时间序列模型概述

目标：

- 金融领域：GDP，各种金融/财务因子
- 互联网：广告点击、流量监测
- 市场营销：销售、促销活动
- 工业：电力负荷，生产线异常检测
- 生物医药：心电图

颗粒度：

- 每个产品
- 一组产品
- 总销量
- 按地区
- 每周、月、年

预测区间：

- 短（排班，生产运输）
- 中（资源安排，采购，招聘）
- 长（战略规划）

时序模式

- 时间序列: 序列观测值$s_t \in R$, 通常按时间排序
- 通用模型:

$$y_{t:t+T} = f(x_{t-1}, x_{t-2}, \dots, x_1, \text{温度}, \text{周天}, \text{月}, \text{价格}, \text{促销}, \dots)$$

其中:

- $x_1, \dots, x_{t-1}$ 历史观测值
- 温度，周天，月，价格，促销等为外部依赖变量。

- 趋势：长期的上升或者下降
- 季节：季节性规律，如周度、月度、年度规律
- 周期：以不固定周期震荡

- 均方误差:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

- 绝对值误差:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$

- 百分比误差:

$$\frac{1}{n}\sum_{i=1}^{n}|\frac{y_i-\hat{y}_i}{y_i}|$$

- 对称百分比误差:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|+|\hat{y}_i|}$$

- 分位数误差

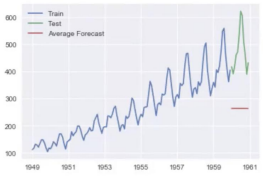
$$\frac{1}{n}\sum_{i=1}^{n}\begin{cases}\alpha|y_i-\hat{y}_i| & if\ y_i>\hat{y}_i\\(1-\alpha)|y_i-\hat{y}_i| & if y_i<\hat{y}_i\end{cases}$$

传统模型

简单平均

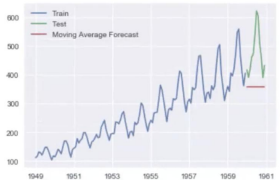
- 简单平均:

$$\hat{y}_{t+h|t}=\frac{1}{t}\sum_{i=1}^{t}y_i$$



- 滑动平均:

$$\hat{y}_{t+h|t}=\frac{1}{T}\sum_{i=t-1}^{t-T}y_i$$

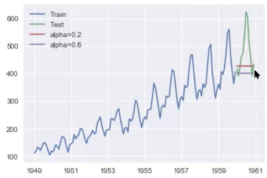


- 加权平均:
- 分量形式:

$$\hat{y}_{t+h|t}=\alpha y_t+(1-\alpha)\hat{y}_{t|t-1}$$

$$\hat{y}_{t+h|t}=l_t$$
$$l_t=\alpha y_t+(1-\alpha)l_{t-1}$$

- α 为平滑指数(衡量有效的历史数据)



Holt 线性趋势模型

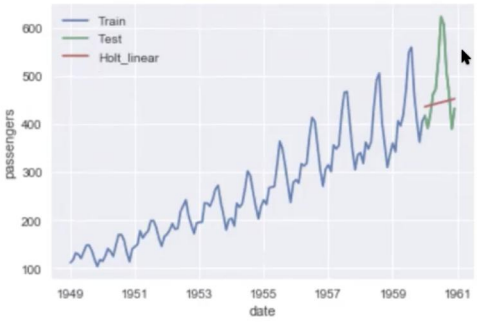
- 分量形式:

$$\hat{y}_{t+h|t}=l_t+hb_t$$
$$l_t=\alpha y_t+(1-\alpha)(l_{t-1}+b_{t-1})$$
$$b_t=\beta^*(l_t-l_{t-1})+(1-\beta^*)b_{t-1}$$

其中:

  - l_t 时刻t的序列水平
  - b_t 时刻t序列的趋势（斜率）
  - α 水平的平滑参数$(0\le\alpha\le1)$
  - β^* 趋势的平滑参数$(0\le\beta^*\le1)$

- 预测考虑了趋势的影响
- 水平和趋势都来自于指数平滑
- h 步的预测 = 最后一步的水平 + $h*$ 最后一步的趋势（斜率）



Holt-Winters 季节性预测模型

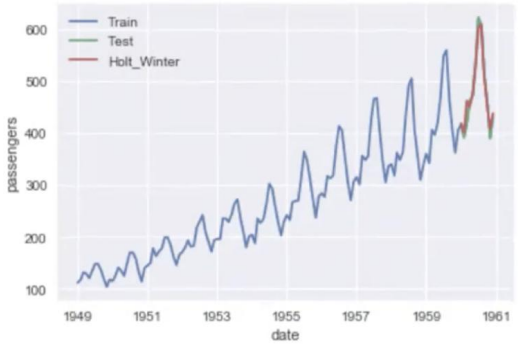
- 分量形式:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$
$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-m}$$

其中:

  - l_t 时刻t的序列水平
  - b_t 时刻t序列的趋势（斜率）
  - s_t 时刻t序列的季节性影响
  - α 水平的平滑参数$(0 \le \alpha \le 1)$
  - β^* 趋势的平滑参数$(0 \le \beta^* \le 1)$
  - γ 季节性的平滑参数$(0 \le \gamma \le 1)$
  - k 为$(h-1)/$m整数部分

- 预测为：趋势 + 季节性
- 水平，趋势，季节性均使用指数平滑（三次指数平滑）



自回归积分滑动平均模型

- ARIMA(p, d, q):

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

其中:

  - y' 为d-阶差分序列
  - p 自回归项
  - q 滑动平均项

- 平稳性和可逆性条件
- ARIMA(p, 0, 0): p-阶自回归模型
- ARIMA(0, 0, q): q-阶滑动平均模型
- 差分:
  - $d = 1$: $y'_t = y_t - y_{t-1}$
  - $d = 2$: $y'_t = y_t - 2y_{t-1} + y_{t-2}$

平稳:

- 假设: 时间序列过去存在的相关性也会在未来重现
- 如果任取两个子序列$X_{t_1}, \ldots, X_{t_n}$， $X_{t_1+k}, \ldots, X_{t_n+k}$ （$\forall n, t_1, \ldots, t_n$ 和k）他们的联合分布都一致，则称序列$X_1, \ldots, X_t$ 为严格平稳。
- 过去发生也会在未来重复!
- 序列的平均值和方差都与时间无关:

$$E[X_t] = \mu, \qquad Var[X_t] = \sigma^2$$

- 自相关系数： $\rho(k) = \frac{Cov[X_t, X_{t+k}]}{\sigma^2} = \frac{\gamma(k)}{\gamma(0)}$
- 偏自相关函数$\pi(k)$：度量X_t 和 X_{t-k} 在剔除中间k-1个变量的影响以后的相关系数

总结：传统时序模型

优点：

线性模型

只依赖于历史数据

参数较少，参数的含义明确

缺点：

很难加入其他特征（价格，产品描述，产品类别）

很难考虑序列间的关联

很难应用于多个序列（百万产品）

需要人为预缺失值处理

需要对每个序列进行统计分析来确定拟合参数

机器学习篇

拟合：

- 函数近似：

$$y = f(x) + \epsilon$$

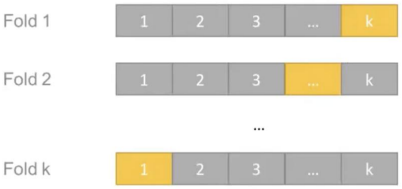
其中：

  - ϵ 为随机噪音
  - $f(\cdot)$ 确定函数（未知）

- 训练数据: $\{(x_i, y_i),\ i = 1, \ldots, n\}$
- 机器学习通过训练数据拟合函数$\hat{f}(x)$ 来逼近$f(x)$
- 参数化模型: $f(x;\ \theta)$ (线性拟合)
- 非参数化模型: $f(x; D)$ (决策时，支持向量机，k-近邻)

- Holdout:

- K-Fold:



时序模型验证：滑动验证一、二

- Move-forward:

- Move-forward:

预测区间：

迭代式分步预测

可能导致误差累积

耗时（模型复杂，预测区间长）

模型无法抓住长时间的依赖过程

特征工程是传统机器学习的基石

手工加入季节性特征

趋势

滞后特征（Xt-1）

类别编码

目标编码

模型选择：

广义线性模型

支持向量机

树模型（随机森林、梯度提升树、XGBoost，Lightgbm）

高斯过程

Case：杂货店销量预测

id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	1/1/13	25	103665	7	
1	1/1/13	25	105574	1	
2	1/1/13	25	105575	2	
3	1/1/13	25	108079	1	
4	1/1/13	25	108701	1	
5	1/1/13	25	108786	3	
6	1/1/13	25	108797	1	
7	1/1/13	25	108952	1	

store_nbr	city	state	type	cluster
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Domingo	Santo Domingo de los Tsachilas	D	4
6	Quito	Pichincha	D	13
7	Quito	Pichincha	D	8
8	Quito	Pichincha	D	8

item_nbr	family	class	perishable
96995	GROCERY I	1093	0
99197	GROCERY I	1067	0
103501	CLEANING	3008	0
103520	GROCERY I	1028	0
103665	BREAD/BAKERY	2712	1
105574	GROCERY I	1045	0
105575	GROCERY I	1045	0

- 预测指标：RMSLE，消除销量大的产品的影响

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i+1)-\log(y_i+1))^2}$$

- 训练区间: 2013-01-01 to 2017-08-15
- 测试区间: 2017-08-16 to 2017-08-31
- # 门店: 54, # 产品: 4036, # 门店 + 产品: 174685
- 节假日 + 地震

- 过去三十天的统计值：产品/门店/门店类别/产品类别
- 过去三十天的统计值：城市/省份/类别/类
- 趋势: (mean_7_days – mean_28_days)
- 对类别特征进行label encoding
- 全局均值: 周天, 促销, 月

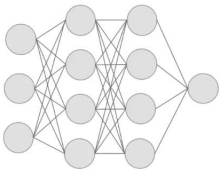
总结：机器学习

优点：很容易处理大数据/模型可解释强

缺点：需要手工特征工程/泛化能力一般/比较难学习序列间的非线性关系

深度学习篇

神经网络



- 线性拟合: $y = wx + b$
- Logistic拟合: $y = \sigma(wx + b)$

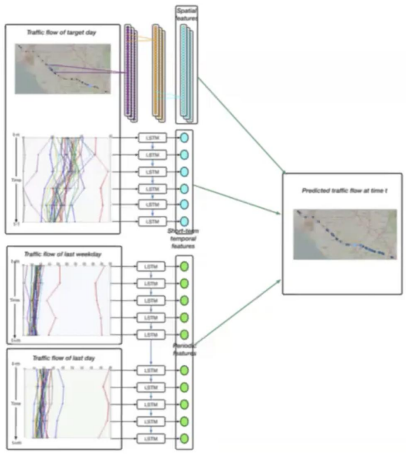
- 通用的函数逼近器
- 不需要手工特征工程
- 可以处理非结构化数据
  - 图像
  - 文本
  - 语音/视频
- 可以准确的学习复杂的非线性特征表示
- 深度学习在图像/自然语言处理领域表现卓越

递归神经网络超参

- 递归单元种类(LSTM, GRU)
- RNN层数(1, 2)
- 隐藏神经元数量(32, 64, 128)
- 激活函数(tanh, relu, selu)
- 正规化(L1, L2, activation or hidden states, dropout)
- 批归一化

卷积递归网络(CNN-LSTM): 交通流

- 交通流同时具有时间和空间关联性
- 交通流具有拓扑局域性
- 交通流具有季节性(周天，节假日，早/晚高峰)
- 一点交通延迟会扩散到周围临近区域
- 卷积提取局域路线信息(空间维度)
- 递归学习长程依赖关系(时间维度)



优点：
端到端解决方案
无需手工提取特征
支持各种类型数据
很容易学习序列非线性关系
缺点：可解释性差/超参数庞大

时序类型	深度学习	传统机器学习	传统时序模型
单一序列/少量多序列	数据量不足以训练模型	可能不适应	• 可解释性强 • 简单易上手
长序列/大量多序列	• 可以学习长程依赖 • 无需手工特征工程	• 需要手工特征工程 • 可以学习复杂关联	只在简单情况下适用(线性依赖)
大量多序列+海量其他信息	• 可以学习长程依赖 • 无需手工特征工程 • 支持所有类型数据 • 可学习非线性关联	• 需要手工特征工程 • 支持所有类型数据	• 无法学习非线性关联 • 不支持其他类型数据 • 很难应用到大量序列