时间序列预测方法汇总：从理论到实践（附Kaggle经典比赛方案）

机器学习算法那些事 2022-04-13 14:32

©作者 | Light
学校 | 中国科学院大学
研究方向 | 机器学习
知乎链接：https://zhuanlan.zhihu.com/p/471014006
编辑| paperweekly

时间序列是我最喜欢研究的一种问题，这里我列一下**时间序列最常用的方法**，包括理论和实践两部分。理论部分大多是各路神仙原创的高赞解读，这里我就简单成呈现在这里，并附上链接。实践部分是质量较高的开源代码，方便大家快速上手。最后，附上一些 kaggle 比赛中比较经典的时序比赛的经典解法链接，供大家参考和学习。

时序问题都看成是回归问题，只是回归的方式（线性回归、树模型、深度学习等）有一定的区别。

01.

传统时序建模

arima 模型是 arma 模型的升级版；arma 模型只能针对平稳数据进行建模，而 arima 模型需要先对数据进行差分，差分平稳后在进行建模。这两个模型能处理的问题还是比较简单，究其原因主要是以下两点：

- arma/arima 模型归根到底还是简单的线性模型，能表征的问题复杂程度有限；
- arma 全名是自回归滑动平均模型，它只能支持对单变量历史数据的回归，处理不了多变量的情况。

原理篇：

写给你的金融时间序列分析：基础篇
重点介绍基本的金融时间序列知识和 arma 模型

https://zhuanlan.zhihu.com/p/38320827

金融时间序列入门【完结篇】 ARCH、GARCH
介绍更为高阶的 arch 和 garch 模型

https://zhuanlan.zhihu.com/p/21962996

实践篇：

【时间序列分析】ARMA预测GDP的 python实现
arma 模型快速上手

https://zhuanlan.zhihu.com/p/54799648

machinelearningmastery.com
arch、garch模型快速建模

https://machinelearningmastery.com/develop-arch-and-garch-models-for-time-series-forecasting-in-python/

总结：如果是处理单变量的预测问题，传统时序模型可以发挥较大的优势；但是如果问题或者变量过多，那么传统时序模型就显得力不从心了。

机器学习模型方法

这类方法以 lightgbm、xgboost 为代表，一般就是把时序问题转换为监督学习，通过特征工程和机器学习方法去预测；这种模型可以解决绝大多数的复杂的时序预测模型。支持复杂的数据建模，支持多变量协同回归，支持非线性问题。

不过这种方法需要较为复杂的人工特征过程部分，特征工程需要一定的专业知识或者丰富的想象力。特征工程能力的高低往往决定了机器学习的上限，而机器学习方法只是尽可能的逼近这个上限。特征建立好之后，就可以直接套用树模型算法 lightgbm/xgboost，这两个模型是十分常见的快速成模方法，除此之外，他们还有以下特点：

- 计算速度快，模型精度高；
- 缺失值不需要处理，比较方便；
- 支持 category 变量；
- 支持特征交叉。

原理篇：

提升树模型：Lightgbm 原理深入探究：
lightgbm 原理

https://blog.csdn.net/anshuai_aw1/article/details/83659932

xgboost 的原理没你想像的那么难：

xgboost 原理

https://www.jianshu.com/p/7467e616f227

实践篇：

在 Python 中使用 Lightgbm：
lightgbm 模型实践

https://zhuanlan.zhihu.com/p/52583923

史上最详细的 XGBoost 实战：
xgboost 模型实践

https://zhuanlan.zhihu.com/p/31182879

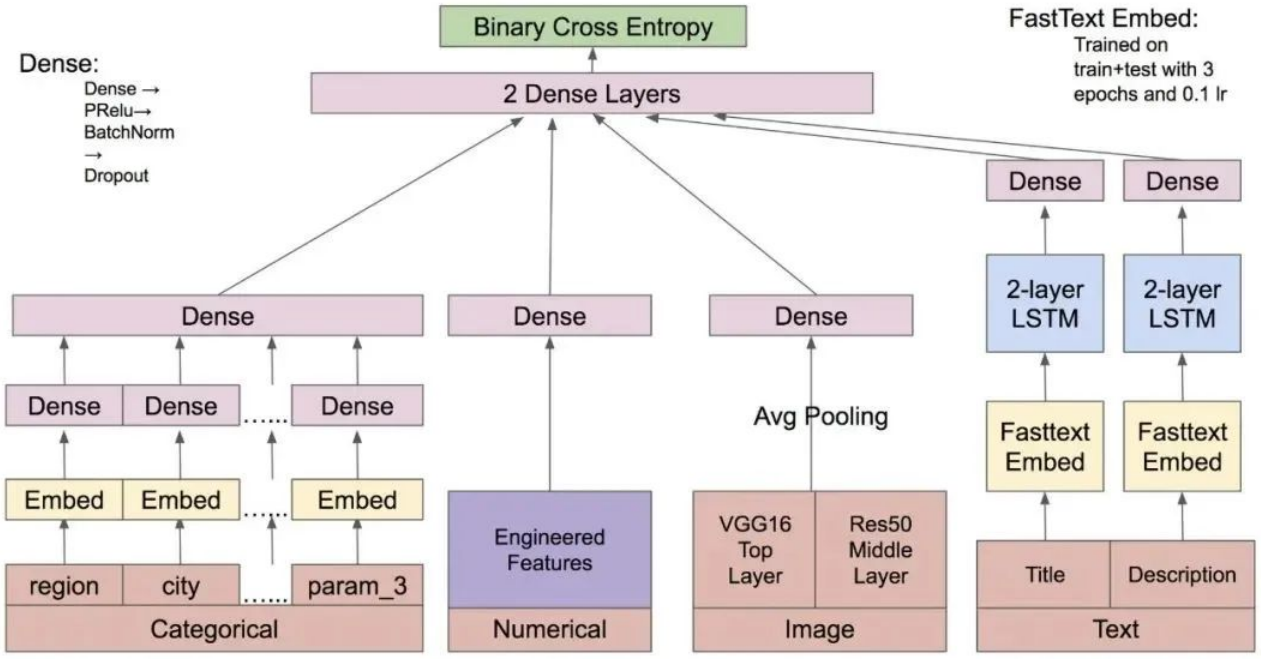
总结：通过一系列特征工程后，直接使用机器学习方法，可以解决大多数的复杂时序问题；不过这方法最大的缺点是特征工程可能会较为繁琐。

深度学习模型方法

这类方法以 LSTM/GRU、seq2seq、wavenet、1D-CNN、transformer为主。深度学习中的 LSTM/GRU 模型，就是专门为解决时间序列问题而设计的；但是 CNN 模型是本来解决图像问题的，但是经过演变和发展，也可以用来解决时间序列问题。总体来说，深度学习类模型主要有以下特点：

- 不能包括缺失值，必须要填充缺失值，否则会报错；
- 支持特征交叉，如二阶交叉，高阶交叉等；
- 需要 embedding 层处理 category 变量，可以直接学习到离散特征的语义变量，并表征其相对关系；
- 数据量小的时候，模型效果不如树方法；但是数据量巨大的时候，神经网络会有更好的表现；
- 神经网络模型支持在线训练。

实际上，基于实际预测问题，可以设计出各式各样的深度学习模型架构。假如我们预测的时序问题（如预测心跳频率），不仅仅只和统计类的数据有关，还和文本（如医师意见）以及图像（如心电图）等数据有关，我们就可以把 MLP、CNN、bert 等冗杂在一起，建立更强力的模型。



▲ 图源：https://www.kaggle.com/c/avito-demand-prediction/discussion/59880

理论篇：

[干货] 深入浅出 LSTM 及其 Python 代码实现：
LSTM 原理

https://zhuanlan.zhihu.com/p/104475016

Seq2Seq 原理详解-早起的小虫子-博客园：
seq2seq 原理

https://www.cnblogs.com/liuxiaochong/p/14399416.html

Wavenet 原理与实现：
wavenet 原理

https://zhuanlan.zhihu.com/p/28849767

CNN 卷积神经网络如何处理一维时间序列数据：
1D-CNN 处理时序数据

https://www.ai8py.com/cnn-in-keras-for-time-sequences.html

Transformer for TimeSeries 时序预测算法详解：
transformer 时序预测

https://zhuanlan.zhihu.com/p/391337035

实践篇：

seq2seq 模型的 python 实现-基于 seq2seq 模型的自然语言处理应用：
seq2seq 模型实现

https://dataxujing.github.io/seq2seqlearn/chapter3/

machinelearningmastery.com：
LSTM 实践

https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-pyth

Conv1d-WaveNet-Forecast Stock price:
wavenet 模型预测股票价格

https://www.kaggle.com/bhavinmoriya/conv1d-wavenet-forecast-stock-price

towardsdatascience.com/:
transformer 时序预测数据

https://towardsdatascience.com/how-to-use-transformer-networks-to-build-a-forecasting-model-297f9270e630

Keras documentation:
Timeseries classification with a Transformer model：transformer 处理时序数据分类

https://keras.io/examples/timeseries/timeseries_transformer_classification/

kaggle.com/fatmakursun/:
CNN 预测模型

https://www.kaggle.com/fatmakursun/predict-sales-time-series-with-cnn

总结：深度学习模型可以解决基本上所有时序问题，而且模型可以自动学习特征工程，极大减少了人工；不过需要较高的模型架构能力。

最后我再附上一些比较经典的数据挖掘比赛链接和解决方案，如果能够理解数据和代码，必会受益匪浅。如果大家对某个比赛解决方案十分感兴趣，我后续会详细解读。

1）网站流量预测：

RNN seq2seq 模型：
https://github.com/Arturus/kaggle-web-traffic

xgboost 和 MLP 模型：
https://github.com/jfpuget/Kaggle/tree/master/WebTrafficPrediction

kalman 滤波：
https://github.com/oseiskar/simdkalman

CNN 模型：
https://github.com/sjvasquez/web-traffic-forecasting

2）餐厅客户量预测

特征工程+lgb：
https://www.kaggle.com/plantsgo/solution-public-0-471-private-0-505

特征工程+lgb：
https://www.kaggle.com/pureheart/1st-place-lgb-model-public-0-470-private-0-502

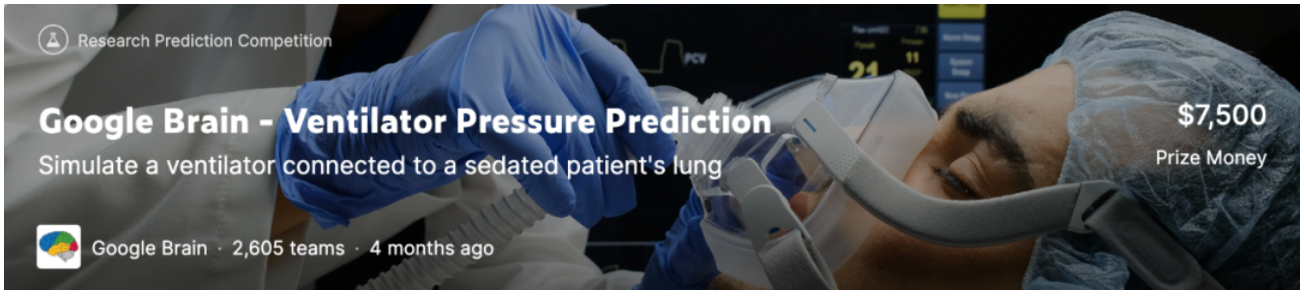
3）开放通道预测

wavenet 模型：
https://www.kaggle.com/vicensgaitan/2-wavenet-swa

1D-CNN 模型：
https://www.kaggle.com/kmat2019/u-net-1d-cnn-with-keras

seq2seq 模型：
https://www.kaggle.com/brandenkmurray/seq2seq-rnn-with-gru

4）肺压力预测



transformer 模型：
https://www.kaggle.com/cdeotte/tensorflow-transformer-0-112

双向 lstm 模型：
https://www.kaggle.com/tenffe/finetune-of-tensorflow-bidirectional-lstm

时间序列问题博大精深，应用场景十分广泛。实际上许多预测问题都可以看做是时间序列问题，比如股票/期货/外汇价格预测，网站/餐馆/旅馆/交通流量预测，店铺商品库存/销量预测等等。掌握了时间序列预测方法，你可能就掌管一把洞见未来的钥匙。

阅读原文

喜欢此内容的人还喜欢

吴恩达登录知乎，亲自回答如何系统学习机器学习

机器学习算法那些事