National Economics University, Vietnam
Faculty of Mathematics Economics
Data Science in Economics and Business
Machine Learning 1

Homework Week 4

Student: Nguyễn Anh Tú - ID: 11207333

1 Problem 1

Fit model Parabola Linear Regression cho tập dữ liệu data_square.csv:

Solution.

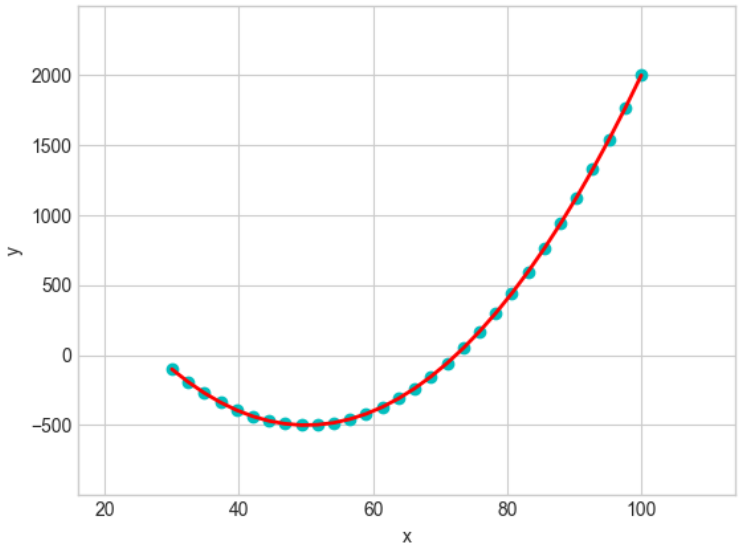
Mô hình được sử dụng là Parabola Linear Regression vì vậy giả thiết (hypothsis) được đưa ra như sau:

$$y = w_0 + w_1 x + w_2 x^2$$

Dataset cho sẵn chỉ chứa hai cột giá trị x và y nên code đã được xây dựng để thêm cột chứa giá trị của x^2. Ta có thể coi x^2 lúc này như một biến độc lập mới trong mô hình của mô hình Linear Regression. Huấn luyện mô hình này sử dụng kết quả của Normal Equation, ta được kết quả các hệ số như sau:

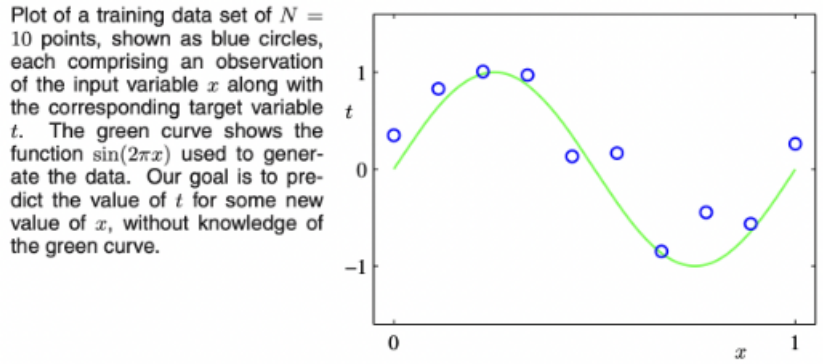
$$w_0 = 2.00000579e + 03$$
$$w_1 = 1.00000199e + 00$$
$$W_2 = -1.00000222e + 02$$

Đường dự đoán và dữ liệu được vẽ trên đồ thị scatter như sau:



2 Problem 2

Tự sinh dữ liệu như hình sau:



a. Fit dữ liệu trên với các mô hình Linear Regression của đa thức bậc 0, 1, 3, 6, 9, vẽ mô hình và nhận xét mô hình.

b. Thêm 15 và 100 điểm dữ liệu vào dữ liệu ban đầu và chạy thử mô hình đa thức bậc 9. Nhận xét mức độ overfitting của mô hình.

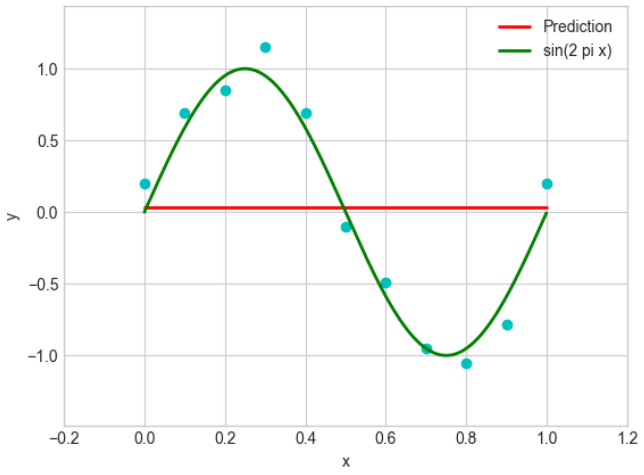
c. Fit đa thức bậc 9 cho 10 điểm dữ liệu ban đầu nhưng dùng Ridge Regression và Lasso Regression để tránh overfitting.

Solution.

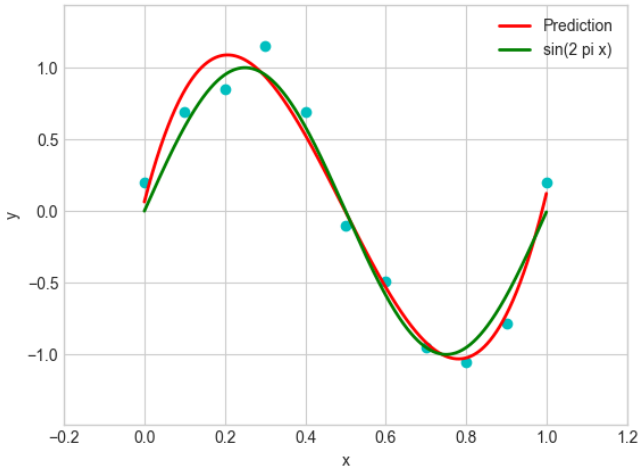
a. Các mô hình được xây dựng tương tự như mô hình bậc 2 của Problem 1, code đã được xây dựng để sinh thêm dữ liệu cho các lũy thừa của x cho những mô hình có

bậc cao. Với mô hình bậc 0, giả thiết được đưa ra là $y = \overline{Y}$, hay chính là giá trị trung bình của dữ liệu biến y.

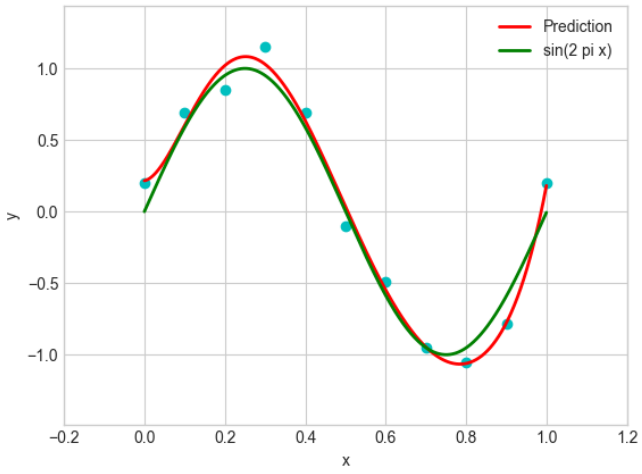
Đường dự đoán và dữ liệu được vẽ trên đồ thị scatte của từng mô hình như sau:



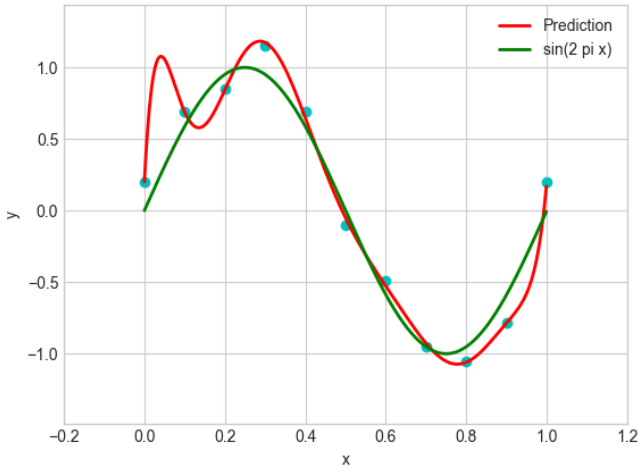
Hình 1: Mô hình bậc 0: $y = \overline{Y}$



Hình 2: Mô hình bậc 3 linear regression, n = 10



Hình 3: Mô hình bậc 6 linear regression, n = 10



Hình 4: Mô hình bậc 9 linear regression, n = 10

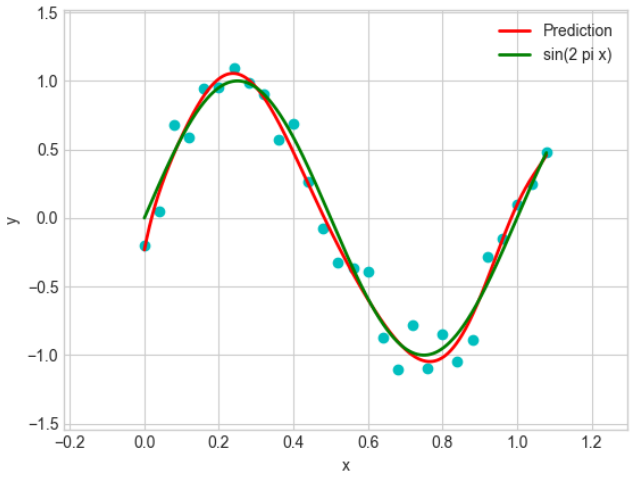
Nhận xét. Do dữ liệu về x và y được xây dựng theo hàm $y = \sin(2\pi x + \text{noise})$, ngoài đường dự đoán màu đỏ thể hiện các giá trị dự đoán của mô hình, các đồ thị còn chứa đường hồi quy chuẩn $y = \sin(2\pi x)$ màu xanh để tiện so sánh kết quả dự đoán:

- Mô hình bậc 0 thể hiện rằng y không bị phụ thuộc bởi x. Đồ thị cho thấy mô hình này không khớp với đường hồi quy chuẩn do quá đơn giản. Mô hình này gặp phải tình trạng underfitting.
- Mô hình bậc 3 và bậc 6 có đường dự đoán trùng khớp ở mức cao với đường hồi quy chuẩn. Tuy nhiên mô hình bậc 3 có sự cân bằng nhiều hơn một chút giữa

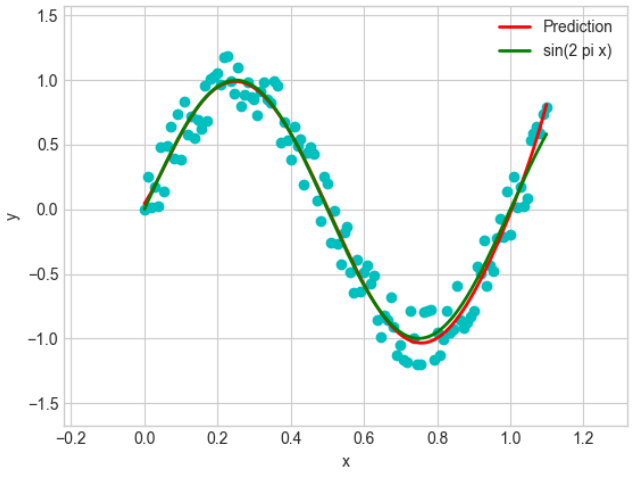
các giá trị của x, mô hình bậc 6 khớp với dữ liệu hơn nhưng đó là dấu hiệu của overfitting nhẹ.

- Mô hình bậc 9 khớp gần như hoàn toàn vớ 10 điểm dữ liệu nhưng không có sự khớp nhiều với đường hồi quy chuẩn như hai mô hình trước. Mô hình này gặp phải tình trạng overfitting.

b. Sau khi bổ sung thêm điểm dữ liệu vào mô hình bậc 9, kết quả đường dự đoán được thể hiện như sau:



Hình 5: Mô hình bậc 9 linear regression, n = 25



Hình 6: Mô hình bậc 9 linear regression, n = 110

Nhận xét. Kết quả dự đoán của mô hình bậc 9 đã được cải thiện rất nhiều khi bổ sung thêm các điểm dữ liệu vào mô hình. Đặc biệt khi bổ sung thêm 100 điểm dữ

liệu, đường dự đoán tuy không khớp nhiều với dữ liệu x nhưng có sự khớp gần như hoàn toàn với đường hồi quy chuẩn.

c. Đầu tiên cần xây dựng lý thuyết cách huấn luyện cho mô hình sử dụng Ridge Regression và Lasso Regression.

Với Ridge Regression, hàm mất mát được định nghĩa lại như sau:

$$\mathcal{L}(w) = \frac{1}{N}||xw - y||_2^2 + \alpha||w||_2^2 = \frac{1}{N}(xw - y)^T(xw - y) + \alpha w^T w$$

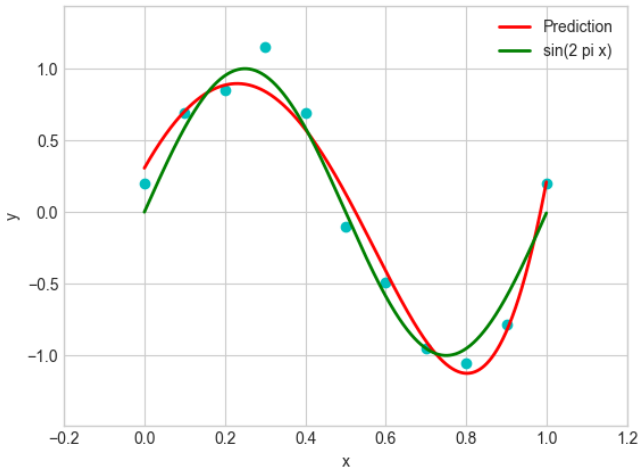
Để tìm được nghiệm tối ưu cho hàm mất mát trên, lấy đạo hàm riêng theo w của $\mathcal{L}(w)$ ta được:

$$\begin{aligned}\frac{\partial \mathcal{L}(w)}{\partial w} &= \frac{1}{N}\frac{\partial (xw - y)^T(xw - y)}{\partial w} + \alpha\frac{\partial w^T w}{\partial w} \\ &= \frac{2}{N}x^T(xw - y) + 2\alpha w \\ &= \frac{2}{N}[(x^T x + N\alpha\mathbf{I})w - x^T y]\end{aligned}$$

Nghiệm tối ưu được tìm ra khi giải phương trình:

$$\begin{aligned}\frac{\partial \mathcal{L}(w)}{\partial w} = 0 &\iff \frac{2}{N}[(x^T x + N\alpha\mathbf{I})w - x^T y] = 0 \\ &\iff (x^T x + N\alpha\mathbf{I})w = x^T y \\ &\iff w = (x^T x + N\alpha\mathbf{I})^{-1}x^T y\end{aligned}$$

Áp dụng kết quả trên để tìm vector hệ số cho mô hình, ta được đồ thị dữ liệu và đường dự đoán như sau:



Hình 7: Mô hình bậc 9 Ridge Regression, n = 10, $\alpha = 0.0001$

Hình trên cho thấy dù với bộ dữ liệu chỉ chứa 10 điểm, tình trạng overfitting của mô hình bậc 9 được cải thiện hoàn toàn.

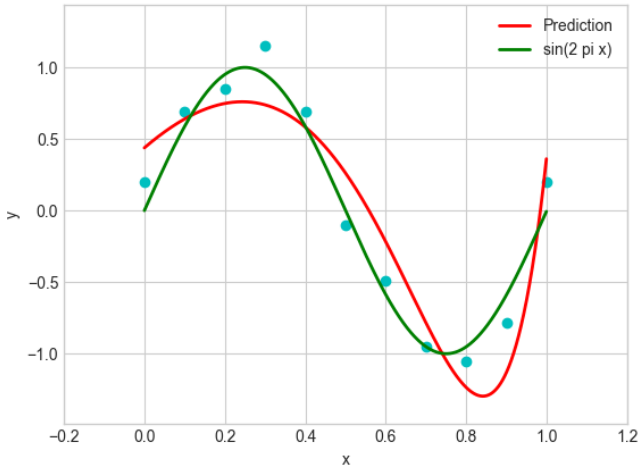
Với Lasso Regression, hàm mất mát được định nghĩa như sau:

$$\mathcal{L}(w) = \frac{1}{N}||xw - y||_2^2 + \lambda||w||_1 = \frac{1}{N}(xw - y)^T(xw - y) + \lambda\sum_i^N |w_i|$$

Hàm mất mát kể trên không có đạo hàm liên tục, vì vậy ta sẽ sử dụng phương pháp Gradient Descent để tìm nghiệm tối ưu cho hàm mất mát trên. Vector gradient của $\mathcal{L}(w)$ như sau:

$$\begin{aligned}\nabla\mathcal{L}(w) &= \frac{2}{N}x^T(xw - y) + \lambda\nabla||w||_1 \\ &= \frac{2}{N}x^T(xw - y) + \lambda\frac{\partial||w||_1}{\partial w_i}\sum_i^N |w_i| \\ &= \frac{2}{N}x^T(xw - y) + \lambda \cdot sign(w)\end{aligned}$$

Cách học của thuật toán Gradient Descent như sau: $w_{k+1} = w_k + \alpha\nabla\mathcal{L}(w)$. Vòng lặp trên sẽ kết thúc khi chuẩn vector (vector norm) của gradient vector nhỏ hơn một sai số ϵ nào đó. Đồ thị và đường dự đoán của mô hình bậc 9 sử dụng Lasso Regression như sau:



Hình 8: Mô hình bậc 9 Lasso Regression, n = 10, $\lambda = 0.045, \epsilon = 0.1, w_0 = 0$, learning rate = 0.1

Đồ thị trên cho thấy vấn đề overfitting đã được giải quyết hoàn toàn, nhưng kết quả chưa tốt bằng Ridge Regression.