Регрессионный анализ точек на плоскости

Баталов Семен

25.02.2021

1. Постановка задачи

Требуется написать программу, которая для некоторых данных на плоскости подбирает оптимальную регрессионную модель, и проверить корректность ее работы на известных примерах.

2. Регрессионный анализ

Регрессионный анализ – набор статистических методов исследования влияния одной или нескольких независимых переменных на зависимую переменную. Наиболее распространенный вид регрессионного анализа – линейная регрессия, когда находят линейную функцию, которая, согласно определённым математическим критериям, наиболее соответствует данным. Например, в методе наименьших квадратов вычисляется прямая (или гиперплоскость), сумма квадратов между которой и данными минимальна.

$$y(w, x) = w_0 + w_1 \cdot x \tag{1}$$

Мы будем рассматривать одномерную задачу и использовать не только линейные функции (1), но и полиномы (2), то есть будем решать задачу не только линейной, но и полиномиальной регрессии.

$$y(w, x) = w_0 + w_1 \cdot x + w_2 \cdot x^2 + \cdots + w_n \cdot x^n \tag{2}$$

Основная идея состоит в том, чтобы взять x^i за новые переменные z_i. Тем самым задача сведется к линейному (многомерному) случаю (3).

$$y(w, z) = w_0 + w_1 \cdot z_1 + w_2 \cdot z_2 + \cdots + w_n \cdot z_n \tag{3}$$

3. Инструменты

Языком разработки является «**Python**». Для работы с алгоритмом линейной регрессии используется библиотека «**sklearn.linear_model**». Для преобразования полинома в функцию многих переменных используется «**sklearn.preprocessing**». Математическое обеспечение – библиотека «**numpy**». Подробнее о программе можно узнать в папке «**source**» проекта.

4. Примеры

Будем решать задачу регрессии для набора точек, который сформирован на плоскости с помощью нормального распределения (4).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big) \tag{4}$$

Набор сотавлен так, чтобы приближать некоторую заранее известную функцию. Зная это, можно оценить корректность работы программы. Специфической особенностью программы можно считать необходимость указания стартовой степени приближающего полинома. Алгоритм устроен так, что он будет перебирать степени полиномов (начиная со стартовой) до тех пор, пока точность очередной модели не будет «сильно» отличаться от точности предыдущей. Под «точностью» подразумевается коэффициент детерминации R^2.

4.1. $y(x) = x^3$

Рассматриваем промежуток $x \in (-3, 3)$. Стартовая степень полинома $deg(p) = 2$. На рис. 4.1.1 изображен результат работы программы.

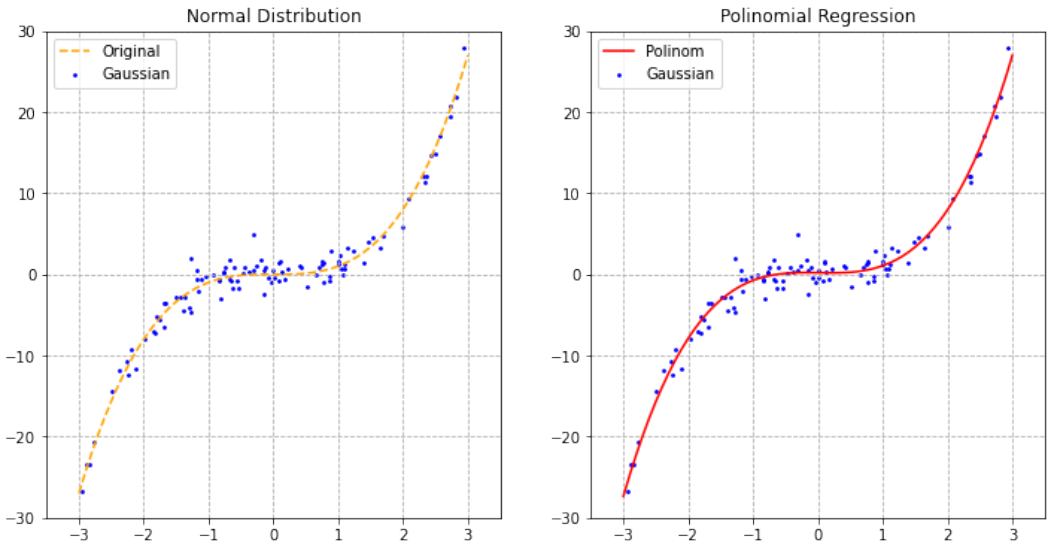

Рис. 1. Регрессионная модель для $y(x) = x^3$; $deg(p_{opt}) = 3$.

Для данной задачи степень оптимального полинома $deg(p_{opt}) = 3$. Коэффициент детерминации $R^2 = 0.976$.

4.2. $y(x) = 5 \cdot x^2$

Рассматриваем промежуток $x \in (-3, 3)$. Стартовая степень полинома $deg(p) = 2$. На рис. 4.2.2 изображен результат работы программы.

Для данной задачи степень оптимального полинома $deg(p_{opt}) = 2$. Коэффициент детерминации $R^2 = 0.988$.

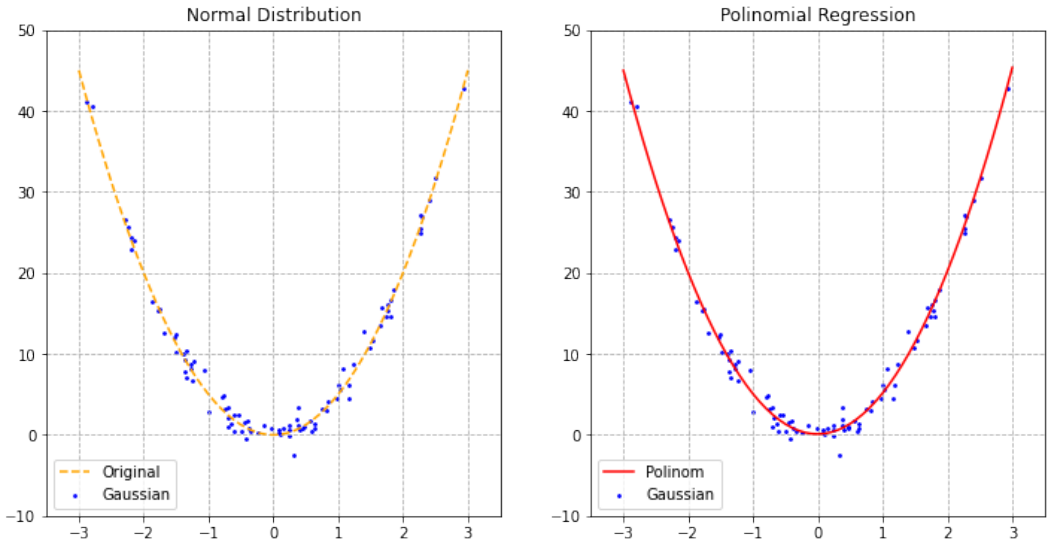

Рис. 2. Регрессионная модель для $y(x) = 5 \cdot x^2$; $deg(p_{opt}) = 2$.

4.3. $y(x) = 0.7 \cdot x + 2$

В данном случае рассмотрим частный случай полиномиальной регрессии – линейную регрессию. Рассматриваем промежуток $x \in (1, 10)$. Стартовая степень полинома $deg(p) = 1$. На рис. 4.3.3 изображен результат работы программы.

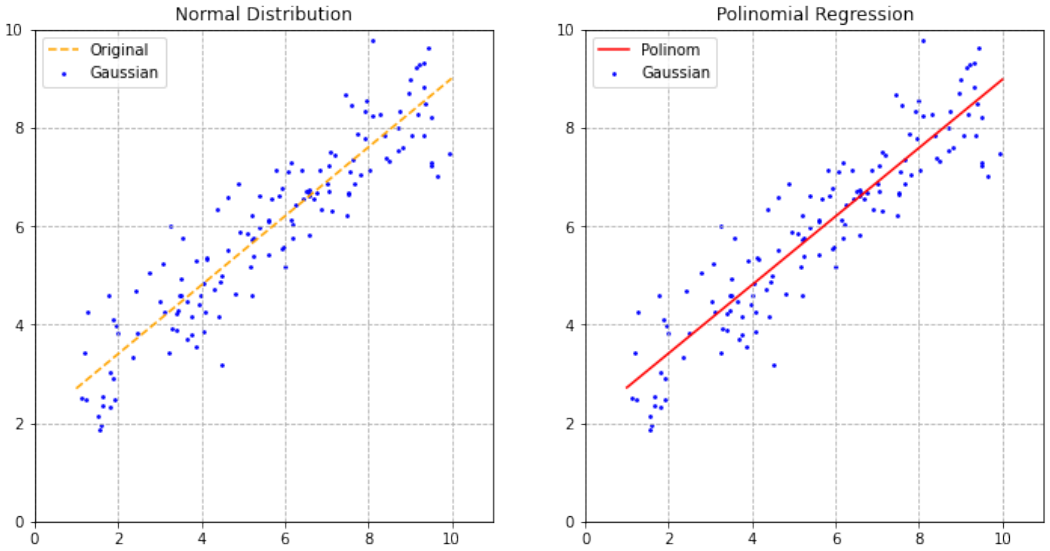

Рис. 3. Регрессионная модель для $y(x) = 0.7 \cdot x + 2$; $deg(p_{opt}) = 1$.

Для данной задачи степень оптимального полинома $deg(p_{opt}) = 1$. Коэффициент детерминации $R^2 = 0.837$.

4.4. $y(x) = sin(x)$

Рассматриваем промежуток $x \in (-3, 3)$. Стартовая степень полинома $deg(p) = 1$. На рис. 4.4.4 изображен результат работы программы.

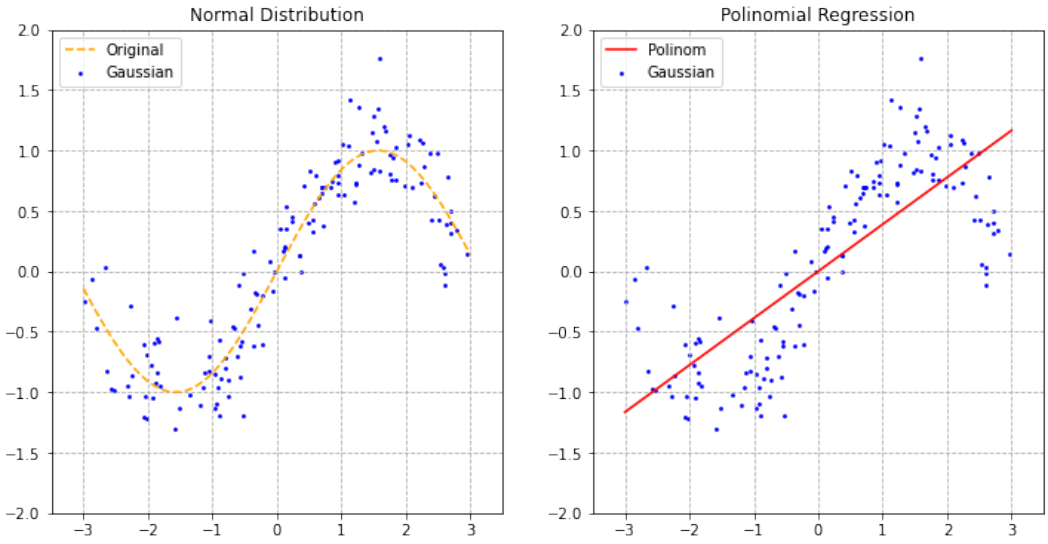

Рис. 4. Регрессионная модель для $y(x) = sin(x)$; $deg(p_{opt}) = 1$.

При установленной точности степень оптимального полинома $deg(p_{opt}) = 1$. Коэффициент детерминации $R^2 = 0.621$.

Теперь потребуем большей точности (начальные условия те же). Результат работы программы изменится (рис. 4.4.5).

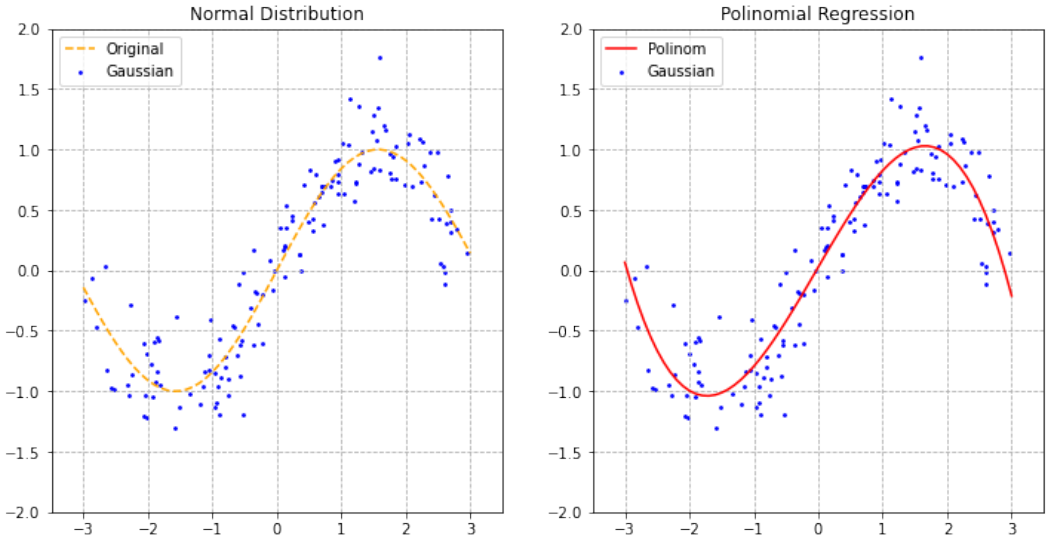

Рис. 5. Регрессионная модель для $y(x) = sin(x)$; $deg(p_{opt}) = 3$.

При установленной точности степень оптимального полинома $deg(p_{opt}) = 3$. Коэффициент детерминации $R^2 = 0.881$.

Вновь потребуем большей точности (начальные условия те же). Результат работы программы изменится (рис. 4.4.6).

При вновь установленной точности степень оптимального полинома $deg(p_{opt}) = 5$. Коэффициент детерминации $R^2 = 0.888$.

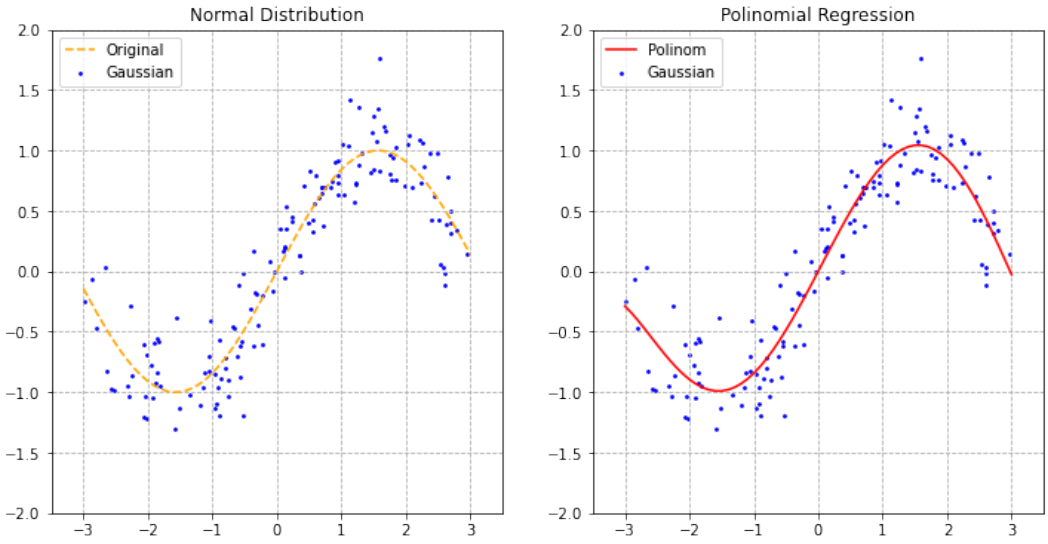

Рис. 6. Регрессионная модель для $y(x) = sin(x)$; $deg(p_{opt}) = 5$.

4.5. $y(x) = exp(x)$

Рассматриваем промежуток $x \in (-3, 5)$. Стартовая степень полинома $deg(p) = 1$. На рис. 4.5.7 изображен результат работы программы.

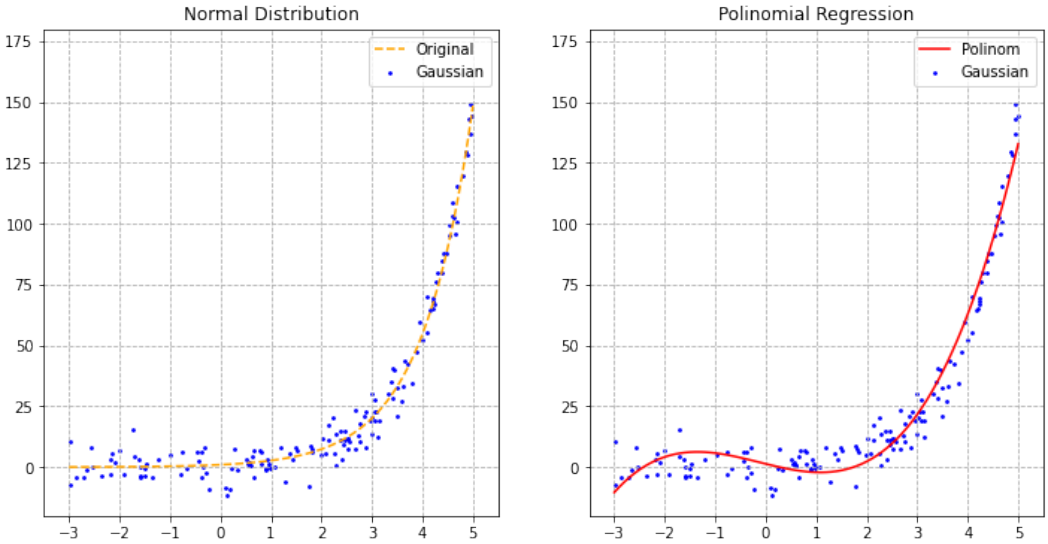

Рис. 7. Регрессионная модель для $y(x) = exp(x)$; $deg(p_{opt}) = 3$.

При установленной точности степень оптимального полинома $deg(p_{opt}) = 3$. Коэффициент детерминации $R^2 = 0.965$.

Теперь потребуем большей точности (начальные условия те же). Результат работы программы изменится (рис. 4.5.8).

При установленной точности степень оптимального полинома $deg(p_{opt}) = 4$. Коэффициент детерминации $R^2 = 0.977$.

Вновь потребуем большей точности (начальные условия те же). Результат работы программы изменится (рис. 4.5.9).

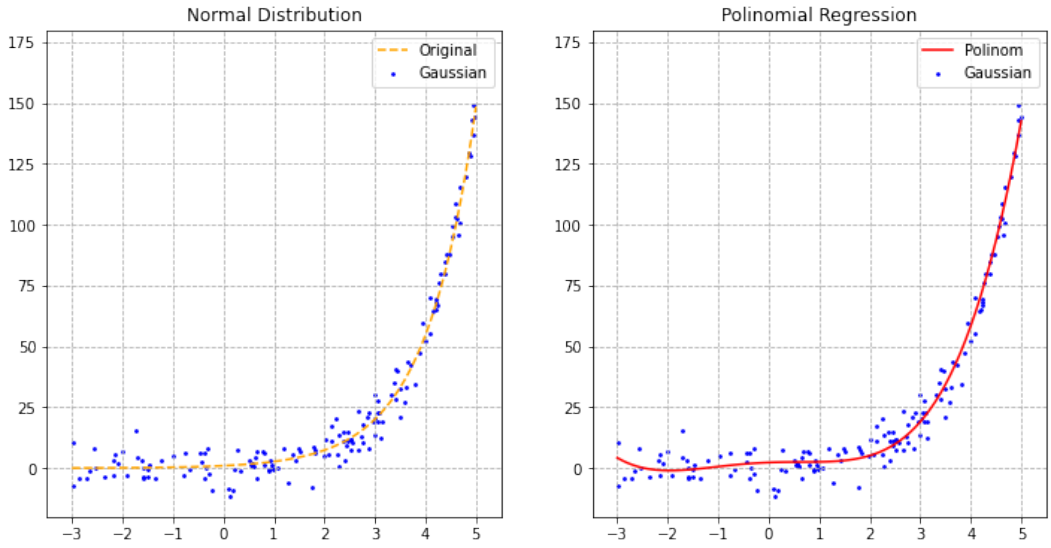

Рис. 8. Регрессионная модель для $y(x) = exp(x)$; $deg(p_{opt}) = 4$.

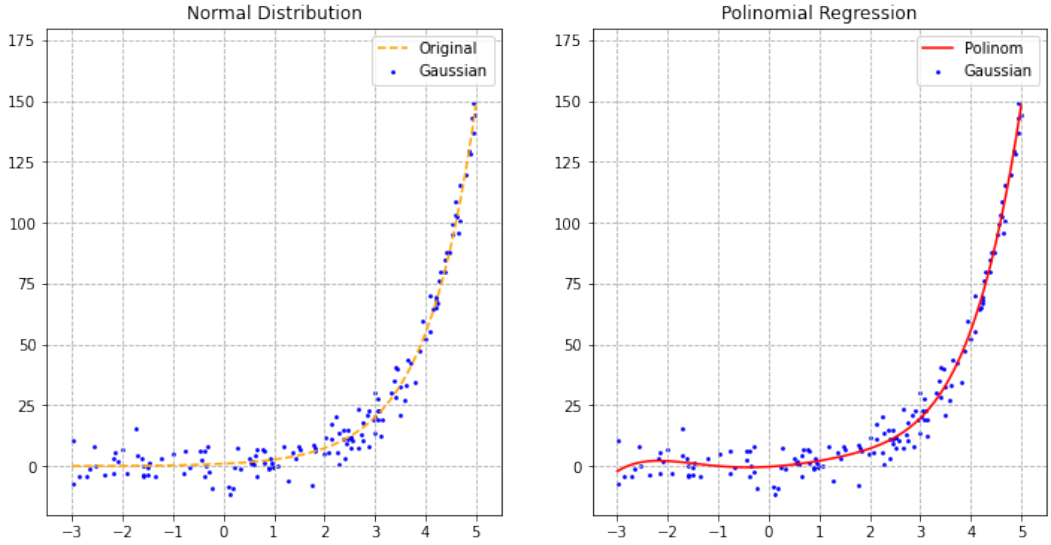

Рис. 9. Регрессионная модель для $y(x) = exp(x)$; $deg(p_{opt}) = 5$.

При вновь установленной точности степень оптимального полинома $deg(p_{opt}) = 5$. Коэффициент детерминации $R^2 = 0.981$.

5. Выводы

В итоге составили программу, которая автоматически подбирает оптимальную регрессионную (полиномиальную) модель. Проверили ее работоспособность (используя нормальное распределенние точек на плоскости) на известных примерах функций, убедились, что она работает корректно. Более того, выяснили, что, меняя входные параметры алгоритма, можно получать разные приближающие полиномы и влиять на точность регрессионной модели (коэффициент детерминации).