否有规律。如果残差有规律，可以通过回归模型，拟合出函数。这样就可以对残差值进行预测，而要预测总值的话，再反向加上原先剔除的趋势因素和周期因素即可（对于趋势因素和周期因素，其实做预测是不难的，因为也有规律性）。
4. 对于第3点，我还有另外的理解是，扣除趋势和周期因素后的残差，是随机的（要满足白噪声检验和DW检验）。如果残差的均值为0，方差为1，则其实可以理解成残差对模型的影响很小（残差趋近于0也是可以的，同样说明残差影响较小）。如残差不满足正态分布，也就是说，其均值较大，对模型的影响较大，这样的话，预测值是偏差较大的，因此不适合用ARIMA模型。----我不知道第3点和第4点，哪种正确，或者说两种都不正确。。
求指点~~

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册， 访问 网站首页。

相关博文:
· 时间序列模式——ARIMA模型
· 预测模型
· Arima模型总结
· 基于R语言的ARIMA模型
· 用R做时间序列分析之ARIMA模型预测