时间序列分析
作业一

502022370071, 庄镇华, zhuangzh@lamda.nju.edu.cn

2022 年 10 月 19 日

作业提交注意事项

(1) 请严格参照教学立方网站所述提交作业，压缩包命名统一为学号 _ 姓名.zip;

(2) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数;

(3) 除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

1 [100pts] 预处理、简单模型和评价指标

ETT (Electricity Transformer Temperature) 数据集是一个多变量时间序列数据集，本次实验选取 ETTh1 中" 油温 (OT)" 组分作为单独的单变量时间序列，该序列记录了每个小时 (h) 的油温变化，学生需要以此序列为实验对象，实现若干预处理方法、简单模型和评价指标。数据集、代码分别保存在 data、code 文件夹下。学生需要完成的任务如下:

(1) 在 transforms.py 中实现归一化（normalization），标准化（Standardization），平均归一化（Mean Normalization），Box-Cox 变换，需继承 Transform 类并实现其抽象方法;

(2) 在 models.py 中实现 $Naive_1$，$Naive_s$(以 24h 为周期)，Drift 模型，需继承 ForecastModel 类并实现其抽象方法;

(3) 在 metrics.py 中实现 MSE，MAE，MAPE，sMAPE，MASE。

(4) 修改并运行 main.py，汇报 (2) 中不同方法，在 (1) 中不同变换下，用 (3) 中不同指标衡量的性能，以表格形式呈现，表格示例如表1所示。绘制并报告 (2) 中模型真实序列与预测序列的曲线图 (变换方式任选，但需在解答中说明)。

注：utils.py 中包含了一些有用的函数，请勿对其进行修改，如有疑问可联系助教。最终需提交的文件为: 1. 修改后的代码，要求附加一个 markdown 格式的文件 README.md，说明如何复现报告中的结果。2. pdf 形式的报告，报告需描述各个功能的实现（例如以数学公式的形式）并报告结果，写于 **solution** 部分即可。

Solution. 此处用于报告（中英文均可）

表 1: 表格示例

Model	Transform	MAE	MSE	MAPE
Drift	None			
	Normalize			
	Box-Cox			
$Naive_1$	None			
	Normalize			
	Box-Cox			

表 2: 不同方法在不同变换下的性能指标

Model	Transform	MSE	MAE	MAPE	sMAPE	MASE
$Naive_1$	None	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Normalize	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Standardization	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	MeanNormalize	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Box-Cox(-1)	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Box-Cox(0)	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Box-Cox(0.5)	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
	Box-Cox(1)	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
$Naive_s$	None	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Normalize	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Standardization	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	MeanNormalize	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Box-Cox(-1)	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Box-Cox(0)	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Box-Cox(0.5)	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
	Box-Cox(1)	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
Drift	None	42.38380405	5.030783023	85.42761343	99.99108322	3.081858115
	Normalize	42.38380405	5.030783023	85.42761343	99.99108322	3.081858115
	Standardization	42.38380405	5.030783023	85.42761343	99.99108322	3.081858115
	MeanNormalize	42.38380405	5.030783023	85.42761343	99.99108322	3.081858115
	Box-Cox(-1)	16.81469076	3.503086382	93.86983314	55.78907566	2.145991021
	Box-Cox(0)	19.77990009	3.735680781	87.22352571	61.86052273	2.28847837
	Box-Cox(0.5)	25.27639353	4.090526961	84.05539909	72.02739739	2.50585717
	Box-Cox(1)	42.38380405	5.030783023	85.42761343	99.99108322	3.081858115

(1) 数据变换即对数据进行规范化处理，以便于后续的信息挖掘。常见的数据变换包括：归一化（*normalization*），标准化（*Standardization*），平均归一化（*Mean Normalization*），*Box-Cox* 变换等。下面以数学公式的形式描述各种变换的实现方法。

归一化（*normalization*）：将时序数据取值限制在 $[0,1]$

$$y'_t = \frac{y_t - y_{min}}{y_{max} - y_{min}}$$

标准化（*Standardization*）/ *Z-Score*：将时序数据变换为 *0* 均值以及标准方差

$$y'_t = \frac{y_t - \mu}{\sigma}$$

平均归一化（*Mean Normalization*）：

$$y'_t = \frac{y_t - \mu}{y_{max} - y_{min}}$$

Box-Cox 变换用于分布“正态”程度矫正：其中 λ 为超参数，实验选取了*-1,0,0.5,1* 四个值

$$y_t^{(\lambda)} = \begin{cases} \frac{(y_t^{\lambda}-1)}{\lambda} & \lambda \neq 0 \\ \log y_t & \lambda = 0 \end{cases}$$

由于实验数据集含有负数，因此使用二参数 *Box-Cox* 变换，由于最小值大于*-5*，因此选取 $\lambda_2 = 5$

$$y_t^{(\lambda)} = \begin{cases} \frac{((y_t+\lambda_2)^{\lambda_1}-1)}{\lambda_1} & \lambda_1 \neq 0 \\ \log(y_t+\lambda_2) & \lambda_1 = 0 \end{cases}$$

(2) 在给定时间序列 $\{y_1, y_2, \cdots, y_n\}$ 的情况下，以数学公式的形式描述各种模型的实现方法。

$Naive_1$：常数预测模型。使用最后一个观察值对后续样本进行预测

$$y'_{n+h} = y_n$$

$Naive_s$：对于周期性时间序列，使用上一个周期同期的观察值作为当前时刻的预测值，实验周期 $m = 24$。

$$y'_{n+h} = y_{n+h-m(\lfloor \frac{h-1}{m} \rfloor + 1)}$$

Drift 方法：充分考虑到时间序列前后的变化。每两个相邻的时间序列可以计算变化值的均值，用这一变化值指导后续的预测。

$$y'_{n+h} = y_n + \frac{h}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1}) = y_n + h(\frac{y_n - y_1}{n-1})$$

(3) 下面以数学公式的形式描述各种指标的计算方法。

MSE (Mean Square Error)

$$MSE = \frac{1}{H}\sum_{i=1}^{H}(y_{n+i} - y'_{n+i})^2$$

MAE (Mean Absolute Error)

$$MAE = \frac{1}{H}\sum_{i=1}^{H}|y_{n+i} - y'_{n+i}|$$

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100}{H}\sum_{i=1}^{H}\frac{|y_{n+i} - y'_{n+i}|}{|y_{n+i}|}$$

sMAPE (symmetric MAPE)

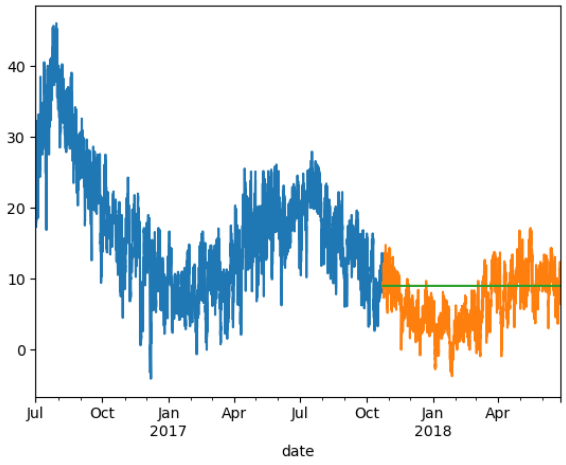
$$sMAPE = \frac{200}{H}\sum_{i=1}^{H}\frac{|y_{n+i} - y'_{n+i}|}{|y_{n+i}| + |y'_{n+i}|}$$

MASE (Mean Absolute Scaled Error)

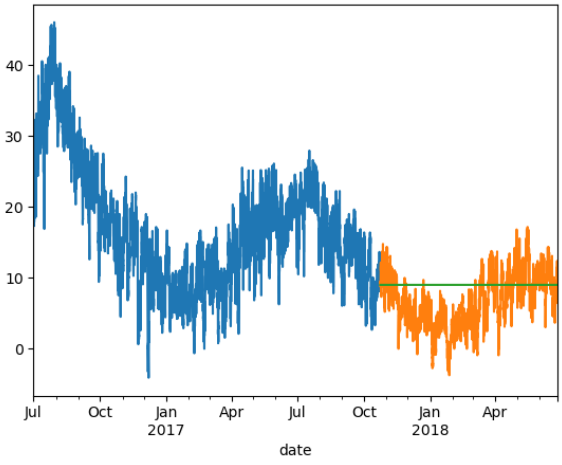
$$MASE = \frac{1}{H}\sum_{i=1}^{H}\frac{|y_{n+i} - y'_{n+i}|}{\frac{1}{n+H-m}\sum_{j=m+1}^{n+H}|y_j - y_{j-m}|}$$

(4) 不同方法在不同变换下的性能指标如表*2*所示，针对二参数 *Box-Cox* 模型，我们选取 $\lambda_1 = -1, 0, 0.5, 1$ 观察实验结果，$\lambda_2 = 5$ 为定值用于调整负值；
三种模型预测序列与真实序列的曲线如图*1*所示，选取两种变换方式，分别为 *Standardization* 和 *Box-Cox(0)* 变换。

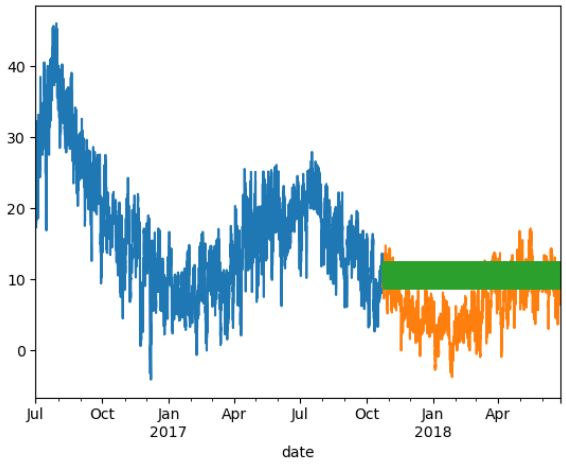
图 1: 三种模型预测序列与真实序列的曲线



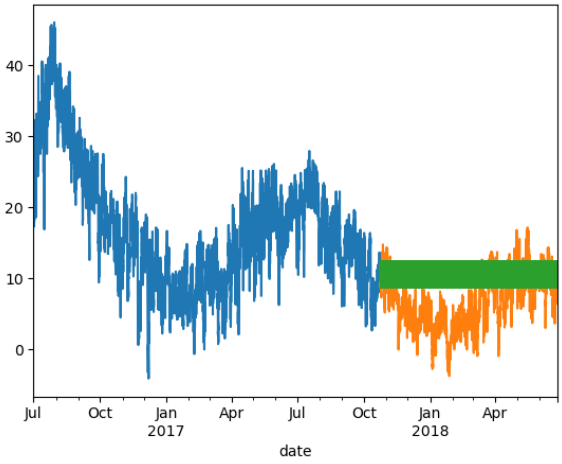
$Naive_1$ 模型 Standardization 变换



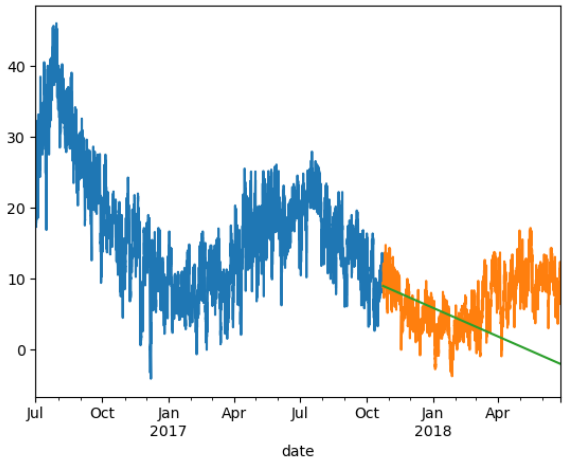
$Naive_1$ 模型 Box-Cox(0) 变换



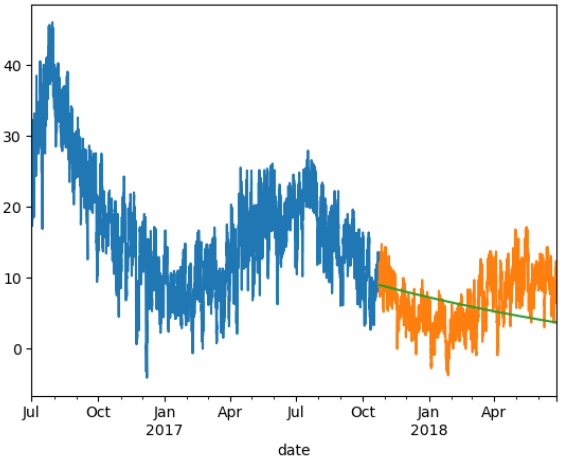
$Naive_s$ 模型 Standardization 变换



$Naive_s$ 模型 Box-Cox(0) 变换



$Drift$ 模型 Standardization 变换



$Drift$ 模型 Box-Cox(0) 变换

表 3: 不同周期对 $Naive_s$ 模型性能的影响

T/h	MSE	MAE	MAPE	sMAPE	MASE
1	18.65752471	3.535363442	111.7803963	53.57583421	2.16576395
2	18.34002584	3.509119488	110.6913492	53.37720748	2.149686902
3	18.02599282	3.483062658	109.5532941	53.18038683	2.133724485
4	19.14936796	3.57654066	113.0188586	53.88935066	2.190989118
5	20.5444861	3.698082518	117.2391948	54.81394785	2.26544567
6	21.89197442	3.808982693	120.693299	55.62994634	2.333383127
9	26.25339061	4.192032928	132.3722014	58.4604395	2.568039734
12	28.08862302	4.350078946	136.9206051	59.60811817	2.664858738
24	28.80465239	4.399937512	138.6976645	59.92193047	2.6954021
36	33.99119983	4.838767039	149.8191526	62.98541309	2.964229106
48	30.40258072	4.52262551	140.6586855	60.69909803	2.770560778
60	31.6850802	4.640361143	145.223024	61.61661936	2.842685637
72	30.35224837	4.513158591	141.5767519	60.65336168	2.764761343
120	26.97899457	4.184694546	132.6652017	58.10934586	2.563544241
168	24.06728953	4.024854047	122.7379699	57.40610562	2.465625937
336	22.83873419	3.907338463	110.5461512	57.45510201	2.39363588
720	34.95754482	4.781684738	141.498658	62.21897926	2.929260484
1440	51.54356031	5.953657632	171.4524685	68.87921892	3.64721118
2160	83.71203853	7.778575769	216.2894885	78.15369892	4.765156188

2 [附加题 20pts] 周期的影响

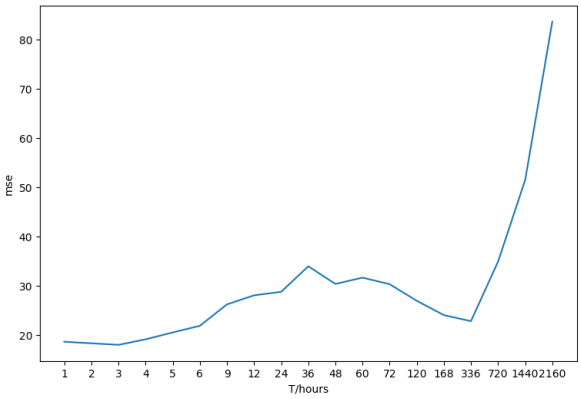
在上一题中，我们指定了 $Naive_s$ 方法中的周期为天 (24h)，本题中学生需要尝试使用不同的周期，考察不同周期下 $Naive_s$ 模型的性能变化。请绘制出模型性能随不同周期的变化曲线。在 ETTh1 的 OT 序列上，最好的周期是什么？你可以得出什么结论?

Solution. 此处用于报告 *(*中英文均可*)*

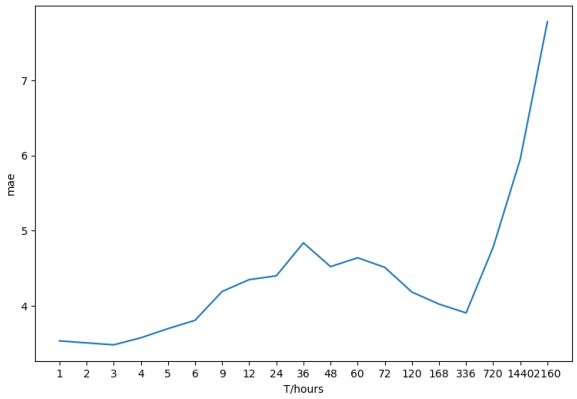
分别计算周期 $T = 1h, 2h, 3h, 4h, 5h, 6h, 9h, 12h(0.5d), 24h(1d), 36h, 48h(2d), 60h, 72h(3d),$ $120h(5d), 168h(7d)$,$336h(14d), 720h(30d), 1440h(60d), 2160h(90d)$ 的模型性能，其中变换方式取 *Standardization*，得到结果如表*3*所示，将其可视化为图*2*，可以得知最好的周期是 3 小时。

可以得出结论：对于 $Naive_s$ 模型，周期对其性能有很大的影响，过短的周期无法把握整体的趋势，过长的周期无法刻画瞬时的变化，只有选取恰当的周期才能达到最好的效果。

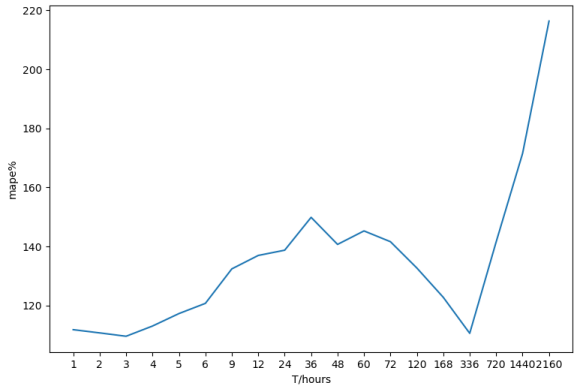
图 2: 不同周期对 $Naive_s$ 模型性能的影响



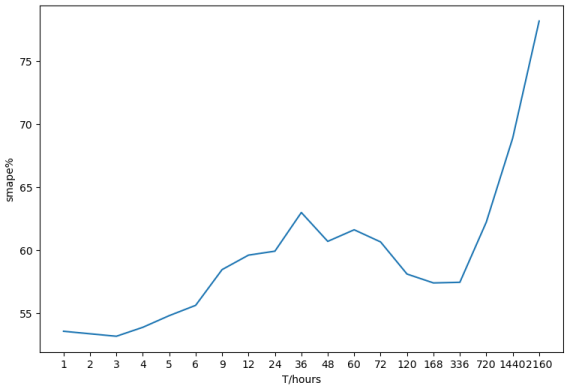
MSE



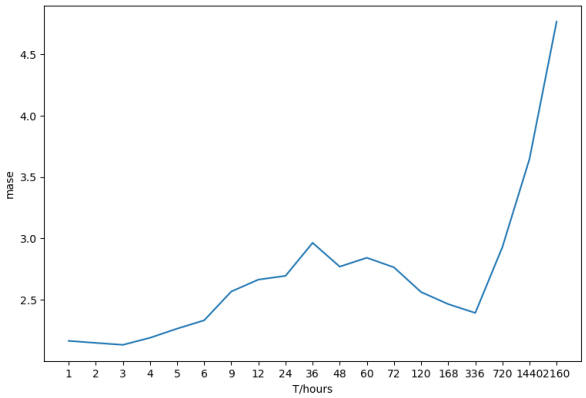
MAE



MAPE



sMAPE



MASE