时间序列分析之理论篇

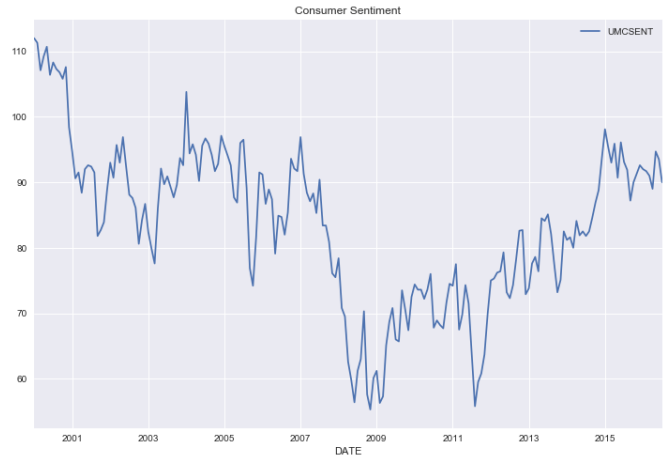
CSDN：http://blog.csdn.net/kicilove/article/

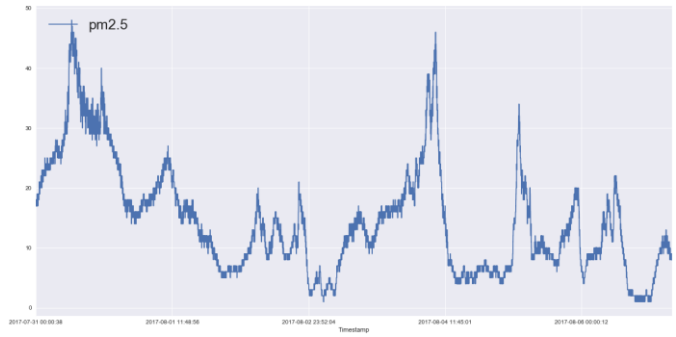
github：https://github.com/zhaohuicici?tab=repositories

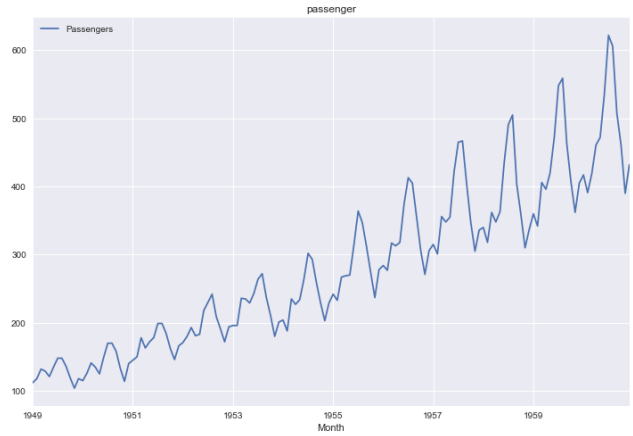
前言：

一说起时间序列大家并不会陌生。每时刻的甲醛浓度变化、每日股票闭盘价格、共享单车每日租车数等等都可以看做一系列时间点上的观测，在一系列时间点上观测获取的数据也就是我们俗称的时间序列数据。本文主要介绍常见的 AR、MA、ARMA、ARIMA 平稳时间序列模型以及时间序列常见的数学特征以及时间序列建模的流程，此篇相对来说偏于理论、偏于公式，下篇会给出一个关于时间序列的 Python 实例。

下面我们通过两幅图来简单看看时间序列的样子。图一是 2000-2016 年美国消费者信心指数【http://data.eastmoney.com/cjsj/xfzxx.html】；图二是某地在某段时间 pm2.5 的浓度变化情况；图三是 1949 年到 1960 年某地某航空公司每月乘客数。







可以看出来时间序列数据的形状真的是千奇百怪，但是不管怎么样，要想预测好数据还得见招拆招，根据不同的数据特点做不同的检验，选择不同的模型，确定不同的参数。

数学特征

在介绍模型之前，先看看时间序列数据有啥部件需要我们知晓，下面介绍时间序列的数学特征。像一般的随机变量一样，时间序列也有随机变量序列，也有相应的均值、自协方差、方差、期望、相关系数等，只不过这里我们要加上函数俩字，也就是均值函数、自协方差函数、方差函数、期望函数、相关系数函数，之所以加上函数二字，是因为时间序列对应的这些数学特征变成了时间的函数。来看看它们具体的数学表达式（公式是一个一个敲上去的，如果有手误请指正）：

一般随机变量的数学特征：

期望：

对于连续型随机变量 X，有概率密度函数 f(x),则定义

$$E(X) = \int_{-\infty}^{\infty} f(x)dx$$

为 X 的数学期望。
对于离散型的随机变量 X，X 的数学期望就是随机变量 X 的取值与发生概率相乘得到的加和，这个是高中知识就不再赘述。

方差：

- 设 X 施一个随机变量，若 $E\{[X - E(X)]^2\}$存在，则称 E{[X-E(X)]}为 X 的方差，记为 D(X) 或 Var(X)，即 D(X) = Var(x) = $E\{[X - E(X)]^2\}$
- $\sqrt{D(X)}$称为 X 的标准差。

- 若 X 是离散型随机变量,则$D(X) = \sum_{k=1}^{\infty}[x_k - E(X)]^2 p_k$
- 若 X 是连续型随机变量，则$D(X) = \int_{-\infty}^{\infty}[x - E(X)]^2 f(x)\,dx$
- $D(X) = E\{[X - E(X)]^2\} = E\{X^2 - 2XE(X) + [E(X)]^2\} = E(X^2) - 2E(X)E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2$

协方差：

- 称$E\{(X - E(X))(y - E(Y))\}$为随机变量 X 与 Y 的协方差，记为$Cov(X, Y)$

$$Cov(X, Y) = E\{(X - E(X))(y - E(Y))\} = E(XY) - E(X)E(Y)$$

相关系数：

称$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}$为随机变量 X 与 Y 的相关系数。

随机变量序列的一些数学特征

类比于随机变量的数学特征，下面是随机变量序列的一些数学特征：

随机变量序列$\{y_t: t = 0, 1, 2, \dots\dots\}$称为一个时间序列模型。

均值函数：

$$u_t = E(y_t), t = 0,1,2, \dots\dots$$

自协方差函数：

$$\gamma_{t,s} = Cov(y_t, y_s) = E[(y_t - u_t)(y_s - u_s)] = E(y_t y_s) - u_t u_s, t, s = 0,1,2, \dots\dots$$

自相关系数：

$$\rho_k = \frac{Cov(y_t,\ y_{t-k})}{\sqrt{Var(y_{t-k})Var(y_t)}}$$

平稳性：

做平稳时间序列模型比知必会的平稳性概念：

1．平稳性就是要求经由样本时间序列所得到的拟合曲线
在未来的一段期间内仍能顺着现有的形态“惯性”地延续下去
2.平稳性要求序列的均值和方差不发生明显变化

严平稳与弱（宽）平稳：

1.严平稳：严平稳表示的分布不随时间的改变而改变。

如：白噪声（正态），无论怎么取，都是期望为 0，方差为 1

2.弱平稳：期望与相关系数（依赖性）不变

未来某时刻的 t 的值$\mathbf{Y_t}$就要依赖于它的过去信息，所以需要依赖性

弱平稳时间序列的数学特征：

1.均值$E(Y_t) = \mu$与时间 t 无关的常数；
2.方差$Var(Y_t) = \gamma$与时间 t 无关的常数；
3.协方差$Cov(Y_t, Y_{t+k}) = \gamma_{0,k}$只与时间间隔 k 有关，与时间 t 无关的常数。
4.自相关系数$\rho_k = \rho_k = \frac{Cov(y_t,\ y_{t-k})}{\sqrt{Var(y_{t-k})Var(y_t)}} = \frac{Cov(y_t,\ y_{t-k})}{Var(y_t)} = \frac{\gamma_k}{\gamma_0}$

差分法：

对于不平稳的时间序列，我们一般会使用差分的方法得到想要的平稳序列，下图是美国消费者信心指数序列，一阶差分和二阶差分后的序列。

其他变换：对数变换，幂变换等

时间序列影响因素（影响因素的叠加）

1.长期趋势 Trend：现象在较长时期内受某种根本性因素作用而形成的总的变动趋势
2. 循环变动\周期性 Cyclic：现象以若干年为周期所呈现出的波浪起伏形态的有规律的变动
3. 季节性变化 Seasonal variation：现象随着季节的变化而发生的有规律的周期性变动
4. 不规则变化 Irregular movement：是一种无规律可循的变动，包括严格的随机变动和不规则的突发性影响很大的变动两种类型

平稳时间序列模型介绍：

自回归模型（AR）：

1.描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测
2.自回归模型必须满足平稳性的要求
3.p 阶自回归过程的公式定义：$y_t = u_t + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$
4.y_t是当前值 u_t是常数项 P 是阶数 γ_i是自相关系数 ϵ_t是误差

自回归模型的限制:
1.自回归模型是用自身的数据来进行预测
2.必须具有平稳性
3.必须具有自相关性，如果自相关系数小于 0.5，则不宜采用
4.自回归只适用于预测与自身前期相关的现象

移动平均模型（MA）：

1.移动平均模型关注的是自回归模型中的误差项的累加
2.q 阶自回归过程的公式定义：$y_t = u + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$
3.移动平均法能有效地消除预测中的随机波动

自回归移动平均模型（ARMA）：

1.自回归 AR 与移动平均 MA 的结合

2.公式定义：$\mathbf{y_t = u + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}\ \epsilon_t}$

ARIMA(p，d，q)模型：

1.全称为差分自回归移动平均模型(Autoregressive Integrated Moving Average Model,简记 ARIMA)
2.AR 是自回归， p 为自回归项； MA 为移动平均，q 为移动平均项数，d 为时间序列成为平稳时所做的差分次数，如一阶差分：时间序列在 t 与 t-1 时刻的差值
3.原理：将非平稳时间序列转化为平稳时间序列然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型

自相关函数 ACF(autocorrelation function)

1.有序的随机变量序列与其自身相比较

2.自相关函数反映了同一序列在不同时序的取值之间的相关性

3.公式：$\mathbf{ACF(k) = \rho_k = \frac{Cov(y_t,\ y_{t-1})}{Var(y_t)}}$

4.$\mathbf{\rho_k}$的取值范围为[-1,1]

偏自相关函数(PACF)(partial autocorrelation function):

1.对于一个平稳 AR(p)模型，求出滞后 k 自相关系数ρ_k时，实际上得到并不是 x(t)与 x(t-k)之间单纯的相关关系；

2. x(t)同时还会受到中间 k-1 个随机变量 x(t-1)、 x(t-2)、 ……、 x(t-k+1)的影响，而这 k-1 个随机变量又都和 x(t-k)具有相关关系，所以自相关系数ρ_k里实际掺杂了其他变量对 x(t)与 x(t-k)的影响；

3.剔除了中间 k-1 个随机变量 x(t-1)、 x(t-2)、 ……、 x(t-k+1)的干扰之后，x(t-k)对 x(t)影响的相关程度；

4.ACF 还包含了其他变量的影响，而偏自相关系数 PACF 是严格这两个变量之间的相关性。

ARIMA(p，d，q) 参数确定：

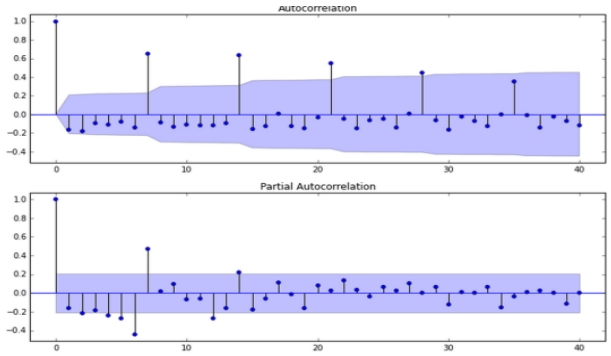
模型	ACF	PACF
AR(p)	拖尾，衰减趋于零	P 阶后截尾
MA(q)	q 阶后截尾	拖尾，衰减趋于零
ARMA(p,q)	拖尾，q 阶后衰减趋于零	拖尾，P 阶后衰减趋于零

截尾：落在置信区间内（95%的点都符合该规则）

AR(p) 看 PACF

MA(q) 看 ACF

下图是 Python 做出的 ACF 和 PACF 图：



ARIMA 建模流程：

1.将序列平稳（差分法确定 d）

2. p 和 q 阶数确定：ACF 与 PACF

3.得到模型 ARIMA（p,d,q）

模型选择 AIC 与 BIC：选择更简单的模型

1.AIC:赤池信息准则（Akaike Information Criterion，AIC）

$$AIC = 2k - 2\ln(L)$$

2.BIC:贝叶斯信息准则（Bayesian Information Criterion，BIC）

$$\mathrm{BIC} = \mathrm{kln(n)} - 2\mathrm{ln\ (L)}$$

k 为模型参数个数，n 为样本数量，L 为似然函数

模型诊断

残差检验分析：

1.残差图肉眼简单查看
2.ARIMA 模型的残差是否是平均值为 0 且方差为常数的正态分布
3.QQ 图：线性即正态分布
4.Ljung-Box 检验：独立性

过度拟合和参数冗余：

1. 在过度拟合时，不要同时增加 AR 和 MA 部分的阶数
2. 如果拟合了 MA(1)模型后，残差在 2 阶滞后处仍存在明显的相关性，那么应该尝试 MA(2)，而不是 ARMA(1,1)模型。

预测

当上面的步骤走完之后，就可以对数据进行预测了，也就是时间序列建模的主要目标之一，预测该序列未来的取值，此外我们要评估预测的精度。一般采用最小均方误差标准。

总结：

本文从时间序列的例子说起，然后从随机变量的数学特征引入随机变量序列的数学特征，之后介绍了平稳性，包括严平稳和宽平稳的概念及特点，对于非平稳时间序列的处理策略，接着给出时间序列的影响因素介绍，随后重点介绍了平稳时间序列的模型，包括 AR、MA、ARMA 和 ARIMA，给出 ARIMA 的建模流程以及相应的参数 p, q 由 PACF、ACF 的确立，以及差分阶数 d 的确立。对于模型一般采用 AIC 、BIC，此外 Python 也给出了 HQIC 的标准。一个完整的时间序列建模还需要对模型进行诊断，包括残差的正态性、独立性等检验，文中给出检验方法策略；为了避免过度拟合及参数冗余也给出一点建模技巧。最后就是对于时间序列的未来值进行预测。

拓展：

时间序列除了文中给出的 AR、 MA、ARMA、ARIMA 还有季节性的 ARIMA 模型、乘法季节 ARMA 模型、非平稳季节 ARIMA 模型，此外在金融领域也有异方差时间序列模型：ARCH、GARCH，门限模型等等，当然啦，像多元线性回归一样，也存在着多元时间序列。此外，对于时间序列数据来说又有面板数据、截面数据、函数型数据等，对于不同的数据又存在着对应的模型，这里就不再详细说明了。