Deep Learning for Time Series Cookbook

Stephen CUI

May 23, 2024

Contents

Chapter 1

时间序列入门

1.1 重新采样时间序列

更改时间序列的频率是分析前的常见预处理步骤。例如，前面的指南中使用的时间序列具有小时粒度。然而，我们的目标可能是研究每日变化。在这种情况下，我们可以将数据重新采样到不同的时间段。重新采样也是处理不规则时间序列（以不规则间隔的时间段收集的时间序列）的有效方法。

我们将讨论重新采样时间序列可能有用的两种不同场景：改变采样频率和处理不规则时间序列。

大多数时间序列分析方法都假设时间序列是规则的；换句话说，它是以规则的时间间隔（例如每天）收集的。但有些时间序列天生就是不规则的。例如，零售产品的销售发生在顾客到达商店的任意时间戳。

不规则的时间序列可以通过重采样转换为规则的频率。

1.2 处理缺失值

缺失值是困扰所有类型数据（包括时间序列）的一个问题。观测值通常由于各种原因而无法获得，例如传感器故障或注释错误。在这种情况下，可以使用数据插补来克服此问题。数据插补的工作原理是根据某种规则（例如平均值或某个预定义值）分配一个值。

在没有时间顺序的数据集中，通常使用中心统计数据（例如平均值或中位数）来填补缺失值。

时间序列插补必须考虑观测的时间性质。这意味着分配的值应遵循序列的动态。时间序

列中更常见的方法是使用最后已知的观测值来填补缺失数据。

1.3 分解时间序列

时间序列分解是将时间序列分解为其基本组成部分（例如趋势或季节性）的过程。

时间序列由趋势、季节性和残差三部分组成：

- 趋势体现了时间序列水平的长期变化。趋势可以是向上（水平增加）或向下（水平减少），也可以随时间而变化。
- 季节性是指固定时间段（例如每天）内的规律变化。
- 时间序列的残差部分（也称为不规则部分）是去除趋势和季节性成分后剩下的部分。

将使用两种方法描述时间序列分解的过程：经典分解方法和基于局部回归的方法。

对于具有年度季节性的每日时间序列，周期应设置为 365，即一年中的天数。需要先确定周期，然后设置每个周期中的样本数据点。

最流行的时间序列分解方法之一是 STL（使用 LOESS 进行季节和趋势分解）。对于 STL 来说，不需要像使用经典方法那样指定模型。通常，时间序列分解方法假设数据集包含单一季节性模式。然而，以高采样频率（例如每小时或每天）收集的时间序列可能包含多个季节性模式。例如，每小时时间序列可以显示常规的每日和每周变化。

MSTL() 方法（Multiple STL 的缩写）扩展了具有多个季节性模式的时间序列的 MSTL 。可以在元组中指定每个季节性模式的周期作为周期的输入参数。

在经典分解中，使用移动平均值来估计趋势，例如过去 24 小时的平均值（对于每小时序列）。通过对每个周期的值取平均值来估计季节性。STL 是一种更灵活的时间序列分解方法。它可以处理复杂的模式，例如不规则趋势或异常值。STL 利用 LOESS （局部加权散点图平滑）来提取每个成分。

1.4 自相关

自相关是衡量时间序列与其自身在不同滞后时间之间的相关性的指标，它有助于理解时间序列的结构，具体来说，可以量化过去的值如何影响未来。

相关性是衡量两个随机变量之间线性关系的统计数据。自相关将这一概念扩展到时间序列数据。在时间序列中，在给定时间步长内观察到的值将与之前观察到的值相似。自相关函数量化时间序列与其自身滞后版本之间的线性关系。滞后时间序列是指在多个时间段内偏移的时间序列。

偏自相关函数是识别自回归模型阶数的重要工具。其思想是选择偏自相关显著的滞后数。

1.5 检测平稳性

平稳性是时间序列分析的核心概念，也是许多时间序列模型做出的重要假设。

如果时间序列的统计特性不变，则该时间序列是平稳的。这并不意味着序列不会随时间而变化，只是它变化的方式本身不会随时间而变化。这包括时间序列的水平，在平稳条件下是恒定的。趋势或季节性等时间序列模式会破坏平稳性。因此，在建模之前处理这些问题可能会有所帮助。

差分是取连续观测值之间的差值的过程。这个过程分为两个步骤：

1. 估计平稳性所需的差分步数。
2. 应用所需次数的差分运算。

差分也可以应用于季节性时期。在这种情况下，季节性差分涉及计算同一季节期间连续观测值之间的差异。

1.6 处理异方差

时间序列的方差会随时间而变化，这也违反了平稳性。在这种情况下，时间序列被称为异方差，通常呈现长尾分布。这意味着数据是左偏或右偏的。这种情况是有问题的，因为它会影响神经网络和其他模型的训练。

处理非常量方差分为两个步骤。首先，我们使用统计检验来检查时间序列是否为异方差。然后，我们使用对数等变换来稳定方差。

我们可以使用 White test or the Breusch-Pagan test 等统计检验来检测异方差。

检验的输出是 p 值，其中零假设假定时间序列具有恒定方差。因此，如果 p 值低于显著性水平，我们将拒绝零假设并假设异方差。

处理非常量方差的最简单方法是使用对数转换数据。

对数是 Box-Cox 变换的一个特例，可在 scipy 库中找到。stats.boxcox() 方法估计一个转换参数 lmbda，它可用于逆操作。

$$y = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases} \tag{1.1}$$

对数转换可稳定时间序列的方差。它们还使数据分布更接近正态分布。这些转换对于神经网络特别有用，因为它们有助于避免饱和区域。在神经网络中，当模型对不同的输入变得不敏感时就会发生饱和，从而损害训练过程。

Yeo-Johnson 幂变换与 Box-Cox 变换类似，但允许时间序列中出现负值。

1.7 重新采样多元时间序列

重采样对于多变量时间序列来说可能有点棘手，因为最终需要对不同的变量使用不同的汇总统计数据。