Авторегрессия. Обучение модели для прогнозирования данных.

1 Поиск данных

С платформы kaggle были взяты данные температуры в последовательные моменты времени. Всего 6676 значений, из них 5341 точка выделена для тренировки модели, а 1335 - для тестов.

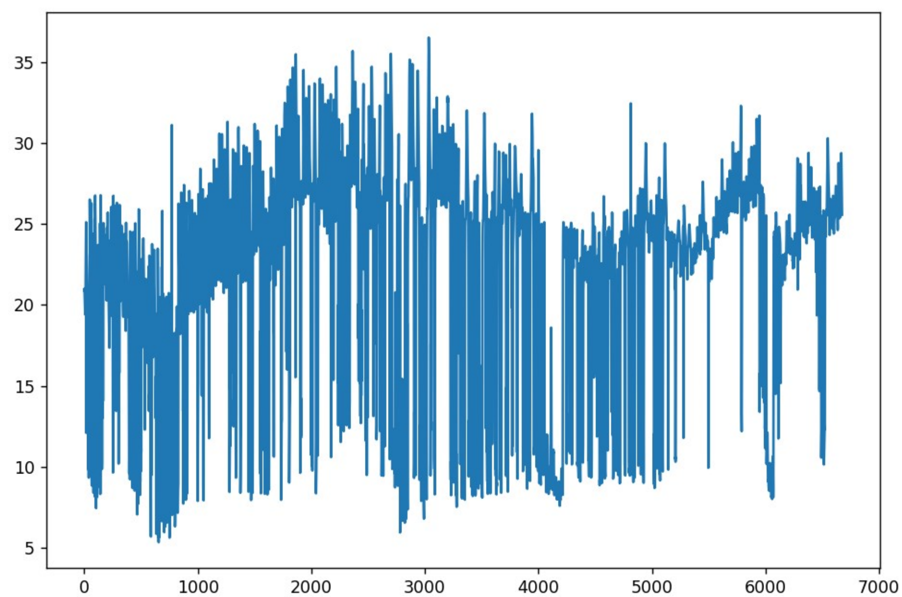

Рис. 1: Данные

2 Выбор авторегрессии

Рассмотрим авторегрессии первого и второго порядка и выберем наиболее точную.

2.1 Авторегрессия первого порядка

$x_i = a_0 + a_1 x_{i-1} + \varepsilon$
Домножим уравнение на x_{i-1} и x_{i-2} и усредним полученные выражения (по сути сложим такие уравнения для всех i и поделим на количество):

$$\begin{cases} \overline{x_i x_{i-1}} = a_0 \overline{x_{i-1}} + a_1 \overline{x_{i-1}^2} + \overline{\varepsilon x_{i-1}} \\ \overline{x_i x_{i-2}} = a_0 \overline{x_{i-2}} + a_1 \overline{x_{i-1} x_{i-2}} + \overline{\varepsilon x_{i-2}} \end{cases}$$

Заметим, что $\overline{\varepsilon x_{i-2}}$ и $\overline{\varepsilon x_{i-1}}$ можно считать равными нулю.
Обозначим:

$sr_1 = \overline{x_i} = \frac{\sum_{i=0}^{num_{train}} data_{train}[i]}{n}$

$sr_2 = \overline{x_i x_{i-1}} = \frac{\sum_{i=1}^{num_{train}} data_{train}[i] data_{train}[i-1]}{n-1}$

$sr_3 = \overline{x_i x_{i-2}} = \frac{\sum_{i=2}^{num_{train}} data_{train}[i] data_{train}[i-2]}{n-2}$

$sr_4 = \overline{x_i x_{i-3}} = \frac{\sum_{i=3}^{num_{train}} data_{train}[i] data_{train}[i-3]}{n-3}$

$sr_5 = \overline{x_i^2} = \frac{\sum_{i=0}^{num_{train}} data_{train}[i]^2}{n}$

Также будем считать, что, например, $\overline{x_i x_{i-1}} = \overline{x_{i-1} x_{i-2}}$ из-за большого количества данных. Верны и аналогичные конструкции, например, $\overline{x_{i-1}} = \overline{x_i}$.

Перепишем систему с использованием введенных обозначений:

$$\begin{cases} \mathrm{sr}_2 = a_0 sr_1 + a_1 sr_5 \\ \mathrm{sr}_3 = a_0 sr_1 + a_1 sr_2 \end{cases}$$

$a_{1_1} = \frac{sr_2 - sr_3}{sr_5 - sr_2}$

$a_{0_1} = \frac{sr_2 - a_1 sr_5}{sr_1}$

Найдем среднее квадратичное отклонение, которое получается при сравнении тестовых данных и прогнозов тестовых данных, которые дает наша модель. Прогнозы тестовых данных записываются в массив $data_{test_{y1}}$

$deviation_1 = \frac{\sum_{i=0}^{num_{test}} (data_{test}[i] - data_{test_{y1}}[i])^2}{num_{test}} \approx 6.2 \cdot 10^{202}$

2.2 Авторегрессия второго порядка

$x_i = a_0 + a_1 x_{i-1} + a_2 x_{i-2} + \varepsilon$
Домножим уравнение на x_{i-1}, x_{i-2} и x_{i-3} и усредним полученные выражения:

$$\begin{cases} \overline{x_i x_{i-1}} = a_0 \overline{x_{i-1}} + a_1 \overline{x_{i-1}^2} + a_2 \overline{x_{i-2} x_{i-1}} + \overline{\varepsilon x_{i-1}} \\ \overline{x_i x_{i-2}} = a_0 \overline{x_{i-2}} + a_1 \overline{x_{i-1} x_{i-2}} + a_2 \overline{x_{i-2}^2} + \overline{\varepsilon x_{i-2}} \\ \overline{x_i x_{i-3}} = a_0 \overline{x_{i-3}} + a_1 \overline{x_{i-1} x_{i-3}} + a_2 \overline{x_{i-2} x_{i-3}} + \overline{\varepsilon x_{i-3}} \end{cases}$$

Перепишем систему с использованием введенных обозначений:

$$\begin{cases} \mathrm{sr}_2 = a_0 sr_1 + a_1 sr_5 + a_2 sr_2 \\ \mathrm{sr}_3 = a_0 sr_1 + a_1 sr_2 + a_2 sr_5 \\ \mathrm{sr}_4 = a_0 sr_1 + a_1 sr_3 + a_2 sr_2 \end{cases}$$

$a_{1_2} = \frac{sr_2 - sr_4}{sr_5 - sr_3}$
$a_{2_2} = \frac{sr_2 - sr_3}{sr_2 - sr_5} + a_{1_2}$
$a_{0_2} = \frac{sr_2 - a_{1_2} \cdot sr_5 - a_{2_2} \cdot sr_2}{sr_1}$
Найдем среднее квадратичное отклонение, которое получается при сравнении тестовых данных и прогнозов тестовых данных, которые дает наша модель. Прогнозы тестовых данных записываются в массив $data_{test_{y2}}$

$deviation_2 = \frac{\sum_{i=0}^{num_{test}} (data_{test}[i] - data_{test_{y2}}[i])^2}{num_{test}} \approx 39.6$

$deviation_1 > deviation_2$, следовательно, выбираем авторегрессию второго порядка, так как она точнее.

3 Расчет ε

3.1 Расчет ε для данных из $data_{train}$

Посчитаем ε (результаты будем записывать в массив $epsilon_{train}$ для -того элемента из $data_{train}$ по формуле:

$epsilon_{train}[i] = a_{0_2} + a_{1_2} \cdot data_{train}[i-1] + a_{2_2} \cdot data_{train}[i-2] - data_{train}[i]$
По полученным данным построим гистограмму при помощи библиотеки matplotlib (рис. 1).
По гистограмме мы видим, что в среднем $-2 <= \varepsilon <= 2$. Значит, для прогнозируемых данных мы можем определять ε по формуле:
$\varepsilon = 4 \cdot random.random() - 2$

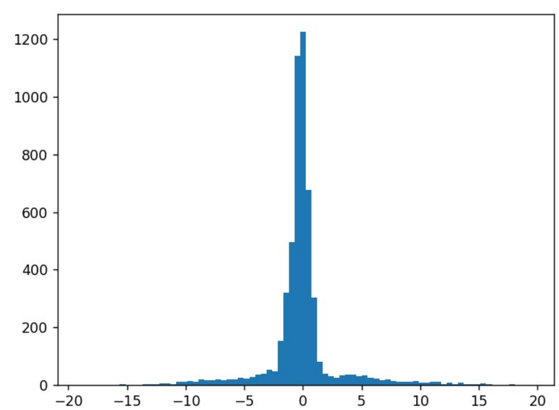

Рис. 2: Распределение значений ε

Воспользовавшись формулой авторегрессии второго порядка, в которой мы теперь умеем определять ε, заполним новый список прогнозируемых тестовых данных $data_{test_y}$ и найдем среднее квадратичное отклонение по аналогичной написанной выше формуле. Проделаем эту операцию 10 раз (для 10-ти различных наборов ε у прогнозируемых данных) и отразим на графике линиями наши результаты, а точками - данные из $data_{test}$. Также определим среднее из всех средних квадратических отклонений.

$deviation \approx 3.74$

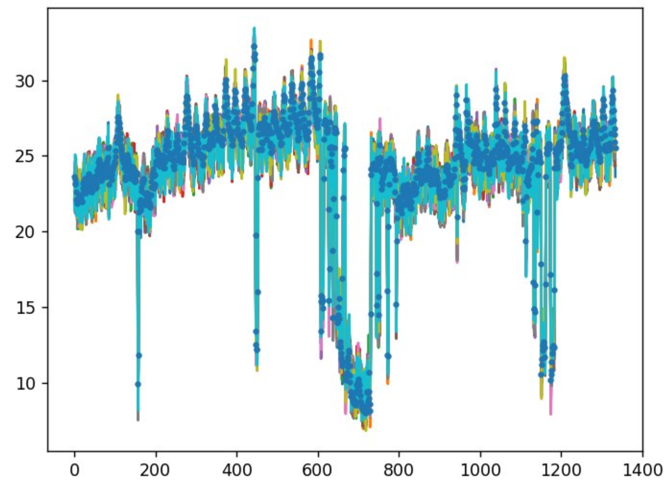

Рис. 3: Результаты обучения модели

4 Вывод

Опираясь на полученные результаты в виде довольно точного прогнозирования данных, что мы видим из графика, и небольших значений отклонений с учетом того, что это квадраты величин, мы можем сказать, что обучить модель получилось в целом хорошо, а данные она предсказывает достаточно точно.