Инструкции

Класс моделей ARIMA недостаточно богат для наших данных: с их помощью, например, никак нельзя учесть взаимосвязи между рядами. Это можно сделать с помощью векторной авторегрессии VARIMA, но её питоновская реализация не позволяет использовать регрессионные признаки. Кроме того, авторегрессионный подход не позволяет учитывать, например, взаимодействия между сезонными компонентами. Вы могли заметить, что форма суточных сезонных профилей в будни и выходные немного разная; явно моделировать этот эффект с помощью ARIMA не получится.

Нам нужна более сложная модель. Давайте займёмся сведением задачи массового прогнозирования рядов к регрессионной постановке!

Вам понадобится много признаков. Некоторые из них у вас уже есть — это:

- идентификатор географической зоны
- дата и время
- количество поездок в периоды, предшествующие прогнозируемому
- синусы, косинусы и тренды, которые вы использовали внутри регрессионной компоненты ARIMA

Кроме того, не спешите выбрасывать построенный вами на прошлой неделе прогнозы — из них может получиться хороший признак для регрессии!

Вы можете попробовать разные регрессионный модели, но хорошие результаты, скорее всего, дадут такие, которые будут позволять признакам взаимодействовать друг с другом.

Поскольку прогноз нужен на 6 часов вперёд, проще всего будет построить 6 независимых регрессионных моделей — одна для прогнозирования $\hat{y}_{T+1|T}$, другая для $\hat{y}_{T+2|T}$ и т.д.

Чтобы сдать задание, выполните следующую последовательность действий.

1. Для каждой из шести задач прогнозирования $\hat{y}_{T+i|T}, i = 1, \ldots, 6$ сформируйте выборки. Откликом будет y_{T+i} при всевозможных значениях T, а признаки можно использовать следующие:

- идентификатор географической зоны — категориальный
- год, месяц, день месяца, день недели, час — эти признаки можно пробовать брать и категориальными, и непрерывными, можно даже и так, и так
- синусы, косинусы и тренды, которые вы использовали внутри регрессионной компоненты ARIMA
- сами значения прогнозов ARIMA $\hat{y}^{ARIMA}_{T+i|T}$
- количество поездок из рассматриваемого района в моменты времени $y_T, y_{T-1}, \ldots, y_{T-K}$ (параметр K можно подбирать; попробуйте начать, например, с 6)
- количество поездок из рассматриваемого района в моменты времени $y_{T-24}, y_{T-48}, \ldots, y_{T-24*K_d}$ (параметр K_d можно подбирать; попробуйте начать, например, с 2)
- суммарное количество поездок из рассматриваемого района за предшествующие полдня, сутки, неделю, месяц

Будьте внимательны при создании признаков — все факторы должны быть рассчитаны без использования информации из будущего: при прогнозировании $\hat{y}_{T+i|T}, i = 1, \ldots, 6$ вы можете учитывать только значения y до момента времени T включительно.

2. Разбейте каждую из шести выборок на три части:

- обучающая, на которой будут настраиваться параметры моделей — всё до апреля 2016
- тестовая, на которой вы будете подбирать значения гиперпараметров — май 2016
- итоговая, которая не будет использоваться при настройке моделей вообще — июнь 2016

3. Выберите вашу любимую регрессионную модель и настройте её на каждом из шести наборов данных, подбирая гиперпараметры на мае 2016. Желательно, чтобы модель:

- допускала попарные взаимодействия между признаками
- была устойчивой к избыточному количеству признаков (например, использовала регуляризаторы)

4. Выбранными моделями постройте для каждой географической зоны и каждого конца истории от 2016.04.30 23:00 до 2016.05.31 17:00 прогнозы на 6 часов вперёд; посчитайте в ноутбуке ошибку прогноза по следующему функционалу:

$$Q_{may} = \frac{1}{R * 739 * 6} \sum_{r=1}^{R} \sum_{T=2016.04.3023:00}^{2016.05.3117:00} \sum_{i=1}^{6} \left| \hat{y}^r_{T|T+i} - y^r_{T+i} \right|.$$

Убедитесь, что ошибка полученных прогнозов, рассчитанная согласно функционалу Q, определённому на прошлой неделе, уменьшилась по сравнению с той, которую вы получили методом индивидуального применения моделей ARIMA. Если этого не произошло, попробуйте улучшить ваши модели.

5. Итоговыми моделями постройте прогнозы для каждого конца истории от 2016.05.31 23:00 до 2016.06.30 17:00 и запишите все результаты в один файл в формате *geoID, histEndDay, histEndHour, step, y.* Здесь geoID — идентификатор зоны, histEndDay — день конца истории в формате id,y, где столбец id состоит из склеенных через подчёркивание идентификатора географической зоны, даты конца истории, часа конца истории и номера отсчёта, на который делается предсказание (1-6); столбец y — ваш прогноз.

6. Загрузите полученный файл на kaggle: https://inclass.kaggle.com/c/yellowtaxi. Добавьте в ноутбук ссылку на сабмишн.

7. Загрузите ноутбук в форму.

Review criteriaменьше

В качестве ответа в этом задании вам нужно загрузить ноутбук; убедитесь, что ход анализа, который вы провели, описан достаточно подробно для того, чтобы ваши сокурсники поняли, что вы делали и почему.