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

Homework Week 3

Student: Nguyễn Anh Tú - ID: 11207333

1 Problem 1

Biến đổi để chứng minh kết quả của Normal Equation:

$$\omega = (X^T X)^{-1} X^T t$$

Solution. Giả sử trong một bài toán hồi quy tuyến tính có tập dữ liệu gồm n quan sát. Các biến độc lập được ký hiệu bởi vector $x = (x_1, x_2, ..., x_n)^T$ và các biến phụ thuộc tương ứng của chúng được ký hiệu bởi vector $t = (t_1, t_2, ..., t_n)^T$. Giả sử các điểm dữ liệu đều độc lập và có cùng phân phối (independent and identically distributed), ta có thể viết được mối quan hệ của hai biến t và x như sau:

$$t = y(x, \omega) + \epsilon$$

Với ϵ là một nhiễu ngẫu nhiên, biểu thị cho sai số của mô hình so với giá trị thực. Giả sử $\epsilon \sim \mathcal{N}(0, \beta)$, vì $t = y(x, \omega) + \epsilon$ nên $t \sim \mathcal{N}(y(x, \omega), \beta) \Rightarrow p(t) = \mathcal{N}(t|y(x, \omega), \beta)$.

Từ những giả thiết trên ta có thể xây dựng được hàm likelihood, hay xác suất của bộ dữ liệu, như sau:

$$p(t|x, \omega, \beta) = \prod^{n} \mathcal{N}(t|y(x, \omega), \beta)$$

Để thuận tiện cho việc tính toán maximum likelihood, thay vào đó ta sẽ tối ưu hàm log-likelihood:

$$\begin{aligned}
\log p(t|x, \omega, \beta) &= \sum^{n} \log(\mathcal{N}(t|y(x, \omega), \beta)) \\
&= \sum^{n} \log \int_{-\infty}^{\infty} \frac{1}{\beta^{1/2}\sqrt{2\pi}} \exp\left(-\frac{(y(x, \omega) - t)^2}{2\beta}\right) \\
&= n \log \frac{1}{\beta^{1/2}\sqrt{2\pi}} - \frac{1}{2\beta} \sum^{n} (y(x, \omega) - t)^2
\end{aligned}$$

Lưu ý rằng β trong bài toán này được xét như một hằng số nên để tìm cực đại của hàm log-likelihood ta cần tìm cực tiểu của $\sum(y(x,\omega)-t)^2$.

Gọi $P=\sum^n(y(x,\omega)-t)^2$ trong đó $y(x,\omega)=\omega_1 x+\omega_0$. Giả sử các vector x,t,ω được ký hiệu như sau:

$$x=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix};\qquad t=\begin{bmatrix}t_1\\ t_2\\ \vdots\\ t_n\end{bmatrix};\qquad \omega=\begin{bmatrix}w_0\\ w_1\end{bmatrix}$$

Vậy:

$$y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}=\begin{bmatrix}w_1x_1+w_0\\ w_2x_2+w_0\\ \vdots\\ w_nx_n+w_0\end{bmatrix}=x\cdot\omega$$

$$t-y=\begin{bmatrix}t_1-y_1\\ t_2-y_2\\ \vdots\\ t_n-y_n\end{bmatrix}$$

$$\begin{aligned}\implies \|t-y\|_2^2&=(t_1-y_1)^2+\cdots+(t_n-y_n)^2\\ &=\sum^n(t_i-y_i)^2=P\end{aligned}$$

$$\implies P=\|t-y\|_2^2=\|t-x\omega\|_2^2=(x\omega-t)^T(x\omega-t)$$

Lấy đạo hàm riêng theo ω của P ta được:

$$\begin{aligned}\frac{\partial(P)}{\partial(\omega)}&=2x^T(t-x\omega)=0\\ &\iff x^t=x^Tx\omega\\ &\iff \omega=(x^Tx)^{-1}x^Tt\end{aligned}$$

2 Problem 2

Viết code numpy, tìm model linear regression cho bài toán dữ đoán giá nhà với dataset data_linear.csv. Sau đó thực hiện các yêu cầu sau:

a. Vẽ model dự đoán (đường thẳng) và dữ liệu (point - scatter).

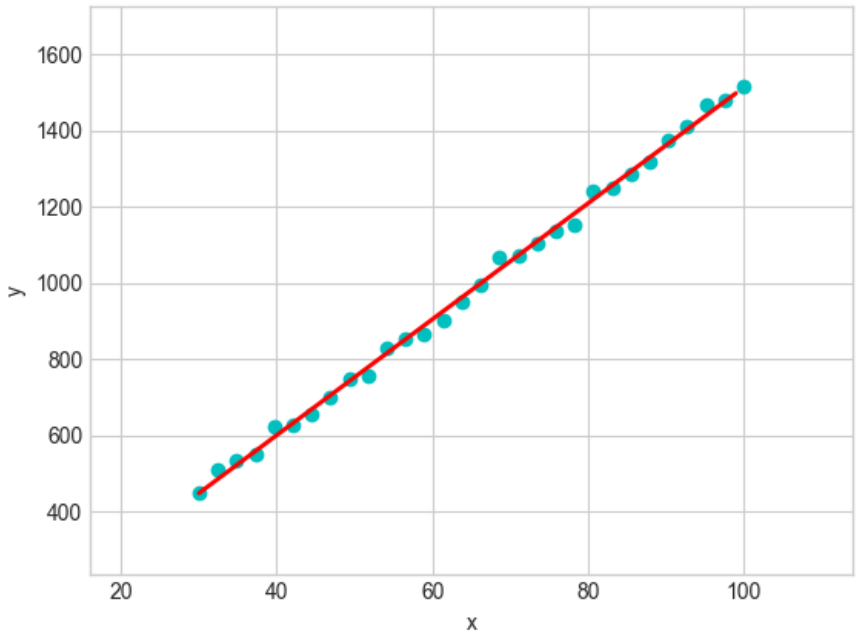
b. Dự đoán giá các căn nhà có diện tích 50, 100, 150.

Solution.

a. Sử dụng kết quả của Problem 1 để viết code tìm hệ số ω cho mô hình hồi quy tuyến tính sử dụng dataset trên, tìm được kết quả ω như sau:

$$\omega_0 = -7.0642686452452494$$
$$\omega_1 = 15.211090799670416$$

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



b. Kết quả dự đoán giá các căn nhà có diện tích 50, 100, 150 như sau:

Diện tích	Giá tiền
50	753.49027
100	1514.04481
150	2274.59935

3 Problem 3

Viết code numpy, tìm model linear regression cho bai toán dữ đoán giá nhà, dataset housing.csv

Trong source code đã sử dụng phương pháp Newton-Raphson để tìm cực tiểu cho hàm mất mát (cost function), qua đó tìm được các tham số cho mô hình hồi quy tuyến

tính cho dataset trên. Lưu ý rằng trong mô hình $y = \theta_0 + \theta_1 x$ có cost function được biểu diễn như sau:

$$J(\theta) = \frac{1}{n}\sum_{i}^{n}(x_i\theta - y_i)^2$$

Gradient vector của $J(\theta)$ là:

$$\begin{aligned}\nabla J(\theta) = \frac{\partial}{\partial\theta}J(\theta) &= \frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i}^{n}(x_i\theta - y_i)^2 \\ &= \frac{2}{m}(x_i\theta - y_i)\frac{\partial}{\partial\theta}(x_i\theta - y_i) \\ &= \frac{2}{m}(x_i\theta - y_i)x_i\end{aligned}$$

Ma trận Hessian của $J(\theta)$ là:

$$\mathbf{H}(J)(\theta) = \frac{\partial^2 J}{\partial\theta_i\partial\theta_j} = \frac{\partial}{\partial\theta}\frac{2}{m}(x_i\theta - y_i)x_i = \sum_{i}^{n}(x_i)^2 = X^T X$$

Cuối cùng, learning rule của phương pháp Newton-Raphson được quy định như sau (lặp lại cho đến khi learning step đủ nhỏ) :

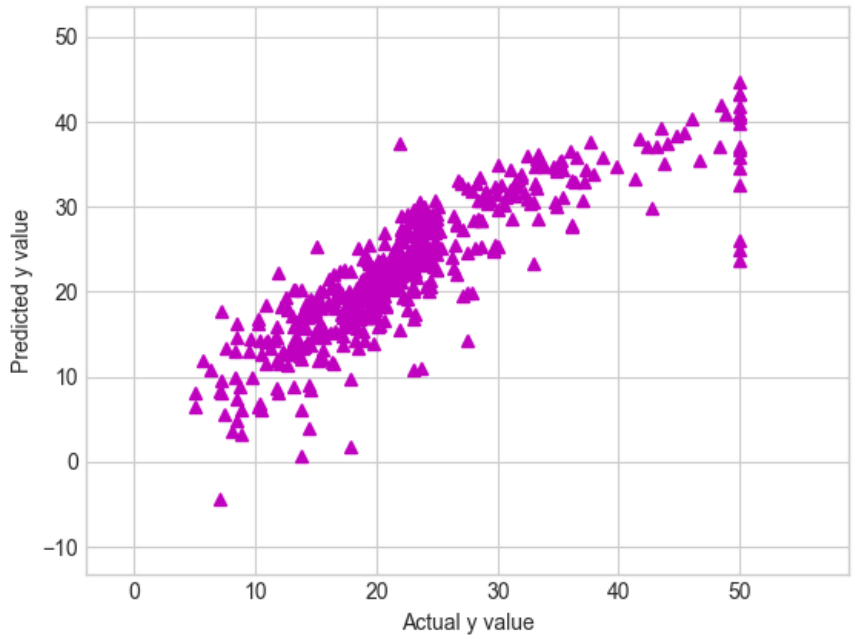
$$\theta_{i+1} = \theta_i \pm \alpha * \mathbf{H}(J)^{-1} \cdot \nabla J$$

Trong đó α là một tham số biểu thị learning rate tùy chọn.

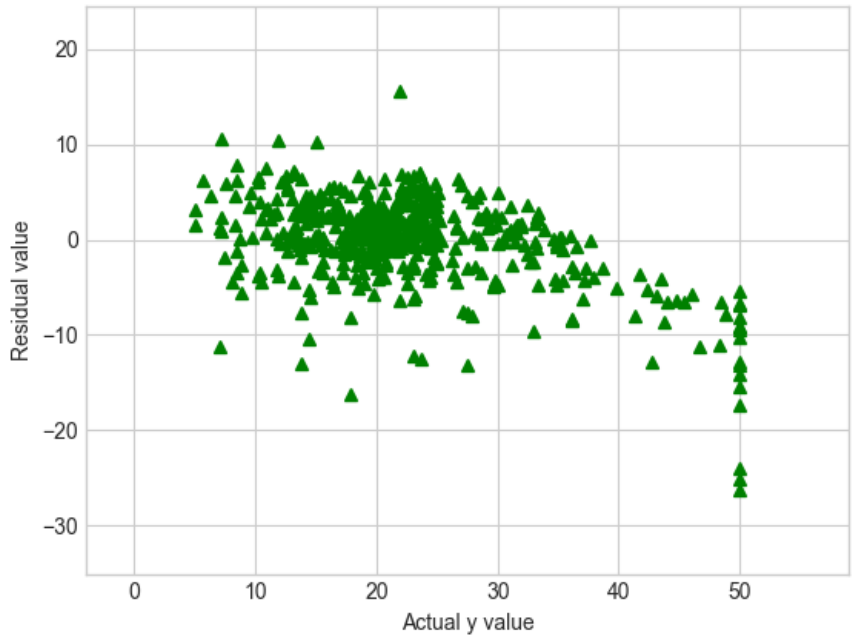
Sau khi áp dụng phương pháp trên, bộ tham số tìm được cho mô hình hồi quy tuyến tính của dataset này là:

Biến	Hệ số
θ_0	3.646e+01
CRIM	-1.08e-01
ZN	4.642e-02
INDUS	2.056e-02
CHAS	2.687e+00
NOX	-1.777e+01
RM	3.81e+00
AGE	6.922e-04
DIS	-1.476e+00
RAD	3.060e-01
TAX	-1.233e-02
PTRATIO	-9.527e-01
B	9.312e-03
LSTAT	-5.248e-01

Đồ thị biểu thị sự khác nhau giữa giá trị thực tế và giá trị dự đoán của mô hình như sau:



Đồ thị biểu thị residual với các giá trị y như sau:



4 Problem 4

Chứng minh rằng với ma trận X thì $X^T X$ khả nghịch khi X full rank.

Solution. X là full rank, ta cần chứng minh X độc lập tuyến tính.

$$\begin{aligned}
&\Rightarrow \vec{v}^T X^T X \vec{v} = \vec{v}^T \overrightarrow{0} \\
&\Rightarrow (X\vec{v})^T X\vec{v} = 0 \\
&\Rightarrow (X\vec{v}) \cdot (X\vec{v}) = 0 \\
&\Rightarrow X\vec{v} = \overrightarrow{0}
\end{aligned}$$

Ta có: nếu $\vec{v} \in N\left(X^T X\right)$:

$$\begin{aligned}
&\Rightarrow \vec{v} \in N(X) \\
&\Rightarrow \vec{v} \text{ chỉ có thể là } \overrightarrow{0} \\
&\Rightarrow N\left(X^T X\right) = N(X) = \{\overrightarrow{0}\}
\end{aligned}$$

$\Rightarrow X^T X$ là độc lập tuyến tính; mà $X^T X$ là ma trận vuông $\Rightarrow X^T X$ khả nghịch.