

# 基于股价和零售的时序项目实战

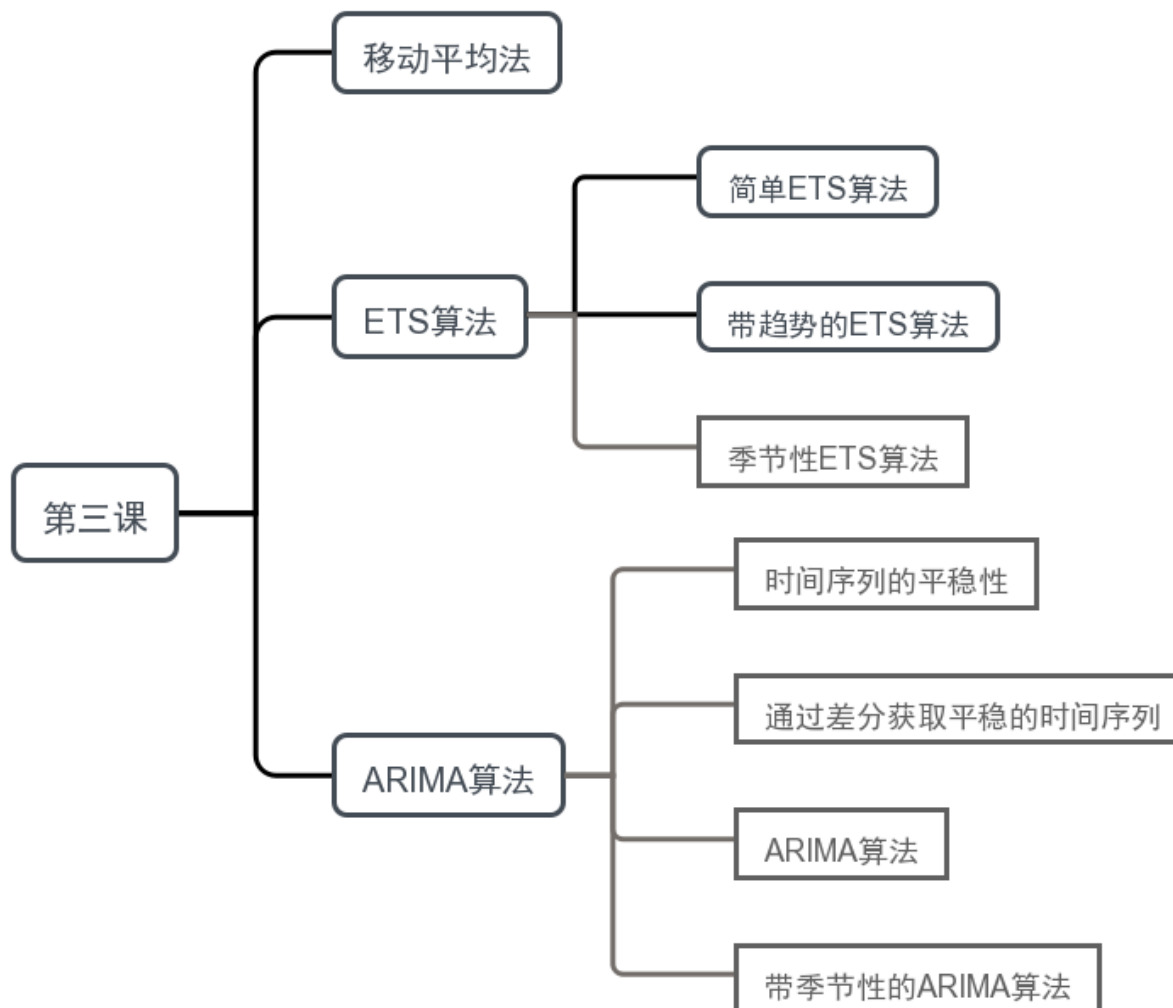
## - 第四课

导师： YY

---



# 第三课回顾



# 第四课学习目标

---

- 学会使用Prophet做时间序列预测
- 了解如何构造时间序列的机器学习样本和特征
- 学会使用LightGBM做时间序列预测
- 完成销售预测项目





# 主要内容

1/ 用Prophet算法预测时间序列

2/ 用机器学习算法预测时间序列

3/ 总结与作业



# 1.用Prophet算法预测时间序列

Predict time series with Prophet algorithms

---

- Prophet算法原理介绍
- Prophet代码实践



# Prophet算法介绍



## 时间序列预测的入门级算法

### Forecasting at Scale

Sean J. Taylor\*†

Facebook, Menlo Park, California, United States

[sjt@fb.com](mailto:sjt@fb.com)

and

Benjamin Letham†

Facebook, Menlo Park, California, United States

[bletham@fb.com](mailto:bletham@fb.com)

- **Facebook**开发的时间序列预测软件包，在高校和企业都具有**广泛的应用**
- 使用者只要具备基本的建模知识即可**快速上手**
- 模型结果有很强的**可解释性和可视化**支持
- 能为**大部分时间序列预测问题**提供一个可靠的**benchmark**结果

# Prophet模型结构



模型结构：关于时间的广义线性模型

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

将观测值 $y$ 看做是时间 $t$ （一般用整数表示）的函数，并将该函数分解成三个部分：

- $g(t)$ : **trend**, 用分段线性函数或逻辑斯蒂(logistic)函数拟合
- $s(t)$ : **seasonality**, 用傅里叶级数拟合。可以叠加多个季节性，如weekly, yearly。
- $h(t)$ : **regressor**, 用线性函数拟合。可以叠加多个外部变量，如节假日、温度、活动

备注：

- $\epsilon_t$ : 模型残差，不用拟合
- 以上方程也可以写成乘法形式：  $y(t) = g(t) * s(t) * h(t) + \epsilon_t$

# Prophet算法原理-趋势

Introduction to the Prophet algorithm

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

1

$g(t)$  趋势

2

$s(t)$ 季节性

3

$h(t)$ 外部变量的影响



# Prophet算法原理-趋势

## 线性趋势函数和分段线性趋势函数

线性趋势函数:

$$g(t) = kt + m$$

分段线性趋势函数:

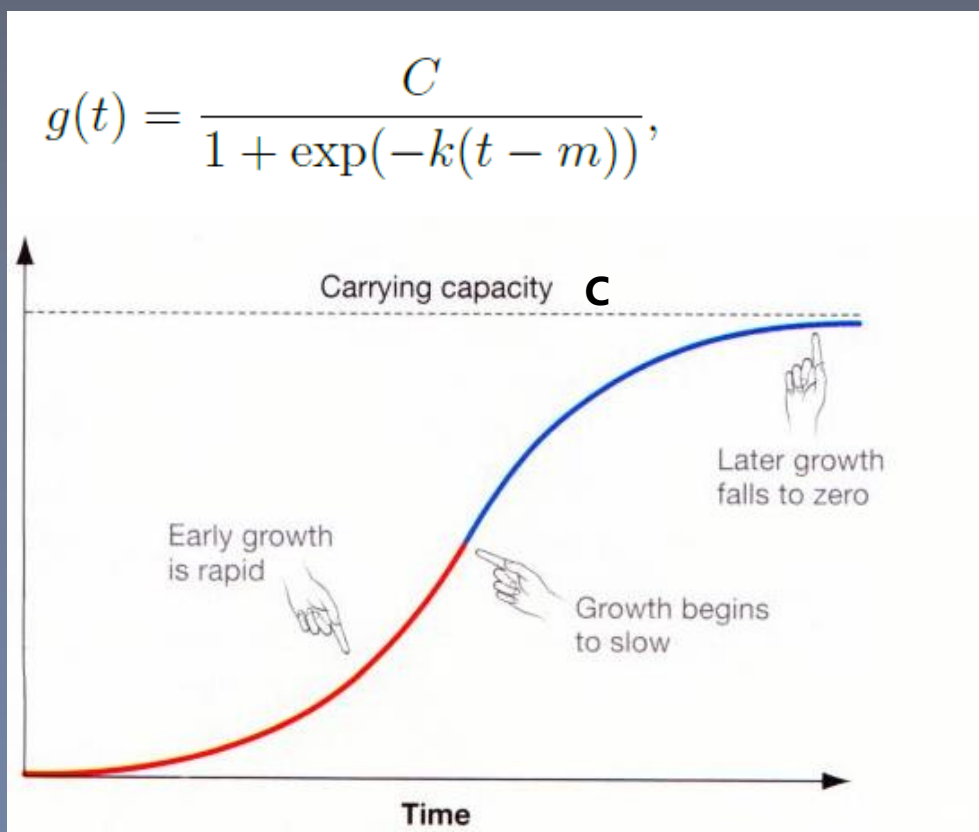
$$\begin{aligned} g(t) &= k_1 t + m_1, t \leq T_1 \\ & k_2 t + m_2, T_1 \leq t \leq T_2 \\ & k_3 t + m_3, T_2 \leq t \leq T_3 \\ & \dots \end{aligned}$$

- 以下超参数, 由用户给定
  - 最多分几段
- 以下参数, 根据历史数据拟合:
  - k: 曲线增长速率
  - m: 曲线的截距

# Prophet算法原理-趋势

## Introduction to the Prophet algorithm

### 逻辑斯蒂增长曲线和分段逻辑斯蒂增长曲线



➤ 以下超参数，由用户给定：

- C: carrying capacity, 曲线的渐近线。
- 一共分几段

➤ 以下参数从历史数据中拟合：

- k: growth rate, 增长的速率。当 $k > 0$ 时，曲线呈上升趋势；当 $k < 0$ 时，曲线呈下降趋势。
- m: 曲线拐点对应的时间，即左图中红色线和蓝色线的分界点。



# Prophet算法原理-季节性

---

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

1

$g(t)$ 趋势

2

$s(t)$ 季节性

3

$h(t)$ 外部变量的影响

# Prophet算法原理-季节性

## 用傅里叶级数拟合季节性

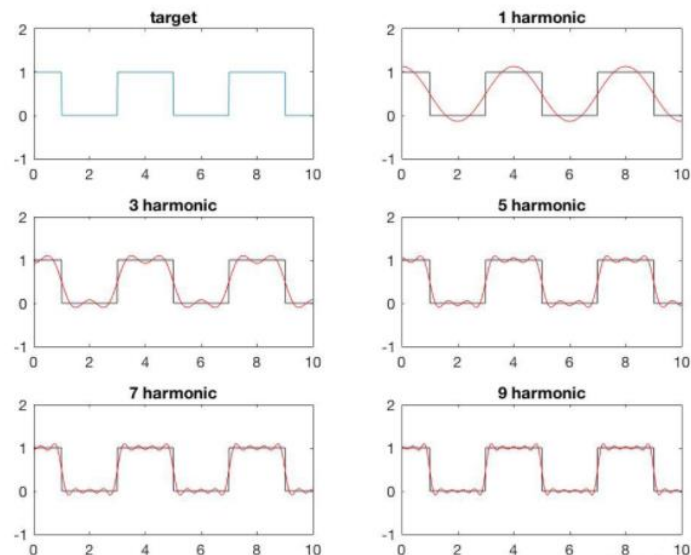
任何周期性函数都可以表示成傅里叶级数：

$$s(t) = \sum_{n=1}^N \left( a_n \cos \left( \frac{2\pi nt}{P} \right) + b_n \sin \left( \frac{2\pi nt}{P} \right) \right)$$

P: 周期长度，常见的周期有7 (weekly)、365 (yearly)

N: 傅里叶级数的阶数，N越大，季节性曲线波动越大

$a_n$ 、 $b_n$ : 系数



➤ 以下超参数，由用户给定：

- P: 周期长度。
- N: 傅里叶级数的阶数，N越大，季节性曲线波动越大，越容易过拟合。

➤ 以下参数，根据历史数据拟合：

- $a_n$ ,  $b_n$ : 傅里叶级数的系数



# Prophet算法原理-外部因素

---

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

1

$g(t)$ 趋势

2

$s(t)$ 季节性

3

$h(t)$ 外部因素的影响



# Prophet算法原理-外部因素

## Introduction to the Prophet algorithm

## 用线性函数拟合外部变量的影响

$$h(t) = Z(t)\kappa.$$

$Z(t)$ : 模型输入, 外部因素在 $t$ 时刻的取值。

$Z$ 可以是0-1变量 (e.g. 是否是法定假日, 是否是春节, 是否有促销), 也可以是连续变量 (e.g. 产品价格, 温度, 降雨量)。

$\kappa$ : 线性回归系数

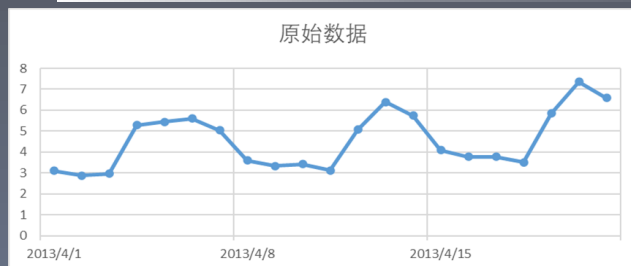
➤ 以下参数, 根据历史数据拟合:

- $\kappa$ : 线性回归系数



# Prophet算法原理-示例

## Introduction to the Prophet algorithm



ds	t	is_holiday	y(t)
2013/4/1	1	0	3.11
2013/4/2	2	0	2.882
2013/4/3	3	0	2.97
2013/4/4	4	1	5.28
2013/4/5	5	1	5.44
2013/4/6	6	1	5.6
2013/4/7	7	0	5.04
2013/4/8	8	0	3.6
2013/4/9	9	0	3.33
2013/4/10	10	0	3.42
2013/4/11	11	0	3.12
2013/4/12	12	0	5.07
2013/4/13	13	0	6.4
2013/4/14	14	0	5.74
2013/4/15	15	0	4.1
2013/4/16	16	0	3.78
2013/4/17	17	0	3.78
2013/4/18	18	0	3.52
2013/4/19	19	0	5.85
2013/4/20	20	0	7.36
2013/4/21	21	0	6.58

示例:

### Step1: 用户设置超参数

$$y(t) = g(t) + s(t) + h(t)$$

$$= kt + m + a \sin(2\pi t/7) + b \cos(2\pi t/7) + w \text{is\_holiday}(t)$$

### Step2: Prophet解参数

$$k = 0.08, m = 3.48, a = -0.91, b = 0.95, w = 1.02$$

$$g(t) = 0.08 * t + 3.48$$

$$s(t) = -0.91 * \sin(2\pi t/7) + 0.95 * \cos(2\pi t/7)$$

$$h(t) = 1.02 * \text{is\_holiday}(t)$$

# 使用Prophet算法做销量预测

---

代码演示





# 2.用机器学习算法预测时间序列

Predict time series with machine learning algorithms

---

- 如何构造训练集、验证集和测试集
- 如何构造特征
- LightGBM算法简介
- LightGBM算法实践



# 通过滑动窗的方式构造样本

## ➤ 符号定义

- $t$ : 时间点
- $y_t$ :  $t$ 时刻的观测值
- $\mathbf{x}_t$ :  $t$ 时刻的外部变量,  $\mathbf{x}_t$ 是一个向量, 长度表示外部变量的个数。当序列不带外部变量时,  $\mathbf{x}$ 不存在。
- $T$ : 基于多久的历史数据作预测
- $h$ : 预测未来多少个时间点的观测值

## ➤ 模型输入与输出

- 输入: 过去 $T$ 个时间点的观测值及外部变量, 以及未来 $h$ 个时间点的外部变量, 即

$[y_1, \mathbf{x}_1, y_2, \mathbf{x}_2, \dots, y_T, \mathbf{x}_T, \mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \dots, \mathbf{x}_{T+h}]$

- 输出: 未来 $h$ 个时间点的观测值, 即

$[y_{T+1}, y_{T+2}, \dots, y_{T+h}]$

## ➤ 用滑动窗口可以获取多个输入-输出样本

时间	观测值
2018-05	10
2018-06	11
2018-07	12
2018-08	13
2018-09	14
2018-10	15
2018-11	16
2018-12	17

以左侧的数据为例, 假设 $T=3, h=2$ , 我们可以通过滑动窗口改变历史数据的开始时间, 获取多个输入输出组合:



输入	输出
[10,11,12]	[13,14]
[11,12,13]	[14,15]
[12,13,14]	[15,16]
[13,14,15]	[16,17]

# 构建训练集、验证集、测试集

- 我们会按**输出**所处的时间将样本划分为训练集和测试集，并且训练集所处的时间一定早于测试集。

训练集      测试集

时间	观测值
2018-05	10
2018-06	11
2018-07	12
2018-08	13
2018-09	14
2018-10	15
2018-11	16
2018-12	17

以左侧的数据为例，假设 $T=3$ ,  $h=2$ , 且用2018-11之后的数据做测试：

输入	输出
[10,11,12]	[13,14]
[11,12,13]	[14,15]
[13,14,15]	[16,17]

➤ 训练集

— 测试集

注：训练集和测试集的输出所处的时间不能有交集

- 当我们需要从多个算法，或者多组超参数中进行选择时，需要将样划分为训练集、验证集和测试集，并且训练集早于验证集、测试集

训练集      验证集      测试集



# 特征构建方式

---

- 最原始特征:  $[y_1, \mathbf{x}_1, y_2, \mathbf{x}_2, \dots, y_T, \mathbf{x}_T, \mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \dots, \mathbf{x}_{T+h}]$
- 对过去某一段窗口的聚合特征:  $y$  的平均值/标准差/最大值/最小值等,  $x$  的平均值/标准差/最大值/最小值, ...
- 特定时间点的取值: 上一个时间点的  $x/y$ , 上一个周期的  $x/y$ , ...
- 时间序列的复合特征: 自相关性系数, STL 分解结果, 差分...

# LightGBM算法介绍

---

- LightGBM 是一个梯度 boosting 框架, 使用基于学习算法的决策树, 相当于XgBoost的强化版
- 由微软开发算法原理, 并提供软件包, 可以通过Python/R调用
- 官方给出的优势如下:
  - 更快的训练效率
  - 低内存使用
  - 更高的准确率
  - 支持并行化学习

# 使用LightGBM算法做销量预测

---

代码演示





# 3.总结与作业

Summary and assignments

---

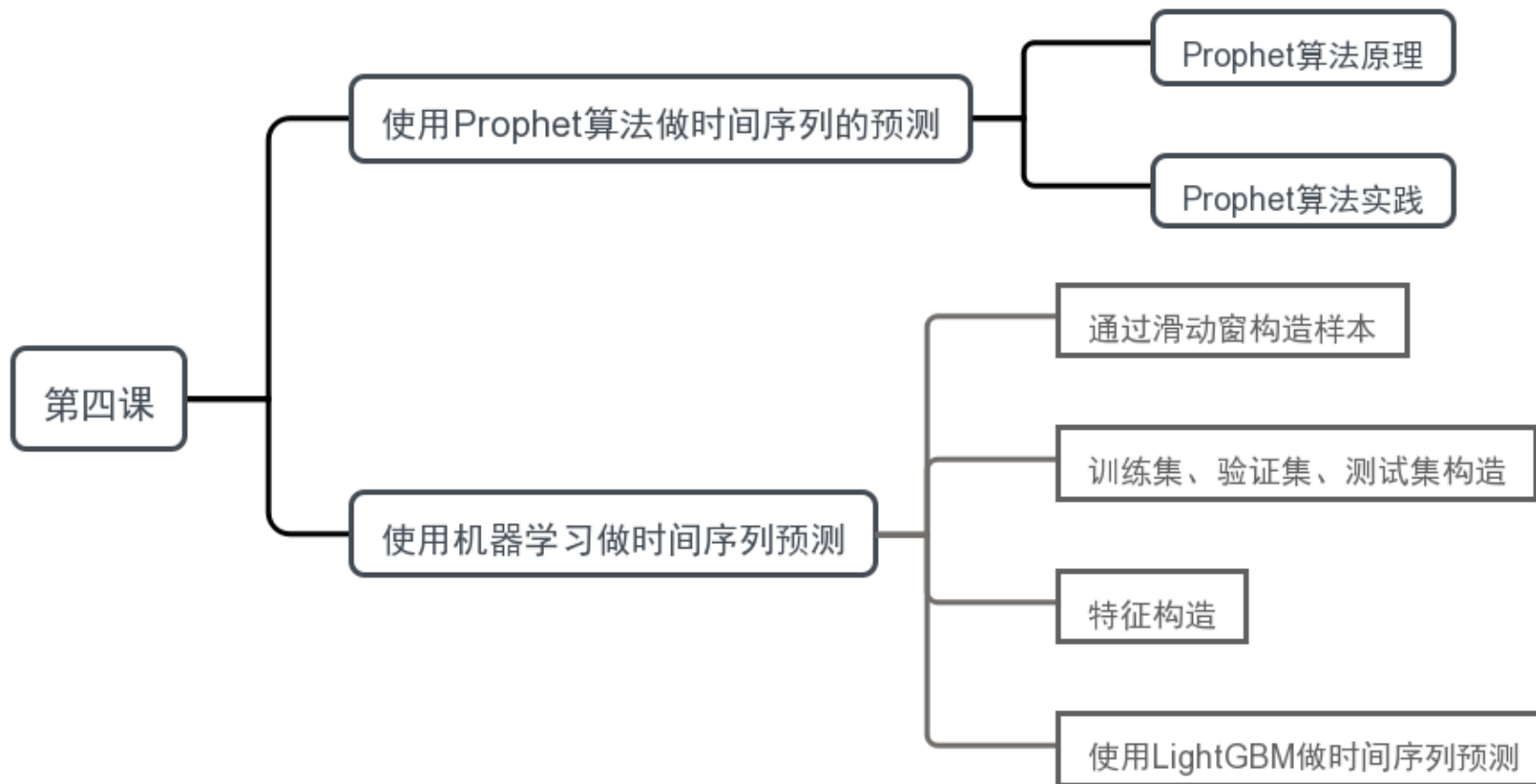
- 本课回顾
- 下节课预告
- 作业



# 本课回顾



深度之眼  
deepshare.net



# ETS/Arima/Prophet/LightGBM方法对比

	ETS	Arima	Prophet	LightGBM
是否适用于短的时间序列	是	否	否	否
是否具备可解释性	是	是	是	否
是否支持多条序列批量预测	否	否	否	是
是否支持外部变量	否	否（Arima的扩展ArimaX支持，但是很难用）	是，但仅限线性的外部变量	是

# 下节预告

## 深度学习用于时序预测导论（进阶课程）

---

- 了解基于RNN的时序预测算法
- 了解基于CNN的时序预测算法
- 理解使用深度学习算法和机器学习算法做时间序列预测的异同点

# 作业

---

分别用Prophet算法、LightGBM算法预测所有门店、所有部门预测2012-07-30及以后的销量，每次往后预测1周，并且计算预测MAPE。





深度之眼  
deepshare.net

联系我们：

电话：18001992849

邮箱：[service@deepshare.net](mailto:service@deepshare.net)

QQ：2677693114



公众号



客服微信

