3. Прогнозирование временных рядов на карте

На прошлой неделе вы научились прогнозировать временной ряд со сложной сезонностью с помощью модели ARIMA на примере одной из географических зон. На этой неделе вы построите такие прогнозы для каждой зоны. Чтобы не подбирать вручную огромное количество моделей, вам понадобится сделать кластеризацию рядов.

Задание, оцениваемое сокурсниками: Прогнозирование большого количества рядов

Срок сдачи прошел июнь 18, 11:59 вечера PDT

Отправить сейчас

Выполненное задание необходимо отправить как можно раньше, чтобы сокурсники могли оценить вашу работу. В противном случае может не оказаться достаточного количества сокурсников для его оценки. Сдавайте работы как можно раньше!

1. **Инструкции**
2. **Моя работа**
3. **Обсуждения**

Инструкции

Процесс подбора модели ARIMA в Питоне достаточно трудоёмок, поэтому вы не сможете вручную подобрать модель для каждого из рядов в выбранных ячейках. Чтобы облегчить ручной перебор, вам предстоит кластеризовать временные ряды и подобрать гиперпараметры модели ARIMA только один раз для всех рядов каждого кластера.

Результатом этой недели будут построенные с помощью ARIMA почасовые прогнозы количества поездок для всех географических зон Нью-Йорка. Модель, которую мы строим, должна делать почасовые прогнозы для всех выбранных непустых ячеек на 6 часов вперёд. Качество модели мы будем оценивать с помощью среднего абсолютного отклонения от истинного количества поездок в июне:

$$Q_{june} = \frac{1}{R * 715 * 6} \sum_{r=1}^{R} \sum_{T=2016.05.31\ 23:00}^{2016.06.30\ 17:00} \sum_{i=1}^{6} \left| \hat{y}^r_{T|T+i} - y^r_{T+i} \right|.$$

R — количество прогнозируемых рядов, 715 — количество перебираемых концов истории.

Построенные прогнозы вам предстоит загрузить на kaggle. К сожалению, в формате kaggle сложно организовать конкурс по прогнозированию временных рядов в традиционном виде, с отложенным тестом и пересчётом лидерборда, поскольку прогнозы необходимо строить со скользящим концом истории. Но цель использования kaggle в этом проекте — не победа в конкурсе; вы всегда можете загрузить истинные данные за июнь и получить первое место. Цель в том, чтобы посмотреть, какие модели, решения и признаки использовали ваши коллеги, и понять, какие из них стоит попробовать и вам.

Чтобы сдать задание, выполните следующую последовательность действий.

1. Составьте из данных о поездках прямоугольную таблицу так, чтобы по строкам было время, а по столбцам идентификатор ячейки (возьмите только те, которые были отобраны на второй неделе). **Не используйте данные за последние имеющиеся месяцы — май и июнь 2016!**
2. Перед проведением кластеризации стандартизуйте столбцы (вычтите выборочное среднее и поделите на выборочную дисперсию). Это необходимо, поскольку при выборе модели ARIMA имеет значение только форма ряда, но не его средний уровень и размах колебаний.
3. Кластеризуйте географические зоны по значениям стандартизованных рядов. Подберите число кластеров так, чтобы оно было не слишком большим, но ряды внутри кластеров имели похожую форму. Постройте графики стандартизованных рядов каждого кластера, чтобы в этом убедиться.
4. В каждом кластере выберите наиболее типичный ряд (например, это может быть ряд, соответствующий центру кластера).

5. Для выбранных географических зон подберите на исходных рядах оптимальную структуру моделей — набор регрессионных признаков и значения гиперпараметров p,d,q,P,D,Q — так, как это делалось на прошлой неделе. **Не используйте данные за последний имеющийся месяц — май и июнь 2016!**
6. Для каждой из R географических зон настройте на данных до апреля 2016 включительно модель ARIMA с гиперпараметрами, соответствующими кластеру этой зоны. Для каждого конца истории от 2016.04.30 23:00 до 2016.05.31 17:00 постройте прогноз на 6 часов вперёд и посчитайте в ноутбуке ошибку прогноза по следующему функционалу:

$$Q_{may} = \frac{1}{R * 739 * 6} \sum_{r=1}^{R} \sum_{T=2016.04.30\ 23:00}^{2016.05.31\ 17:00} \sum_{i=1}^{6} \left| \hat{y}^r_{T|T+i} - y^r_{T+i} \right|.$$

7. Для каждой из R географических зон настройте на данных до мая 2016 включительно модель ARIMA с гиперпараметрами, соответствующими кластеру этой зоны. Для каждого конца истории от 2016.05.31 23:00 до 2016.06.30 17:00 постройте прогноз на 6 часов вперёд и запишите все прогнозы в файл 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.
8. Загрузите полученный файл на kaggle: https://inclass.kaggle.com/c/yellowtaxi. Добавьте в ноутбук ссылку на сабмишн.
9. 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.