Содержание

1. Классификация задач предсказания
2. Проблема ML и DL подхода в задачах прогнозирования временных рядов
3. Алгоритм действий перед решением задачи
4. Где место ML, DL в домене?

Классификация задач предсказания

- По формату входных и выходных данных
- По наличию эндогенных и экзогенных данных
- По наличию тривиального и не тривиального паттерна
- По размерности ряда - многомерные и одномерные
- По количеству точек, которые нужно предсказать

Входные и выходные данные

Размерность вход и размерность выхода

Отвечаем на вопрос - сколько выходных точек по скольки входным мы будем учиться предсказывать

1) Классический подход - n входных, 1 выходная
   Предсказание на несколько точек вперед моделируется только подходом out-of-sample

2) Supervised + DL подход - n входных, m выходных
   Предсказание на несколько точек впредь моделируется как out-of-sample подходом, так и нативно

По наличию эндогенных и экзогенных данных

Эндогенные данные - зависят от других параметров системы

Экзогенные данные - внешние по отношению к системе

Работа с экзогенными данными ограничена в классическом подходе

По наличию или отсутствию тривиального паттерна

Назовем ряд с тривиальным паттерном такой ряд, который не имеющие сезонности или цикличности

Ряды с тривиальным паттерном как правило не имеет смысл предсказывать ML или DL подходом

Пример - индекс Доу-Джонса, sin(t), y(t) = const, белый шум

По размерности ряда

Сами эндогенные данные могут быть представлены несколькими измеряемыми компонентами

Пример - N датчиков одной системы, снимаемых во времени

По количеству точек предсказания

1) One step ahead
   Как правило удовлетворительные данные можно получить любой моделью, однако для эндогенных одномерных данных классические подходы немного опережают в качестве.

2) Multistep ahead
   Хорошие предсказания для эндогенных данных как правило получаются только классическими подходами

Проблема ML и DL для time series

По результатам большого количества тестов была показана неэффективность ML и DL подхода для большинства типичных задач прогнозирования*

В частности:
1) Прогнозирование одномерных эндогенных рядов на одну и несколько точек вперед.
2) Прогнозирование эндогенных многомерных рядов “из коробки”.

* https://machinelearningmastery.com/findings-comparing-classical-and-machine-learning-methods-for-time-series-forecasting/

Возможный ответ на вопрос, почему так просходит?

1) Для рядов с короткой историей как правило недостаточно данных.

   ML модель либо переобучается, либо не может выявить закономерность.

Возможный ответ на вопрос, почему так просходит?

2) Для рядов с тривиальным паттерном, как правило существует более простая модель, оптимальность которой доказана математически.

Так, например, для рядов случайного блуждания лучшим прогнозом будет наивное.

Для белого шума $y(t) \sim (\mu, \sigma^2)$, матожидание μ.

Возможный ответ на вопрос, почему так просходит?

Для всех остальных рядов:

Дело в том, что классический подход предполагает работу со стационарными рядами, для которых по теореме Вольда может быть математически найдена наилучшая модель прогнозирования.

Операция приведения к стационарному виду - это довольно сложное умение, требующее определённой интуиции эксперта, которую на данном этапе развития машинного обучения еще не представляется возможным автоматизировать.

Алгоритм действий

1) Выбрать подходящую модель
2) Выбрать метрику сравнения
3) Построить baseline модель
4) Провалидировать по метрике сравнения вашу модель с валидируемой

Выбор модели

Чтобы выбрать нужную модель, вы должны ответить на следующие вопросы

1) Какова размерность входных и выходных данных?
2) Многомерный или одномерный ряд мы предсказываем?
3) Есть ли у нас дополнительные экзогенные признаки?
4) Есть ли нетривиальный паттерн?
5) На сколько точек вперед нужно сделать предсказание?

Выбор метрики сравнения

MAE, RMSE, MAPE - подойдут любые

Однако удобно использовать MASE

Построение baseline модели

ARIMA, наивное предсказание, скользящие статистики вроде модель Хольта-Винтерса

Место ML и DL в задаче прогнозирования

1) Если при выборе модели вы получили задачу предсказания с одной точкой выходных данных, эндогенный одномерный ряд с нетривиальным паттерном, скорее всего вам нужен классический авторегрессионный подход (arima).

2) Если при выборе модели вы предсказываете ряд с тривиальным паттерном, скорее всего вам нужные простые статистические модели. Модель взвешенного среднего, наивное предсказание и т.п.

Место ML и DL в задаче прогнозирования

3) Если при выборе модели вы получили задачу предсказания с одной точкой выходных данных, ряд с нетривиальным паттерном, большим количеством внешних признаков, и предсказанием на одну точку вперед, скорее всего вам нужен переход к supervised задаче и классические ML алгоритмы.

Место ML и DL в задаче прогнозирования

4) Если при выборе модели вы получили задачу предсказания многомерного ряда с одной точкой выходных данных и возможностью ручной генерации признаков, скорее всего вам будет достаточно перехода к supervised задаче и классических ML алгоритмов.

Место ML и DL в задаче прогнозирования

CNN умеют автоматически генерировать большое количество признаков

5) Если при выборе модели вы получили задачу предсказания многомерного ряда с одной точкой выходных данных и невозможностью ручной генерации признаков, то имеет смысл попробовать CNN.

Место ML и DL в задаче прогнозирования

Нейронные сети могут принимать разный формат входных и выходных данных

6) Если при выборе модели вы получили задачу предсказания ряда с вектором выходных данных, вам стоит попробовать полносвязную сеть или RNN

Место ML и DL в задаче прогнозирования

Нейронные сети могут принимать нефиксированный формат входных и выходных данных

7) Если при выборе модели вы получили задачу предсказания ряда с нефиксированной длиной входных или выходных данных, стоит попробовать LSTM.

Место ML и DL в задаче прогнозирования

Нейронные сети могут моделировать любую нелинейную зависимость

8) Если при выборе модели вы получили задачу предсказания ряда с большим количество экзогенных признаков, у которых вы предполагаете отсутствие линейной зависимости с целевой переменной

Место ML и DL в задаче прогнозирования

Таким образом, в задаче прогнозирования временных рядов наибольшую пользу от использования DL можно получить в следующих задачах:

1) Предсказание вектора точек - полносвязная сеть
2) Сложный feature-engineering - CNN
3) Задача с нефиксированным форматов входных и выходных данных - LSTM