时间序列分析算法

Tianxiang Yang yang20202027@hotmail.com

2025.02

时间序列定义

时间序列预测是一种基于历史数据对未来数值进行预测的方法----按照时间顺序排列的一系列观测值，如股票价格、气温变化、销售额等；这些数据中隐含了时间上的依赖关系，即当前值往往受到过去值的影响；时间序列预测的目标是捕捉这些依赖关系并预测未来数值

顺序

TIME	Feature 1	Feature 2	…	Feature N	Label
S_1	X_1(S_1)	X_2(S_1)	…	X_N(S_1)	y(S_1)
S_2	X_1(S_2)	X_2(S_2)	…	X_N(S_2)	y(S_2)
…	…	…	…	…	…
S_M	X_1(S_M)	X_2(S_M)	…	X_N(S_M)	y(S_M)
S_M+1	X_1(S_M+1)	X_2(S_M+1)	…	X_N(S_M+1)	?

滞后特征

时间特征

时间戳（TIME）最好按“年-周（月）-日-时”进行切片

所有可用时间解释的变量；可只考虑较重要的特征维度，或采用降维后特征

时间序列定义

例如，美国东北部和中部地区的小时用电量（单位：MW，AEP_hourly.csv）可看作时间相关变量，就可假设当前值往往受到过去值的影响，基于历史小时用电量对未来情况进行预测

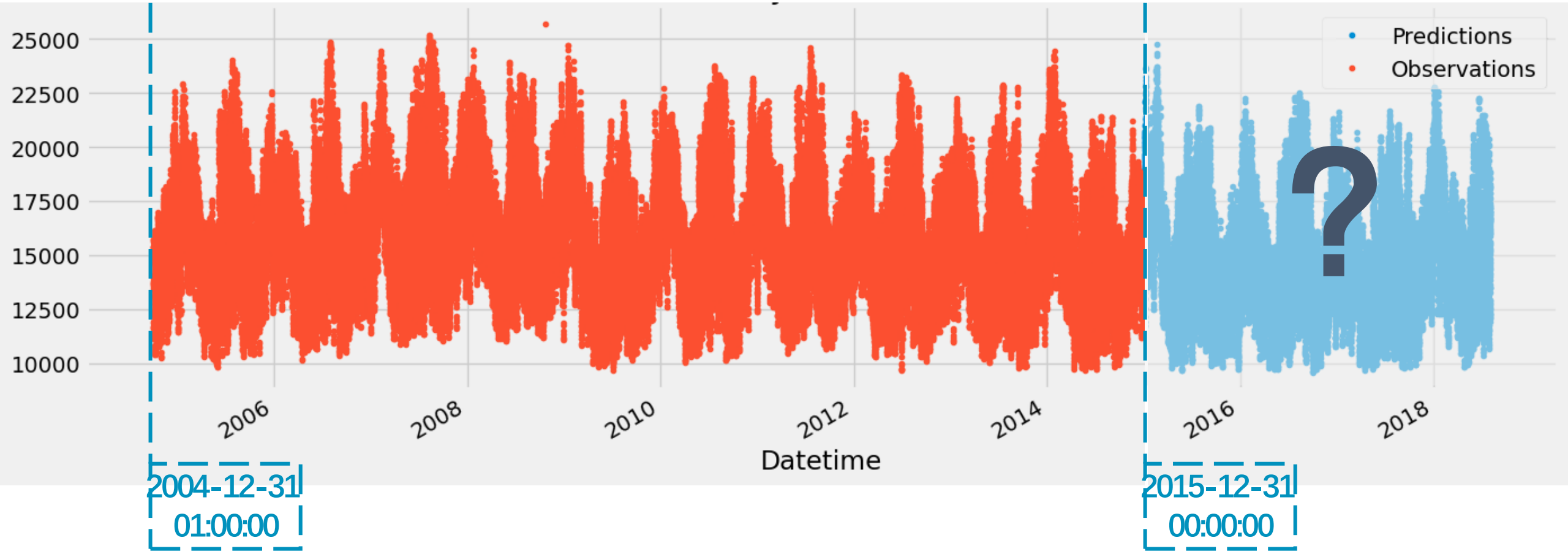
Datetime	AEP_MW	Datetime	AEP_MW	.....	Datetime	AEP_MW
2004-12-31 01:00:00	13478.0	2005-01-01 01:00:00	14097.0		2014-12-31 01:00:00	15993.0
2004-12-31 02:00:00	12865.0	2005-01-01 02:00:00	13667.0		2014-12-31 02:00:00	15780.0
2004-12-31 03:00:00	12577.0	2005-01-01 03:00:00	13451.0		2014-12-31 03:00:00	15769.0
2004-12-31 04:00:00	12517.0	2005-01-01 04:00:00	13379.0		2014-12-31 04:00:00	15738.0
2004-12-31 05:00:00	12670.0	2005-01-01 05:00:00	13506.0		2014-12-31 05:00:00	15993.0
.....	.....	.....	.....		.....	.....
2005-01-01 00:00:00	12892.0	2005-01-02 00:00:00	14316.0		2015-01-01 00:00:00	16375.0

每小时变化

每日变化(可能存在的规律：按周、按月、按季节……)

时间序列定义

美国东北部和中部地区的小时用电量（单位：MW）随时间序列的变化曲线：



时间序列预测

思路1


自回归AR
移动平均MA
自回归移动平均ARMA
差分
季节性差分
差分自回归移动平均ARIMA
季节性差分自回归移动平均SARIMA

时间序列预测：自回归AR

自回归模型（Auto Regressive，AR）假设当前时间点的观测值是其过去几个时间点观测值的线性组合，规定输出变量线性依赖于其自身之前的值和一个随机项（符合线性回归一般假设）；**适用于当前值主要依赖于其过去值的场景**（在前值预测基础上不做改变继续往后预测）：

$$y[t] = c + \phi[1] \cdot y[t-1] + \phi[2] \cdot y[t-2] + \cdots + \phi[p] \cdot y[t-p] + \epsilon[t]$$

更普遍的一种形式是向量自回归模型（VAR），可视作多变量回归，不仅考虑到y在过去的数据，**还可包含另一个变量系列（features）X的数据（其他基于状态的回归同理）**：

$$y[t] = c + \phi[1] \cdot y[t-1] + \phi[2] \cdot y[t-2] + \cdots + \phi[p] \cdot y[t-p] + \epsilon[t]$$
$$+ \psi[1] \cdot X[t-1] + \psi[2] \cdot X[t-2] + \cdots + \psi[p] \cdot X[t-p]$$

特点：能抓住一些大趋势，对局部异动变化较难适应（无法演进模型）

时间序列预测：移动平均MA

移动平均模型（Moving Average，MA）假设当前时间点的观测值是过去几个时间点随机误差项的线性组合（即用白噪声做线性组合），误差是服从正态分布并且相互独立的；**适用于那些当前值主要受过去随机波动影响的场景**（数值受周期/不规则变动的影响较大）：

$$y[t] = c + \epsilon[t] + \theta[1]\cdot\epsilon[t-1] + \theta[2]\cdot\epsilon[t-2] + \cdots + \theta[q]\cdot\epsilon[t-q]$$

特点：能有效地消除预测中的随机波动，对大趋势较难适应

时间序列预测：自回归移动平均ARMA

自回归移动平均模型（Auto Regressive Moving Average，ARMA）是AR模型和MA模型的结合，同时考虑时间序列的自身历史值和随机误差项对当前值的影响；ARMA模型适用于那些既包含自相关性又包含移动平均特性的时间序列数据：

$$y[t] = c + \phi[1]\cdot y[t-1] + \phi[2]\cdot y[t-2] + \cdots + \phi[p]\cdot y[t-p] + \epsilon[t]$$
$$+ \theta[1]\cdot \epsilon[t-1] + \theta[2]\cdot \epsilon[t-2] + \cdots + \theta[q]\cdot \epsilon[t-q]$$

也可包含另一个变量系列（features）X的向量自回归移动平均模型（VARMA）

→即　状态空间模型

特点：结合自回归和移动平均，能兼顾整体趋势和随机波动，但不能预测非平稳时间序列

【TSF_VARMA.ipynb】

时间序列预测：差分自回归移动平均ARIMA

差分自回归滑动平均模型（Auto Regressive Integrated Moving Average Model，ARIMA）是在ARMA模型的基础上增加了差分步骤，以处理非平稳时间序列；通过差分将非平稳时间序列转化为平稳时间序列，然后再用ARMA模型进行拟合，以预测股票价格、经济指标等：

$$\phi p(B) \cdot (1-B)X[t] = \theta q(B) \cdot \epsilon[t]$$

自回归项$\phi p(B) = 1 - \phi[1] \cdot B - \phi[2] \cdot B[2] - \ldots - \phi[p] \cdot B[p]$

移动平均项$\theta q(B) = 1 + \theta[1] \cdot B[1] + \theta[2] \cdot B[2] + \ldots + \theta[q] \cdot B[q]$

差分项(1阶)$(1-B)X[t] = X[t]-X[t-1]$

特点：对处理非平稳时间序列具有预测能力

时间序列预测：季节性差分自回归移动平均SARIMA

季节性差分自回归移动平均（Seasonal Auto Regressive Integrated Moving Average Model，SARIMA）是ARIMA模型的扩展，用于处理具有其他周期性变化数据；SARIMA模型通过引入季节性差分来消除周期性影响，并结合ARIMA模型以预测销售数据、气象数据等：

$$\Phi P(BS) \cdot \phi p(B) \cdot \Delta^{S} D \cdot \Delta^{d} X[t] = \Theta Q(BS) \cdot \theta q(B) \cdot \epsilon[t]$$

$$\text{自回归项} \phi p(B) = 1 - \phi[1] \cdot B - \phi[2] \cdot B[2] - \ldots - \phi[p] \cdot B[p]$$

$$\text{移动平均项} \theta q(B) = 1 + \theta[1] \cdot B[1] + \theta[2] \cdot B[2] + \ldots + \theta[q] \cdot B[q]$$

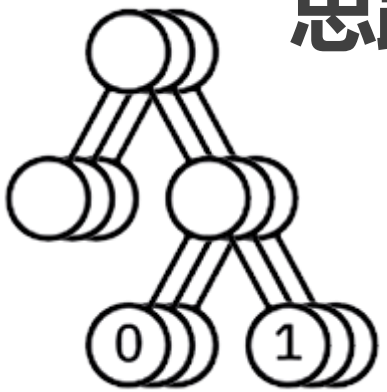
$$\text{季节性自回归项} \Phi P(BS) = 1 - \Phi[1] \cdot B[S] - \Phi[2] \cdot B[2S] - \ldots - \Phi[P] \cdot B[PS]$$

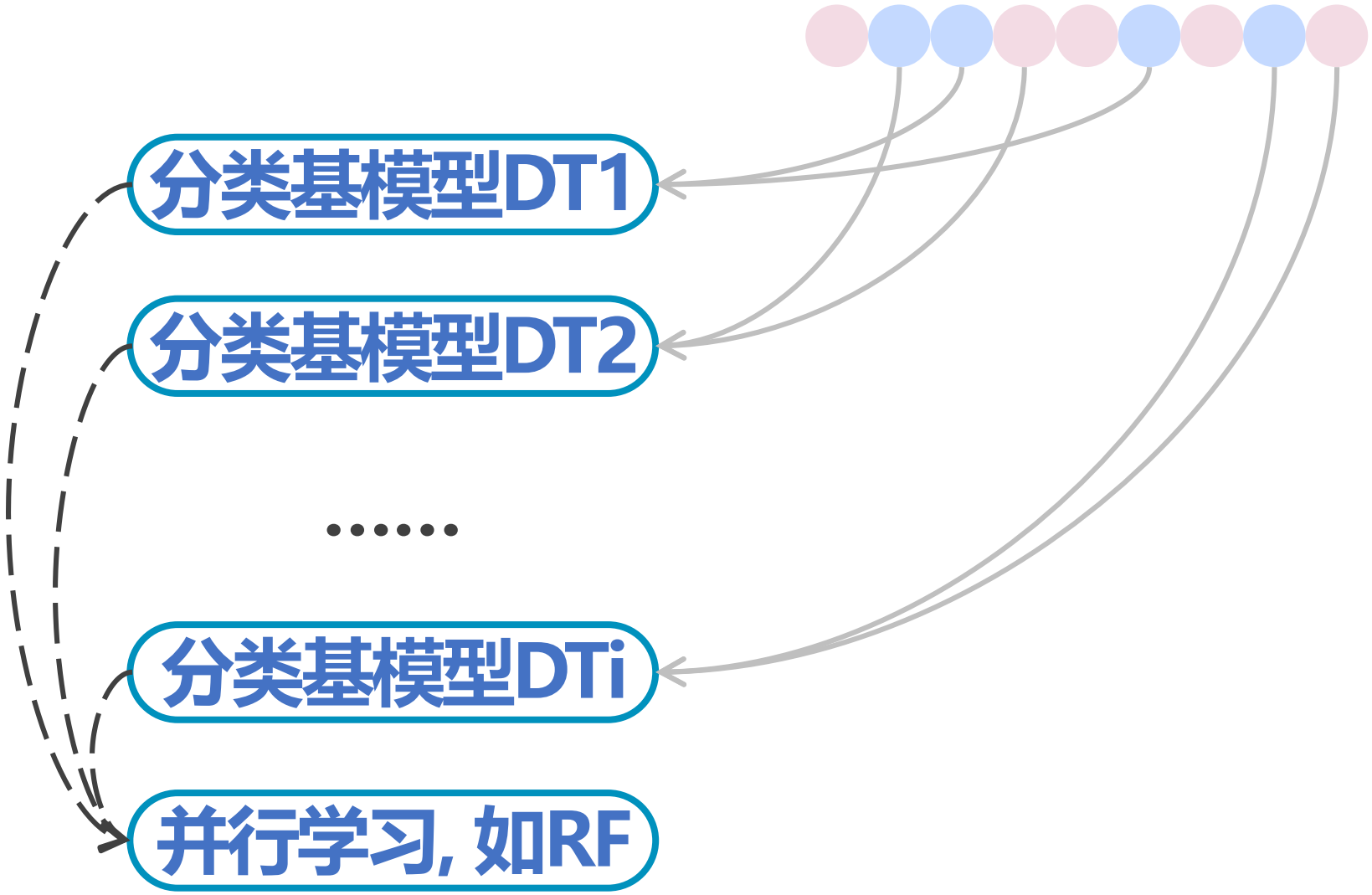
$$\text{季节性移动平均项} \Theta Q(BS) = 1 + \Theta[1] \cdot B[S] + \Theta[2] \cdot B[2S] + \ldots + \Theta[Q] \cdot B[QS]$$

特点：对处理具有更多周期性特征的非平稳时间序列具有预测能力【TSF_SARIMA.ipynb】

时间序列预测

思路2-1



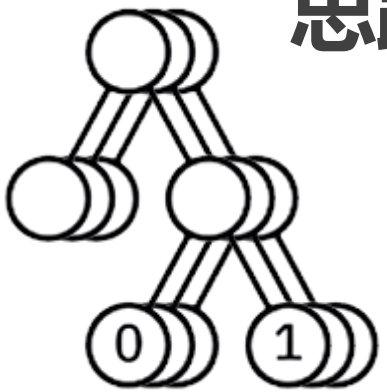


时间序列预测：并行集成模型（如RF）

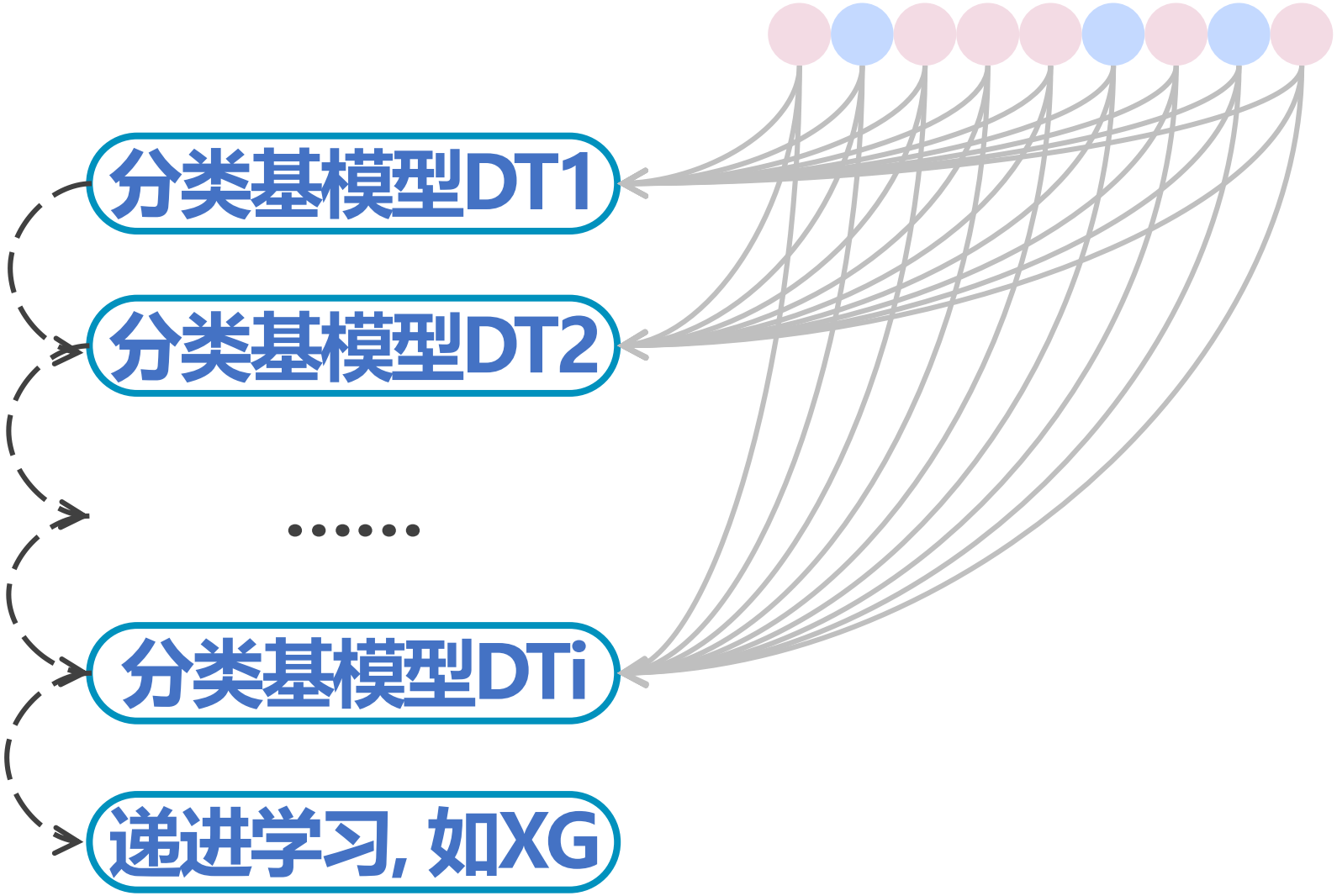
每个弱学习器使用不同的子训练集进行训练，基模型之间没有依赖关系；例如，随机森林（Random Forests，RF）等机器学习模型在时间序列预测中展现出强大的能力；构建随机森林模型进行模型的训练和预测，计算出各个特征的重要性，进行排序

特点：分类器擅长提取时间特征（事先需提取）；和其他并行式bagging类似，未引入特征强化，存在欠拟合缺陷【TSF_RF.ipynb】；对滞后特征缺乏学习能力

时间序列预测



思路2-2

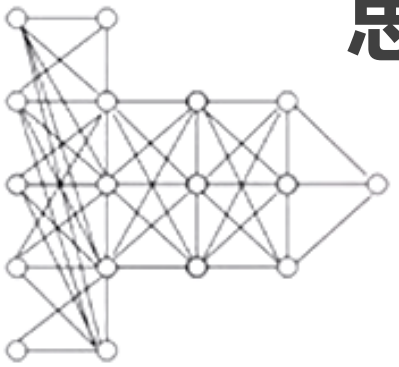


时间序列预测：递进集成模型（如XGBOOST）

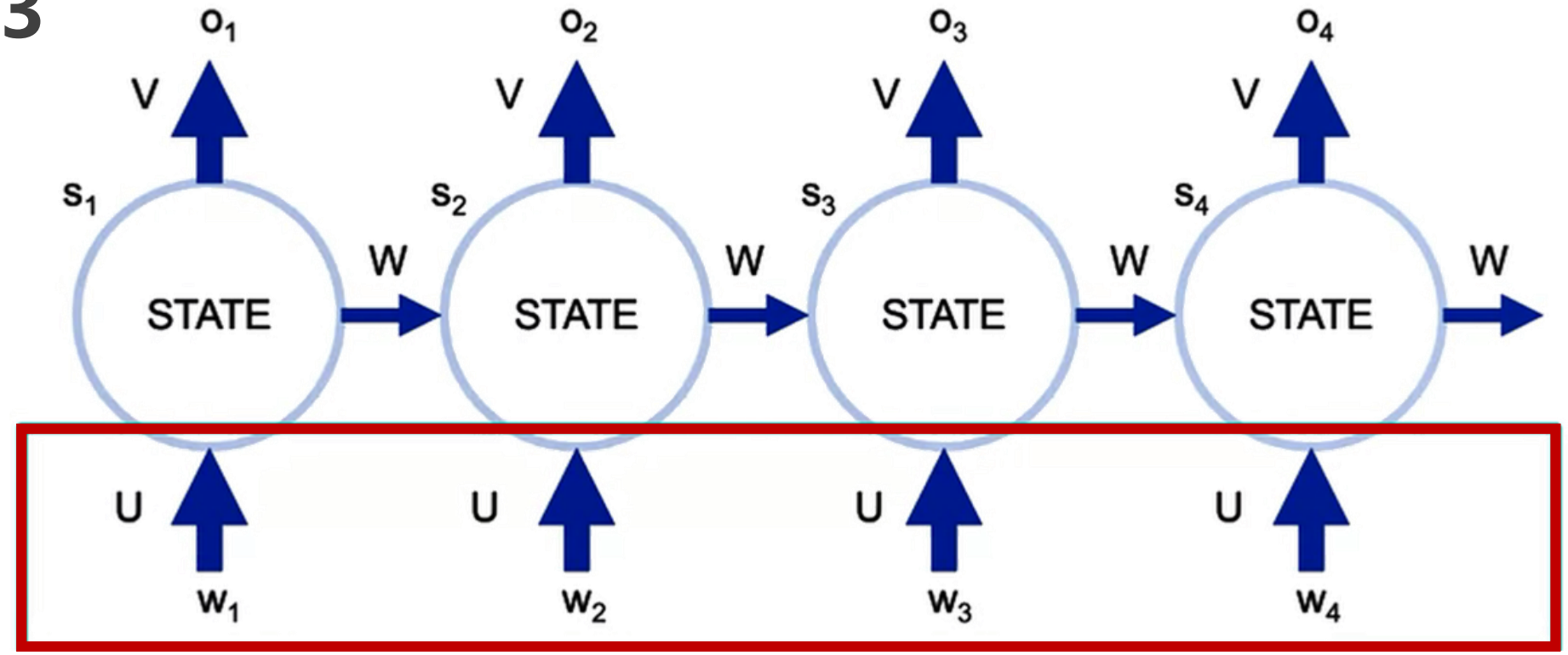
增强学习（Boost）体现在训练部分，即每个基模型学习之前所有基模型所欠缺的能力；例如，如在极端梯度提升（XGBOOST）等模型中，前面的基模型是基础，之后的基模型是修补增强，预测时则直接将所有基模型的预测结果相加；在数据分布复杂的情况下，先忽略局部凸函数，以近似全局最优逐步求解，求解过程中，当前的全局最优针对的是上一步被忽略的局部凸函数，既可减弱基模型的过拟合也可减弱Bagging的欠拟合

特点：分类器擅长提取时间特征（事先需提取）；和RF相比，总体准确率较高，但对特殊样本不敏感【TSF_BOOST.ipynb】； 对滞后特征缺乏学习能力【_TSF_BOOST_滞后特征.ipynb】

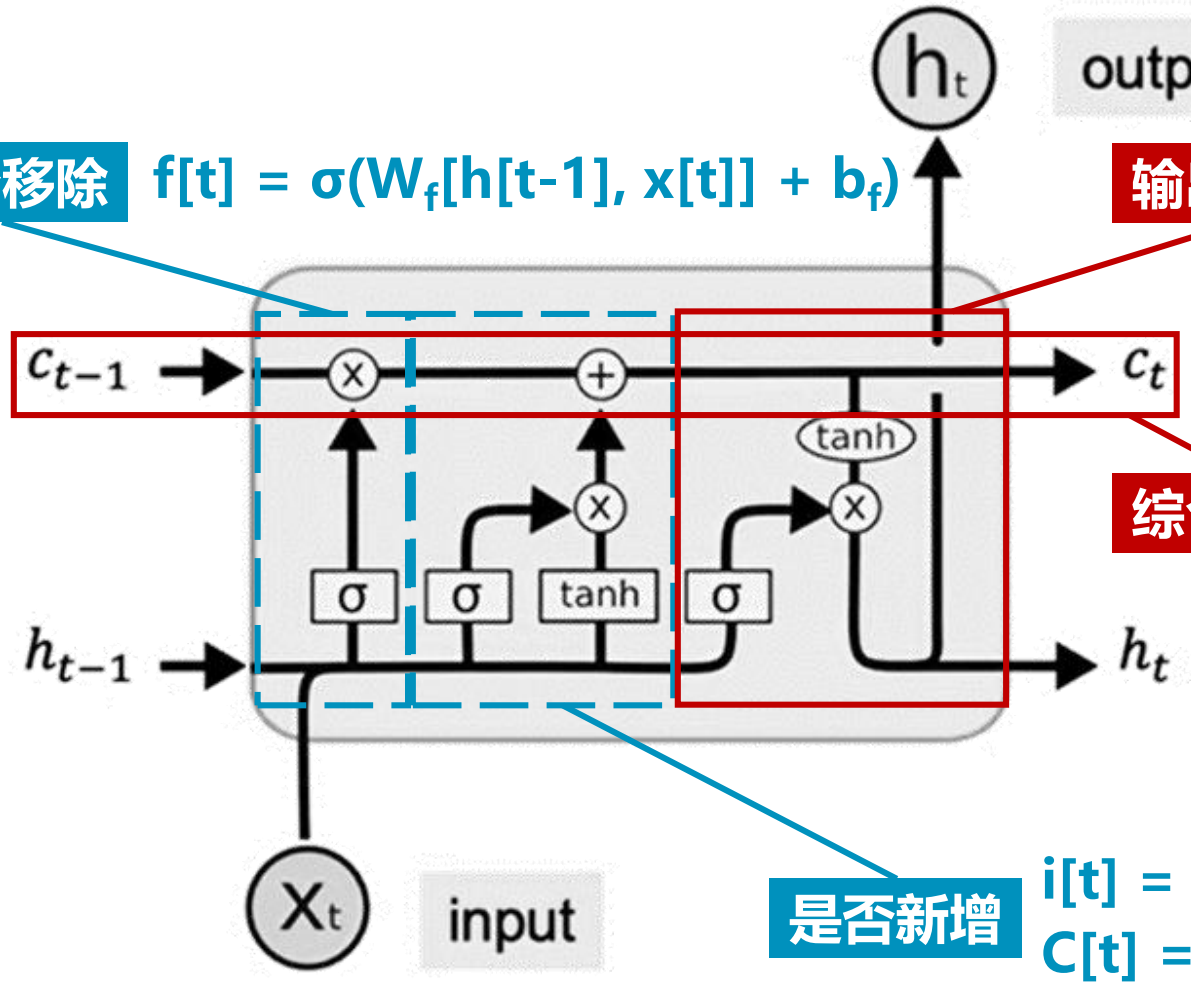
时间序列预测



思路3


o1
o2
o3
o4
V
V
V
V
s1
s2
s3
s4
W
W
W
W
STATE
STATE
STATE
STATE
U
U
U
U
w1
w2
w3
w4

时间序列预测：长短期记忆网络LSTM



是否移除 $f[t] = \sigma(W_f[h[t-1], x[t]] + b_f)$

输出判断 $o[t] = \sigma(W_o[h[t-1], x[t]] + b_o)$

$h[t] = o[t] * \tanh(C[t])$

综合判断 $C[t] = f[t] \cdot C[t-1] + i[t] \cdot C[t]'$

是否新增 $i[t] = \sigma(W_i[h[t-1], x[t]] + b_i)$ 判断程度

$C[t] = \tanh(W_C[h[t-1], x[t]] + b_C)$ 判断可否

时间序列预测：长短期记忆网络LSTM

长短期记忆网络（Long Short-Term Memory，LSTM）等深度学习模型在时间序列预测中展现出强大的能力；LSTM模型通过引入门控机制来捕捉时间序列数据中的长期依赖关系，适用于股票价格预测、气象预报等复杂场景

特点：通过滞后特征即可实现高精度预测（事先需提取序列列表）【TSF_LSTM.ipynb】；对时间特征缺乏学习能力【_TSF_LSTM_时间特征.ipynb】

算法比较

对示例数据（AEP_hourly.csv），从有效性和速度等方面横向比较目前构建的5个算法：

算法名称		VARMA	SARMA	RF	XGBOOST	LSTM
ipynb文件		TSF_VARMA	TSF_SARMA	TSF_RF	TSF_BOOST	TSF_LSTM
特征工程		滞后/时间特征 (粒度=天)	滞后特征 (粒度=天)	时间特征 (粒度=小时)	时间特征 (粒度=小时)	滞后特征 (粒度=小时)
模型参数		(p, q)=(3, 3)	(p, q)=(7, 7), (P,Q,S)=(7,7,14)	树数量1000, 最大深度3	树数量1000, 早停轮数50	LSTM50X2+ DENSE25
有效性	MAE	91390.5	1528.1	1852.8	1365.3	136.67
	MAPE	602.31	10.531	13.088	9.373	0.941
	定性评价	只能捕捉均值	波动捕捉有限	预测精度有限	预测精度较高	预测精度高
训练速度		慢；需降采样	慢；需降采样	快（<5min）	快（<5min）	中（10min）

谢 谢

Tianxiang Yang yang20202027@hotmail.com

2025.02