时序数据处理方法

Created	@December 15, 2022 10:08 AM
Tags	
Property	

时序数据处理方法

- 平稳性检验
  - 平稳性
    1. **严平稳**

       所谓严平稳（strictly stationary），就是一种条件比较苛刻的平稳性定义，它认为只有当序列所有的统计性质都不会随着时间的推移而发生变化时，该序列才能被认为平稳。

    2. **宽平稳**

       E(X) Var(x) 对于t≥0 都是恒定的

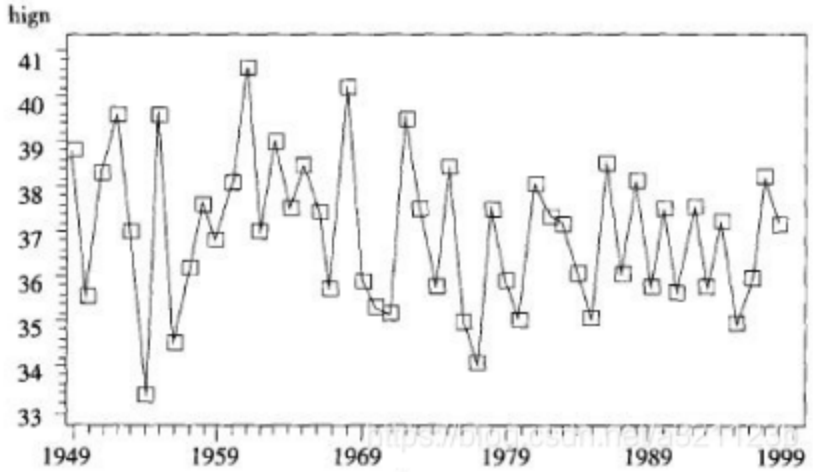
       x(t) 、x(t-h)的协方差对于t≥0 都是恒定的

       weak stationary：由序列的特征统计量判定平稳性。它认为序列的统计性质主要由它的低阶矩决定，所有只要保证序列低阶矩平稳（二阶），就能保证序列的主要性质近似稳定。

       比如服从柯西分布的严平稳序列就不是宽平稳序列，因为它不存在一、二阶矩，所以无法验证它二阶平稳。严格地讲，只有存在二阶矩的严平稳序列才能保证它一定也是宽平稳序列。

       宽平稳一般推不出严平稳，但当序列服从多元正态分布时，则二阶平稳可以推出严平稳。

  - 平稳性检验
    1. 时序图：根据平稳时间序列均值、方差为常数的性质，平稳时间序列的时序图应该是显示出该数列始终在一个常数值附近随机波动，而且波动的范围有界的特点。



2. 自相关图acf 偏自相关图pacf：

   自相关图就是一个平面二维坐标垂线图，一个坐标轴表示延迟时期数，另一个坐标轴表示自相关系数，通过以垂线表示自相关系数的大小。平稳时间序列通常具有短期相关性，该性质使用自相关系数来描述就是随着延迟期数k的增加，平稳时间序列的自相关系数ρ会很快地衰减为0；反之，非平稳序列的自相关系数ρ衰减向0的速度通常会比较慢。

3. 单位根检验

   DF ADF PP DF-GLS KPSS检验

▼ 数据预处理

1. 纠正错误

   错误数据是数据源环境中经常出现的一类问题。数据错误的形式包括：
   数据值错误： 数据直接是错误的，例如超过固定域集、超过极值、拼写错误、属性错误、源错误等。
   数据类型错误： 数据的存储类型不符合实际情况，如日期类型的以数值型存储，时间戳存为字符串等。
   数据编码错误： 数据存储的编码错误，例如将UTF-8写成UTF-80。
   数据格式错误： 数据的存储格式问题，如半角全角字符、中英文字符等。
   数据异常错误： 如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期越界、数据前后有不可见字符等。
   依赖冲突： 某些数据字段间存储依赖关系，例如城市与邮政编码应该满足对应关系，但可能存在二者不匹配的问题。
   多值错误： 大多数情况下，每个字段存储的是单个值，但也存在一个字段存储多个值的情况，其中有些可能是不符合实际业务规则的。这类错误产生的原因是业

务系统不够健全，尤其是在数据产生之初的校验和入库规则不规范，导致在接收输入后没有进行判断或无法检测而直接写入后台数据库造成的。
处理非等间隔：内插法，最常用的是线性内插法。然而内插法往往会带来显著的且不易量化的偏差，因为分析拟合误差时往往不能区分哪些是模型本身带来的误差，哪些是由于内插带来的误差。另一类是直接对数据建模，例如可以把处理等间隔时间序列方法通过某种变换再应用之，或者直接考虑新的适用于非等间隔序列（MIDAS 混频向量自回归 动态多因子）

2. 处理缺失值
   丢弃含缺失数据的记录 or 新值替代缺失数据
   a.替代法： 均值替代、最近邻域替代法
   b.内插法： 线性内插法、K-最近距离法
   c.统计模型：样条法、回归模型
   d.多重插补：先估计出待插补的值，然后再加上不同的噪声，形成多组可选插补值；对每个插补数据集合，都用针对完整数据集的统计方法进行统计分析，从而得到不同的结果；并根据评分函数进行选择，产生最终的插补值。
   e.基于EM算法的替代模型，最大似然估计，外推法

3. 离群值（Outlier）

   类型

   a. 加性离群点: 该干扰只影响发生的那一个时刻t时的序列值而不影响该时刻以后的序列值；
   b. 更新离群点: 造成离群点的干扰不仅作用于Xt，且影响t时刻以后序列的所有观测值；
   c. 水平移位离群点: 造成这种离群点的干扰是在某一时刻t系统的结构发生了变化，并持续影响t时刻以后的所有观测值，在数列上往往表现出t时刻前后的序列均值发生水平位移；
   d. 暂时变更离群点: 造成这种离群点的干扰是在t时刻干扰发生时具有一定初始效应，以后随时间根据衰减因子的大小呈指数衰减。

   处理方法

   基于统计的方法：

   不等式检验（马尔可夫不等式、切比雪夫不等式、统计置信度检验、t检验、f检验、卡方检验、Grubbs' Test 、ESD）

   孤立森林

K-means聚类

▼ 数据降维

直接降维：特征选择

线性降维：PCA MDS

主成分分析(pca):一个中心：原始特征空间的重构 两个基本点：最大投影方差 最小重构代价 主要作用：简化运算、 去除数据噪音（增加数据的信噪比） 利用散点图实现多维数据可视化、发现隐性相关变量

非线性降维: 流形 Isomap LLE

▼ 相似度计算

1. 基于欧几里德空间的 LI L2 LP L∞范数计算相似度

$$d_{L^1}(X_T, Y_T) = \sum_{t=1}^{T} |x_t - y_t|,$$

$$d_{L^p}(X_T, Y_T) = (\sum_{t=1}^{T} |x_t - y_t|^p)^{1/p},$$

$$d_{L^2}(X_T, Y_T) = (\sum_{t=1}^{T} |x_t - y_t|^2)^{1/2},$$

$$d_{L^\infty}(X_T, Y_T) = \max_{1 \le t \le T} |x_t - y_t|.$$

2. 基于相关性的相似度计算 ：**Pearson Coefficient 、The First Order Temporal Correlation Coefficient**
3. 基于自相关系数的距离
4. 基于周期性的相似度计算：周期图表：通过 Fourier 变换得到一组参数，然后通过这组参数来反映原始的两个时间序列时间的距离

5. 基于模型的相似度计算：用一个模型和相应的一组参数去拟合某条时间序列，然后得到最优的一组参数，计算两个时间序列所得到的最优参数的欧几里德距离
    a. **Piccolo 距离**
    b. **Maharaj 距离：增加样本方差矩阵修正piccolo距离**
    c. **cepstral distance：LPC距离**
6. 模式距离

    PLR（piecewise linear representation)分段线性表示：相邻的相同模式进行合并（关于PLR算法的点的分割：自底向上的搜索方法）等模式数化：使用共同的分割点，以获得最后长度相等的模式序列
7. 形状距离 模式距离*振幅
8. DTW 动态规划

▼ 表示与信息提取

1. 熵特征

    Binned Entropy

    Approximate Entropy

    Sample Entropy
2. 分段特征

    分段聚合逼近（Piecewise Aggregate Approximation）

    符号逼近（Symbolic Approximation）

▼ 时间序列分解

1.朴素分解

加法模型 乘法模型 混合模型 移动平均 趋势外推

2. X11
3. STL分解
4. 滤波
5. 状态空间模型