时间序列预测

李泽荃

2017.8.18

主要内容

- 时间序列预测概述

  什么是时间序列预测？ 时间序列有什么特点？ 时间序列预测应考虑什么因素？

  为什么时间序列必须是稳定的？如何检验时间序列的稳定性？ 如何让时间序列稳定？

- 时间序列预测的两个方法

  ARIMA and LSTM

- 案例分析

  广东省大豆价格预测

时间序列预测（Time Series Forecasting）

- 什么是时间序列预测？
  - 利用过去一段时间内某事件的时间特征来预测未来一段时间内该事件的特征。
  - 同样，也是根据统计规律构造X（t）的最佳数学模型。
  - 通常来说，时间序列预测是比较难的。
- 时间序列预测有什么特点？
  - 时间序列表现出季节趋势
  - 基于线性回归模型的假设，对于时间序列预测是不成立的
  - 时间序列模型依赖于事件发生的先后顺序

时间序列预测（Time Series Forecasting）

- 时间序列预测主要考虑的因素是什么？
  - 长期趋势：稳定或随时间呈现某种趋势
  - 季节性变动：与日期、年周期或者气候有关
  - 周期性变动：主要与经济周期有关
  - 随机影响：影响因素较多，比如突发事件
- 为什么必须是平稳序列？
  - 每一个统计学问题都需要进行一定的假设，同样时间序列预测也是
  - 一条时间序列里长期稳定不变的规律，是基本模型
  - 平稳的基本思想：时间序列的行为并不随时间改变

时间序列预测（Time Series Forecasting）

- 时间序列稳定性检验
  - 一些统计特征随着时间保持不变，可以认为它是稳定的，如平均值，方差，自协方差等
  - 稳定性检验方法：观察法和单位根检验法（ADF检验）
- 稳定性处理技术
  - 对数变换：减小数据的振动幅度
  - 平滑技术：移动平均（一段时间内的均值均为估计值）、指数平均（通过变权来提高最近值的权重）
  - 差分技术：等周期间隔的数据进行求减
  - 分解技术：将数据分离成不同的成分，比如长期趋势、季节趋势和随机成分等

差分自回归移动平均模型（ARIMA）

- **ARIMA（p,d,q）模型（Auto Regressive Integrated Moving Averages）**

是指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进行回归建立的模型。平稳时间序列的预测其实就是一个线性方程。

AR——自回归，p为自回归项，AR条件是因变量的滞后；

MA——移动平均，q为移动平均数，MA条件是预测方程的滞后预测误差；

d——差分次数

差分自回归移动平均模型（ARIMA）

（1）$AR(p)$模型（Auto regression Model）——自回归模型

p 阶自回归模型：

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + e_t$$

式中，y_t为时间序列第t时刻的观察值，即为因变量或称被解释变量；y_{t-1}，y_{t-2}，…，y_{t-p}为时序y_t的滞后序列，这里作为自变量或称为解释变量；e_t是随机误差项；c，$\emptyset_1$，$\emptyset_2$，…，$\emptyset_p$为待估的自回归参数。

差分自回归移动平均模型（ARIMA）

（2）$MA(q)$模型（Moving Average Model）——移动平均模型

q阶移动平均模型：

$$y_t = \mu + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

式中，μ为时间序列的平均数，但当$\{y_t\}$序列在0上下变动时，显然$\mu=0$，可删除此项；e_t，e_{t-1}，e_{t-2}，…，e_{t-q}为模型在第t期，第$t-1$期，…，第$t-q$期的误差；θ_1，θ_2，…，θ_q为待估的移动平均参数。

差分自回归移动平均模型（ARIMA）

（3） $ARMA(p,q)$ **模型——自回归移动平均模型（Auto regression Moving Average Model）**

模型的形式为：

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

显然，$ARMA(p,q)$ 模型为自回归模型和移动平均模型的混合模型。当 $q=0$，时，退化为纯自回归模型 $AR(p)$；当 $p=0$ 时，退化为移动平均模型 $MA(q)$。

差分自回归移动平均模型（ARIMA）

（4）ARIMA（p,d,q）模型（Auto Regressive Integrated Moving Averages）

这里的d是对原时序进行逐期差分的阶数，差分的目的是为了让某些非平稳（具有一定趋势的）序列变换为平稳的，通常来说d的取值一般为0,1,2。 对于具有趋势性非平稳时序，不能直接建立ARMA模型，只能对经过平稳化处理，而后对新的平稳时序建立ARMA（p，q）模型。这里的平稳处理可以是差分处理，也可以是对数变换，也可以是两者相结合，先对数变换再进行差分处理。

差分自回归移动平均模型（ARIMA）

- 模型判断

自相关图	偏自相关图	模型
拖尾	截尾	AR模型
截尾	拖尾	MA模型
拖尾	拖尾	ARMA模型

Facebook开源的新预测工具—Prophet

- 特点
  - 包含众多预测技术，比如ARIMA，exponential smoothing；
  - 拥有方便的调参工具；
  - 可以设定不规则日期去掉周期性影响，比如感恩节、超级碗；
  - 自带数据驱动的置信区间;
  - 异常值/离群点检测。
- 支持语言
  - Python
  - R

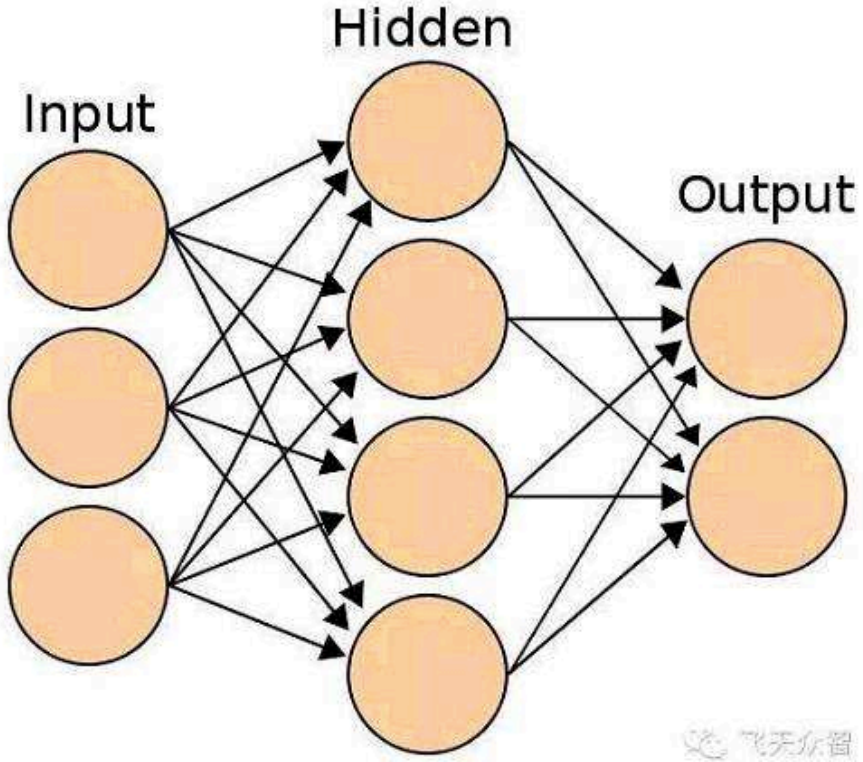
Github网址：https://github.com/facebookincubator/prophet

长短期记忆网络（LSTM）

- 循环神经网络RNN（Recurrent Neural Network）
  - 普通的神经网络结构（NN）

    1.层与层之间全连接；

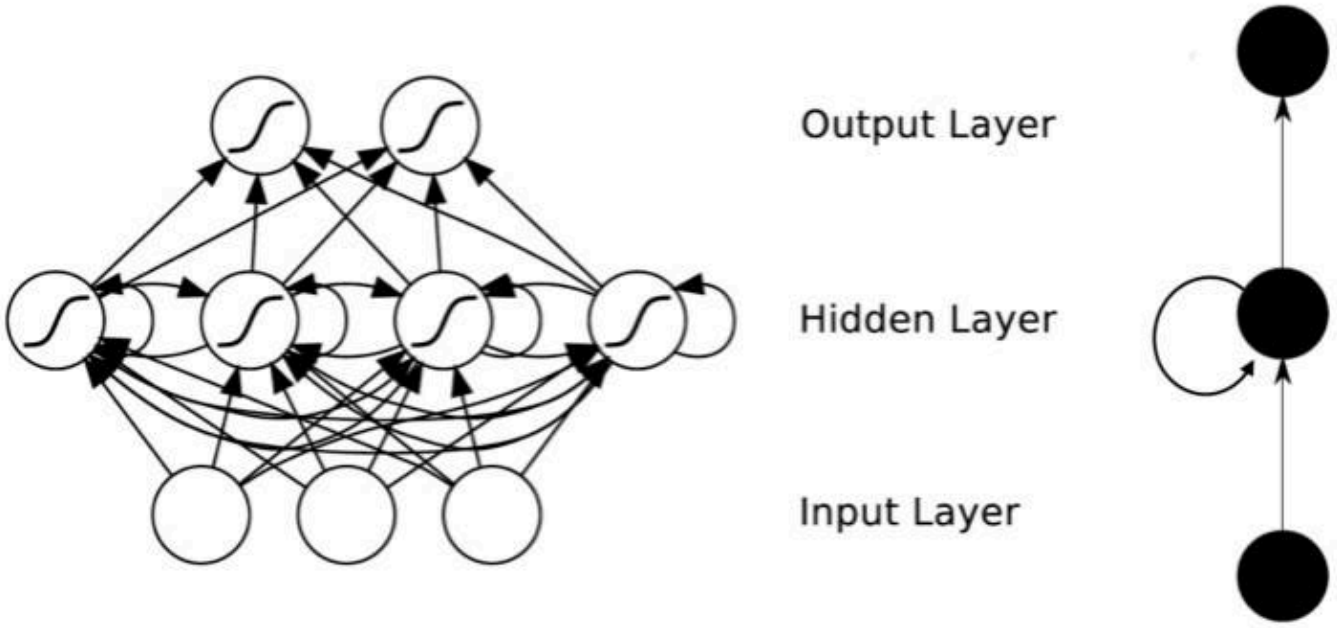
    2.层内节点之间无连接。



长短期记忆网络（LSTM）

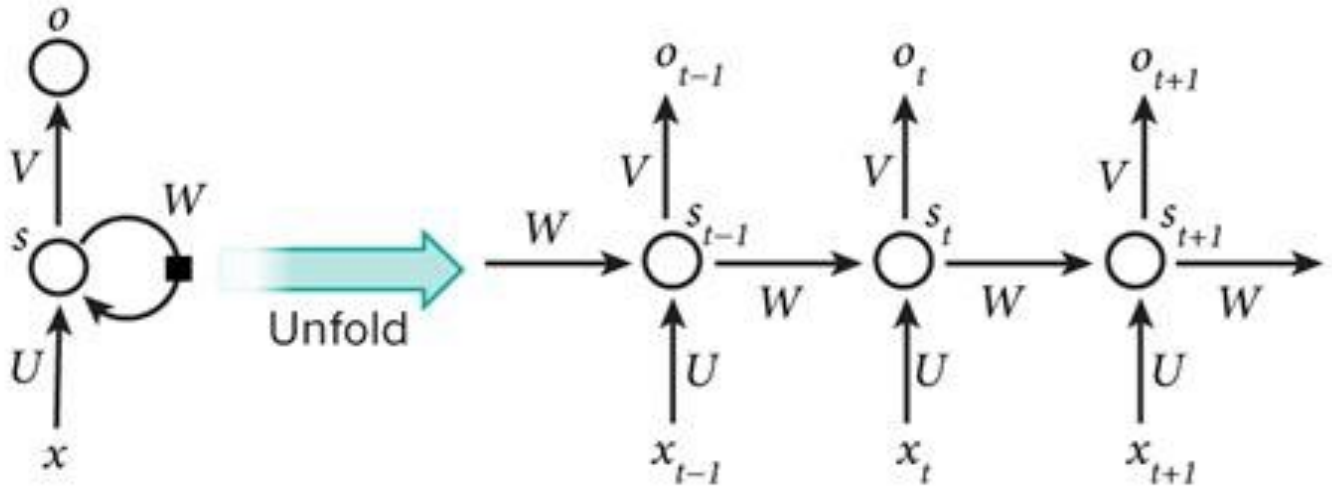
- 循环神经网络RNN（Recurrent Neural Network）

  隐藏层内节点之间也有了连接。



长短期记忆网络（LSTM）

- 循环神经网络RNN（Recurrent Neural Network）



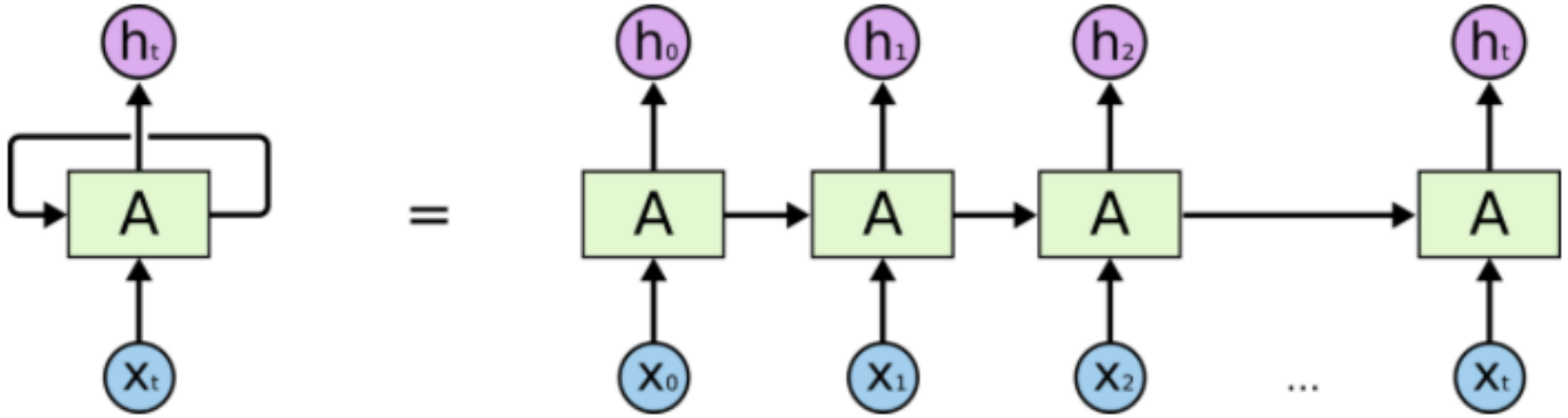
- x_t是时间t处的输入，也可以认为是第t次的计算输入；
- s_t是时间t处的“记忆状态”，$s_t=f(Ux_t+Ws_{t-1})$；
- o_t是时间t处的输出。

RNN的局限：

- 如果需要实现长期记忆的话，RNN需要将当前隐藏状态的计算与前n次的计算挂钩，模型训练时间将大幅增加；
- 在RNN中整个神经网络都共享一组参数（U,V,W）,极大减小了训练和预估的参数量。

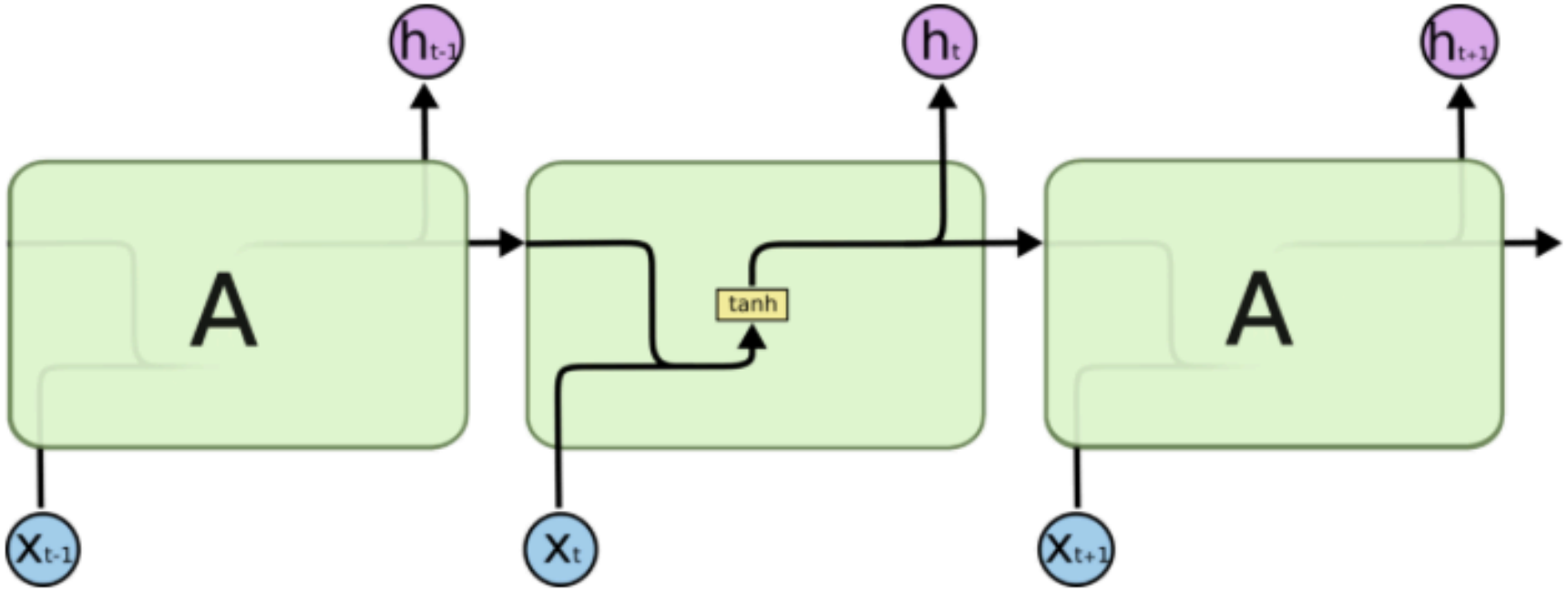
长短期记忆网络（LSTM）

- 展开的RNN结构



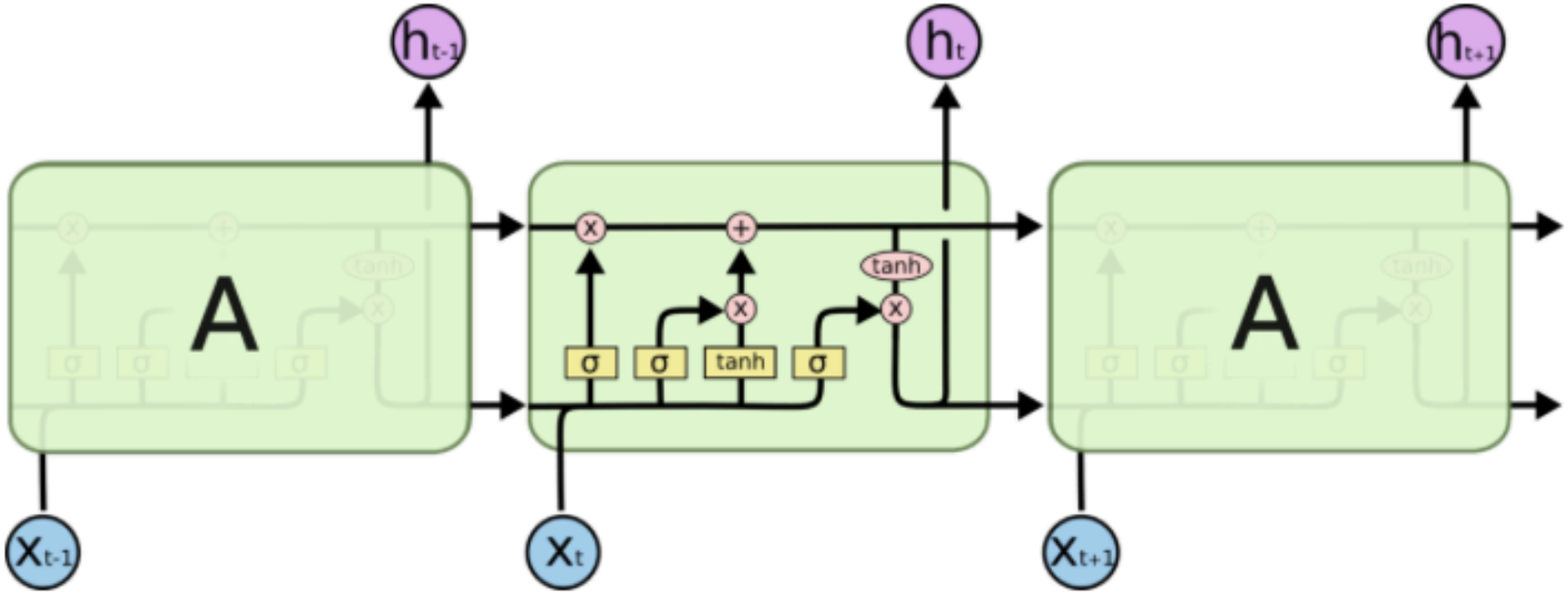
长短期记忆网络（LSTM）

- 单个RNN模块内部构造



长短期记忆网络（LSTM）

- 单个LSTM模块内部构造

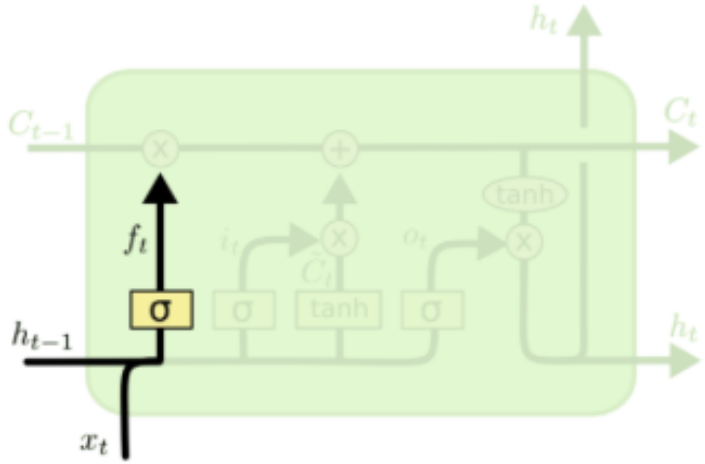


最大区别就在于它在算法中加入了一个判断信息是否有用的“处理器Cell”,增加了三个“门”。

长短期记忆网络（LSTM）

- **LSTM结构**

忘记门层

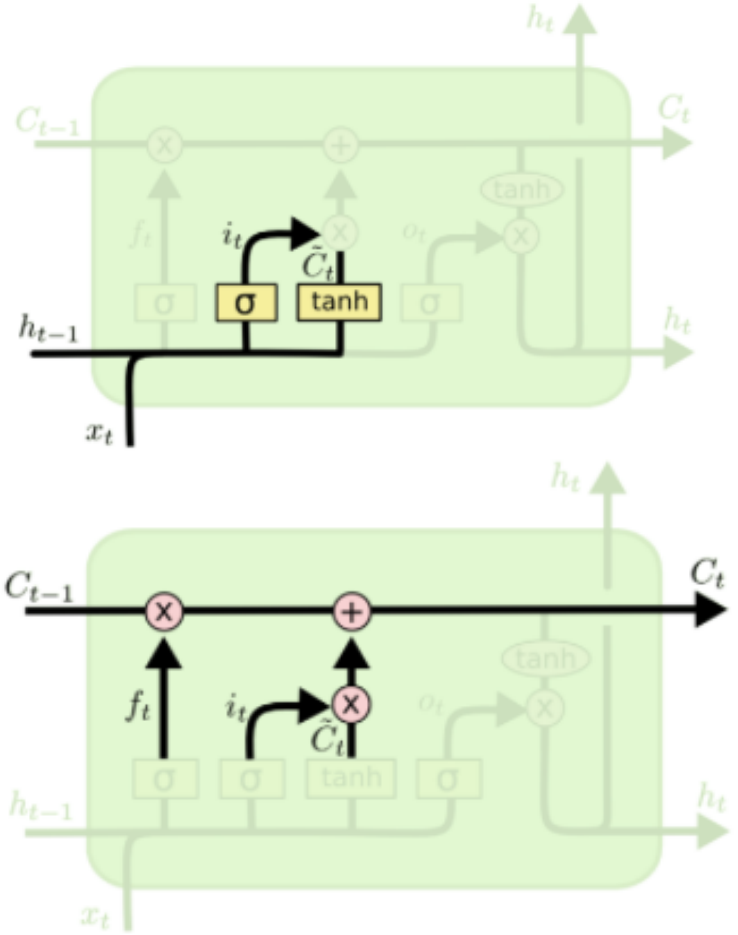


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

长短期记忆网络（LSTM）

- **LSTM结构**

  输入门层



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

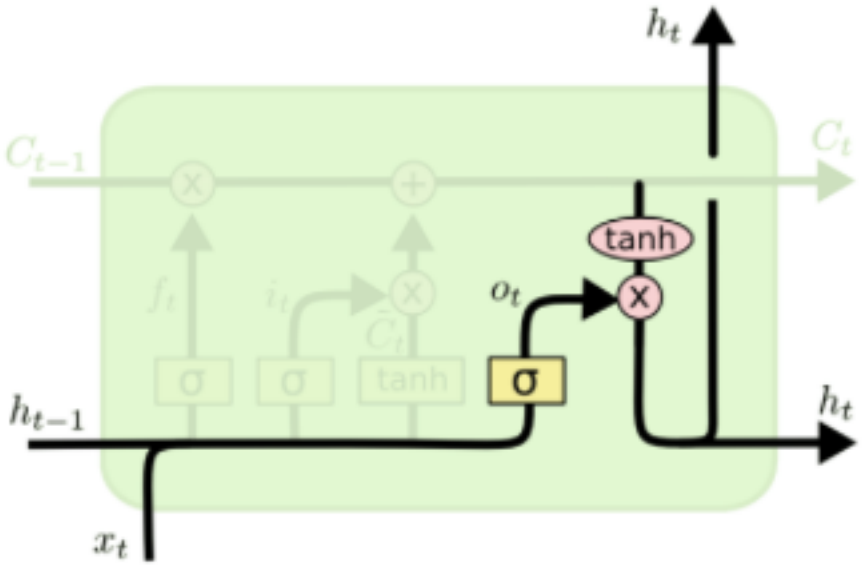
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

长短期记忆网络（LSTM）

- **LSTM结构**

  输出门层



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

时间序列 ⟶ 有监督学习

时间序列

0
1
2
3
4
5
6
7
8
9

有监督学习

X	y
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9

案例分析

广东大豆价格预测