MÔ HÌNH HỒI QUY

Tiết 1: Mô hình hồi quy tuyến tính

Ths Phạm Việt Anh

Viện Công nghệ HaUI
Trường Đại học Công nghiệp Hà Nội

Ngày 21 tháng 3 năm 2024

Nội dung

Nội dung

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Trả lời: Là phân nhánh của Khai phá dữ liệu quan tâm đến việc phát triển các thuật toán học cho máy tính nhằm giúp máy tính có thể trích rút ra thông tin và tri thức từ dữ liệu thông qua các phương pháp "học".

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Trả lời: Là phân nhánh của Khai phá dữ liệu quan tâm đến việc phát triển các thuật toán học cho máy tính nhằm giúp máy tính có thể trích rút ra thông tin và tri thức từ dữ liệu thông qua các phương pháp "học".

- Các bài toán cơ bản trong học máy là gì?

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Trả lời: Là phân nhánh của Khai phá dữ liệu quan tâm đến việc phát triển các thuật toán học cho máy tính nhằm giúp máy tính có thể trích rút ra thông tin và tri thức từ dữ liệu thông qua các phương pháp "học".

- Các bài toán cơ bản trong học máy là gì?

Trả lời: Hai bài toán cơ bản trong học máy là học giám sát và học không giám sát.

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Trả lời: Là phân nhánh của Khai phá dữ liệu quan tâm đến việc phát triển các thuật toán học cho máy tính nhằm giúp máy tính có thể trích rút ra thông tin và tri thức từ dữ liệu thông qua các phương pháp "học".

- Các bài toán cơ bản trong học máy là gì?

Trả lời: Hai bài toán cơ bản trong học máy là học giám sát và học không giám sát.

- Các em hãy truy cập và đường link sau và trả lời các câu hỏi dưới dạng trắc ngiệm

Kiểm tra kiến thức bài cũ

- Các em hãy cho biết học máy là gì?

Trả lời: Là phân nhánh của Khai phá dữ liệu quan tâm đến việc phát triển các thuật toán học cho máy tính nhằm giúp máy tính có thể trích rút ra thông tin và tri thức từ dữ liệu thông qua các phương pháp "học".

- Các bài toán cơ bản trong học máy là gì?

Trả lời: Hai bài toán cơ bản trong học máy là học giám sát và học không giám sát.

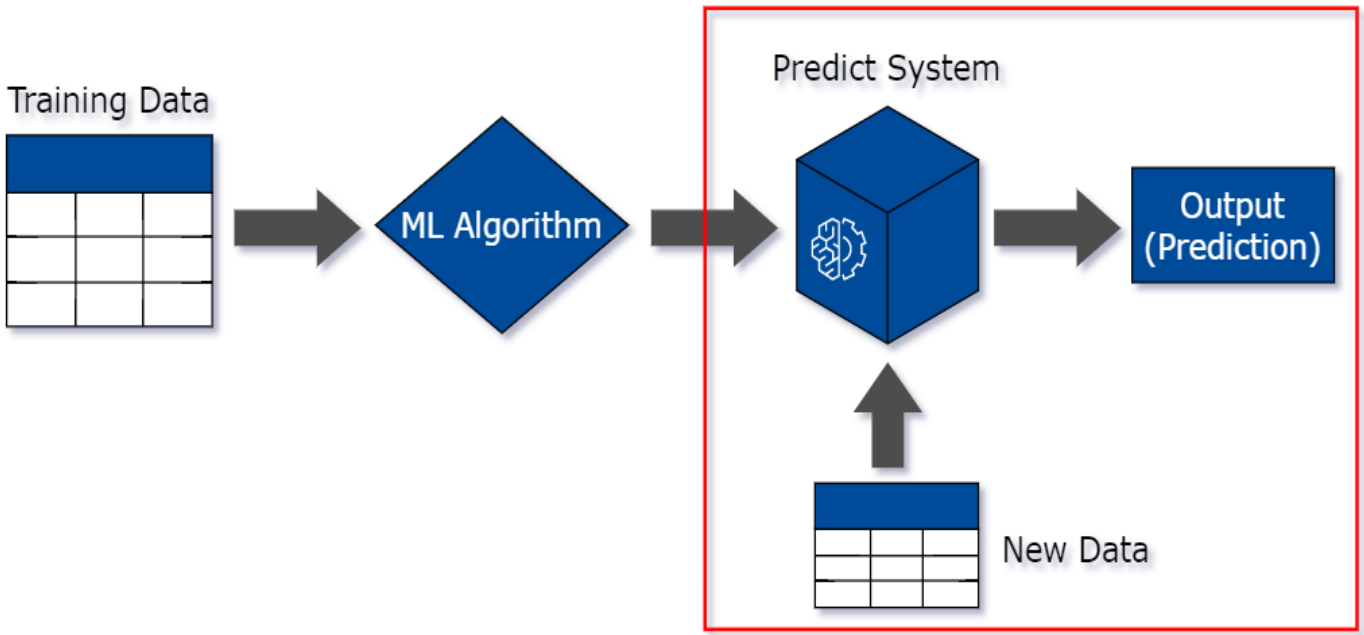
- Các em hãy truy cập và đường link sau và trả lời các câu hỏi dưới dạng trắc ngiệm

Link: https://bit.ly/ktktb1

Nội dung

Ý nghĩa và mục đích

- Có rất nhiều ứng dụng phục vụ dự đoán
  - Ứng dụng phục vụ trong đời sống xã hội
  - Trong chiến lược của mỗi công ty, tổ chức



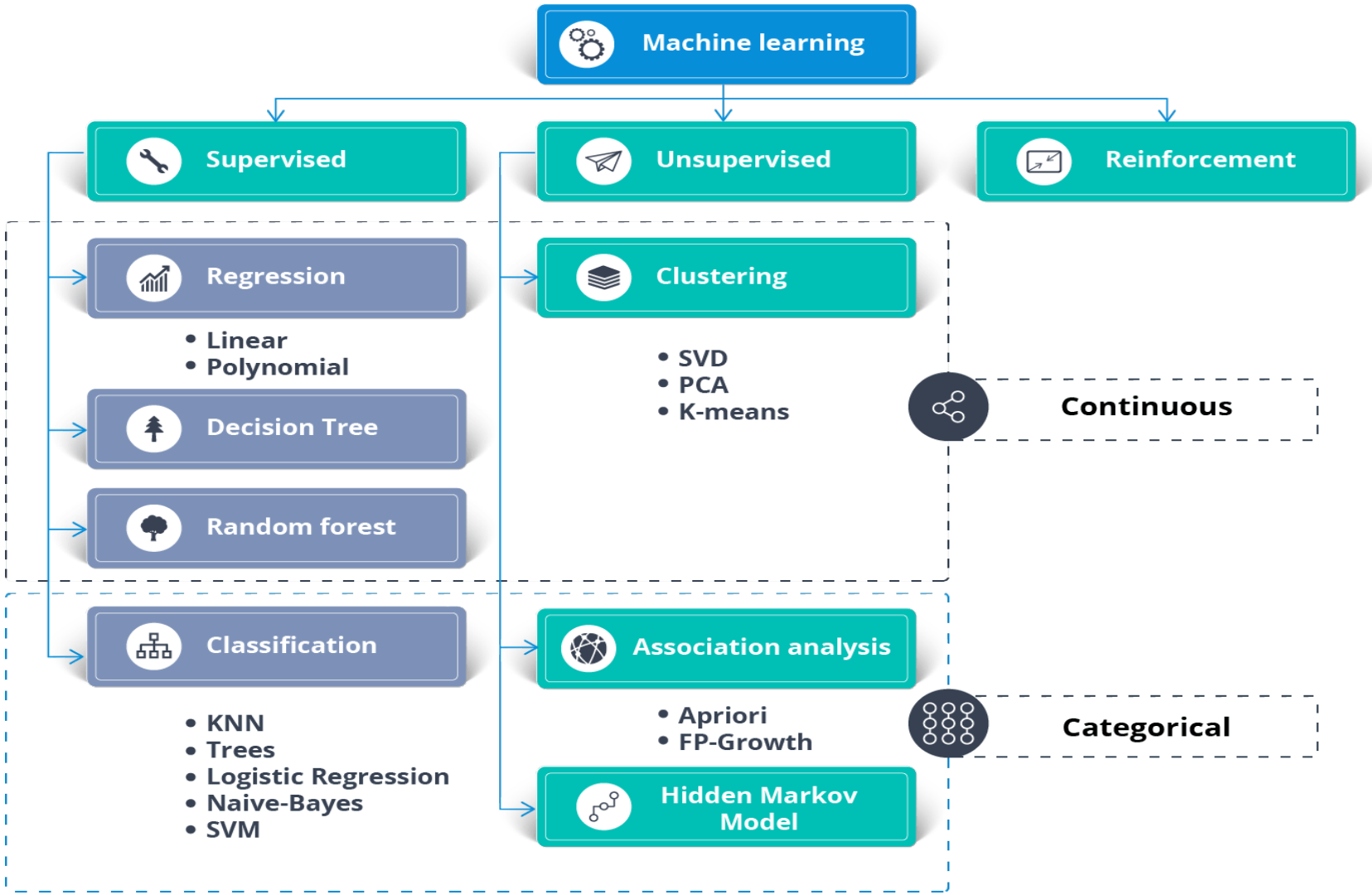
- Dự đoán các thông tin trong tương lai
  - Các thông tin chưa xảy ra
  - Các thông tin xảy ra nhưng cần kiểm chứng
  - Dự đoán giá nhà, giá vàng,...

- Dự đoán các thông tin trong tương lai
  - Các thông tin chưa xảy ra
  - Các thông tin xảy ra nhưng cần kiểm chứng
  - Dự đoán giá nhà, giá vàng,...
- Dựa trên các dữ liệu đã có
  - Các thông tin đã được xử lý
  - Các thông tin được giả định và có tính chắc chắn
  - Dựa trên số lượng lớn các thông tin
  - Dựa trên các thông tin xảy ra có tính quy luật

- Dự đoán các thông tin trong tương lai
  - Các thông tin chưa xảy ra
  - Các thông tin xảy ra nhưng cần kiểm chứng
  - Dự đoán giá nhà, giá vàng,...
- Dựa trên các dữ liệu đã có
  - Các thông tin đã được xử lý
  - Các thông tin được giả định và có tính chắc chắn
  - Dựa trên số lượng lớn các thông tin
  - Dựa trên các thông tin xảy ra có tính quy luật

Nhận xét: Các mô hình dự đoán có hiệu quả phụ thuộc rất nhiều vào các dữ liệu đã biết.

Lớp bài toán


Machine learning
Supervised
Unsupervised
Reinforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Trees
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Hidden Markov Model
Categorical

- Dữ liệu gồm các cặp đầu vào và nhãn tương ứng

  - Các $x_1, x_2, \ldots, x_m$ được gọi là các đặc trưng/thuộc tính của dữ liệu đầu vào/bản ghi/đối tượng/quan sát $\boldsymbol{x}$.
  - Trong thuật ngữ kinh tế lượng các $x_1, x_2, \ldots, x_m$ được gọi là biến độc lập và y gọi là biến phụ thuộc.
  - y là dữ liệu đầu ra/biến mục tiêu.

	x_1	x_2	y
$\boldsymbol{x}_1$	6	87837	787
$\boldsymbol{x}_2$	7	78	5415
$\boldsymbol{x}_3$	545	778	7507
$\boldsymbol{x}_4$	545	18744	7560
$\boldsymbol{x}_5$	88	788	6344

- $Input = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$
- $Ouput : y = f(\boldsymbol{x}) | y_i \cong f(\boldsymbol{x}_i)$
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$

Nhận xét: Việc xác định hàm $y = f(\boldsymbol{x})$ là việc xây dựng mô hình dự đoán.

Định nghĩa

Mô hình hồi quy

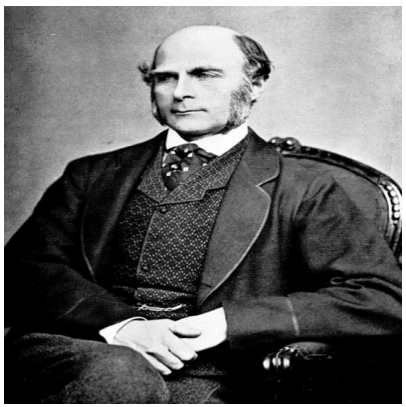
Một mô hình được gọi là hồi quy nếu mô hình đó mô tả được mối quan hệ giữa một hoặc nhiều biến độc lập với biến phụ thuộc.

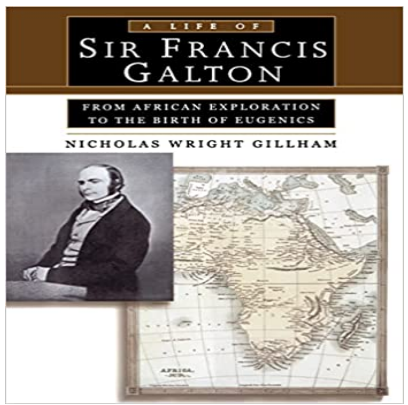
Định nghĩa

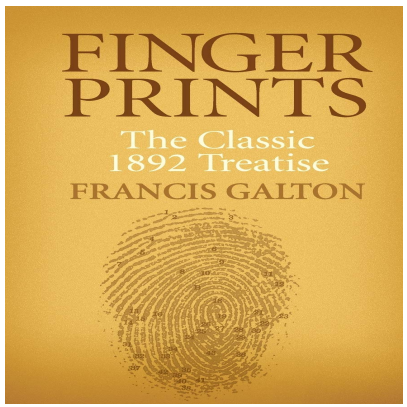
Mô hình hồi quy

Một mô hình được gọi là hồi quy nếu mô hình đó mô tả được mối quan hệ giữa một hoặc nhiều biến độc lập với biến phụ thuộc.

- Được Francis Galton (1822-1911) đưa ra lần đầu tiên.
- Mô hình hồi quy có thể dự đoán hoặc ước lượng giá trị của một biến số từ các giá trị của một hay nhiều biến số khác.
- Phân tích hồi quy là cơ sở cho nhiều loại dự đoán và xác định sự tác động lên các biến mục tiêu

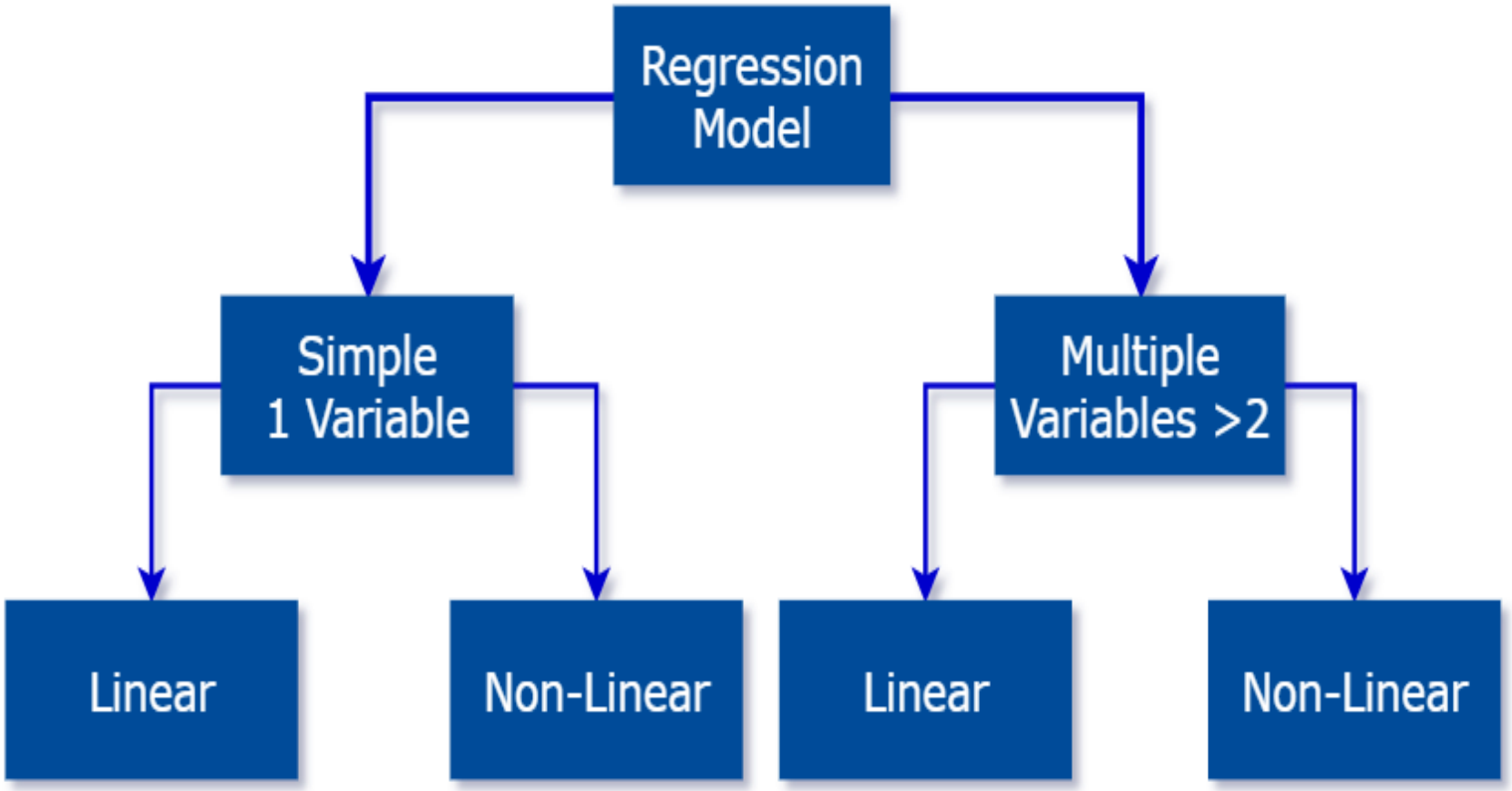






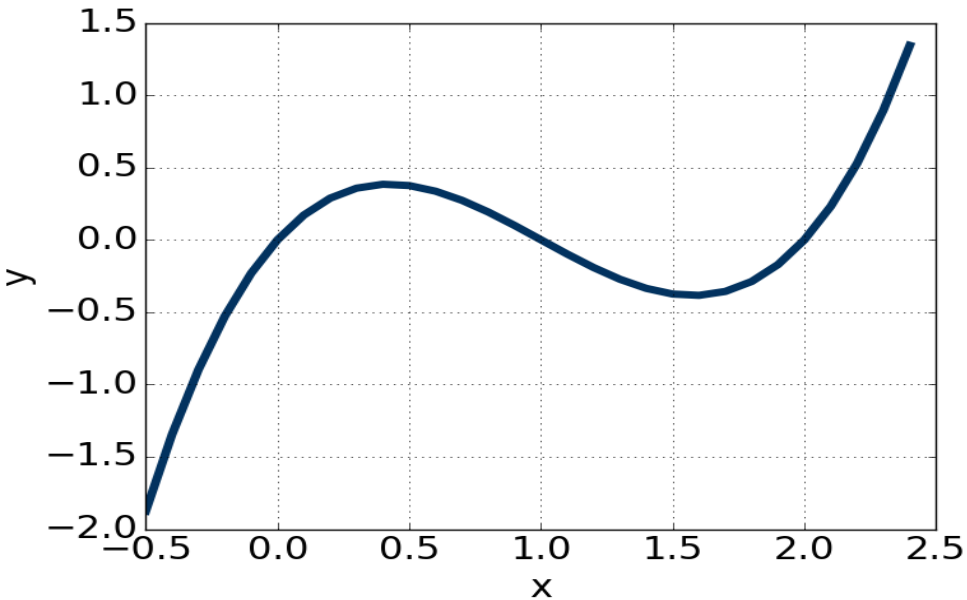
Các mô hình hồi quy

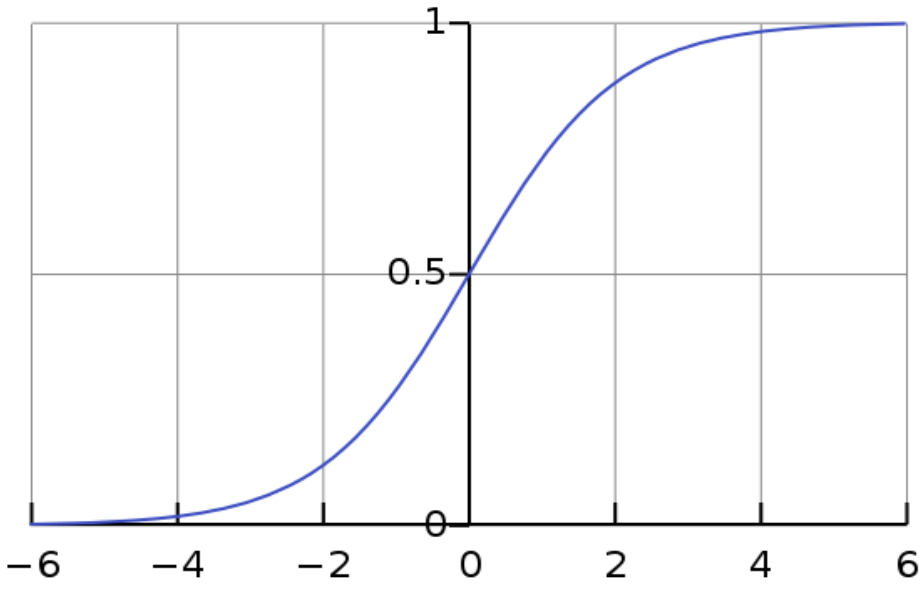
- Có rất nhiều cách để phân loại các mô hình hồi quy.



- Có 6 mô hình hồi quy thường được sử dụng.

- Hồi quy tuyến tính
- Hồi quy Ridge
- Hồi quy Lasso
- Hồi quy Logistic
- Hồi quy đa thức
- Hồi quy bayesian





Nhận xét: Các mô hình hồi quy được xây dựng dựa trên sự phân bố dữ liệu đầu vào.

Mô hình hồi quy tuyến tính

Định nghĩa

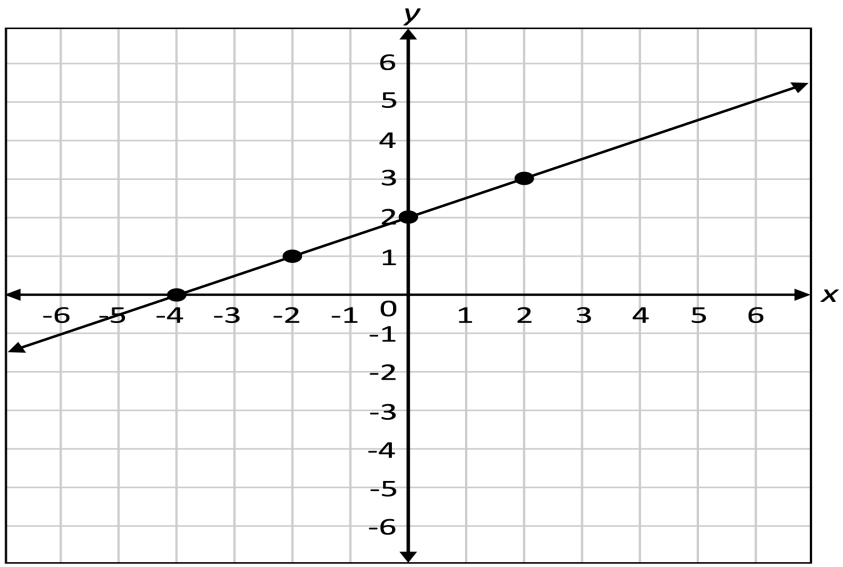
Là mô hình mô tả được mối quan hệ giữa một hoặc nhiều biến độc lập với biến phụ thuộc **dựa trên một hàm tuyến tính**.

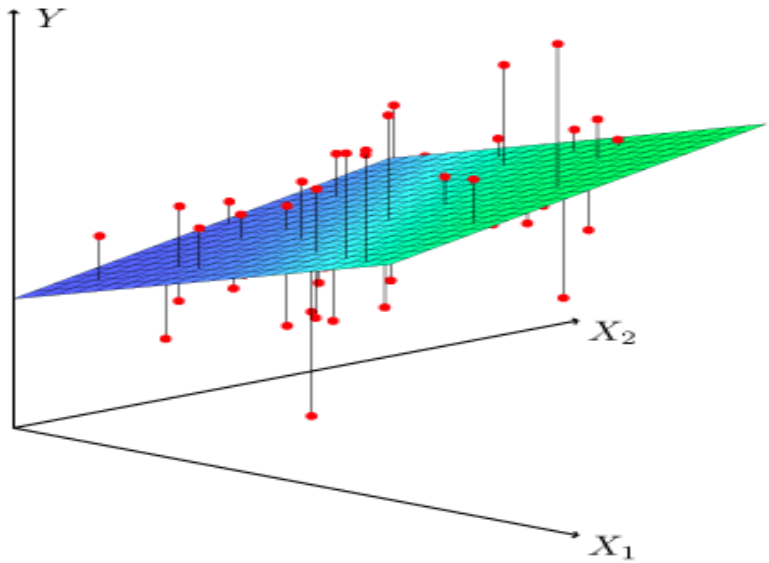
Mô hình hồi quy tuyến tính

Định nghĩa

Là mô hình mô tả được mối quan hệ giữa một hoặc nhiều biến độc lập với biến phụ thuộc **dựa trên một hàm tuyến tính**.

- Tuyến tính là các hàm số dạng bậc 1 (đường thẳng, mặt phẳng, siêu phẳng).

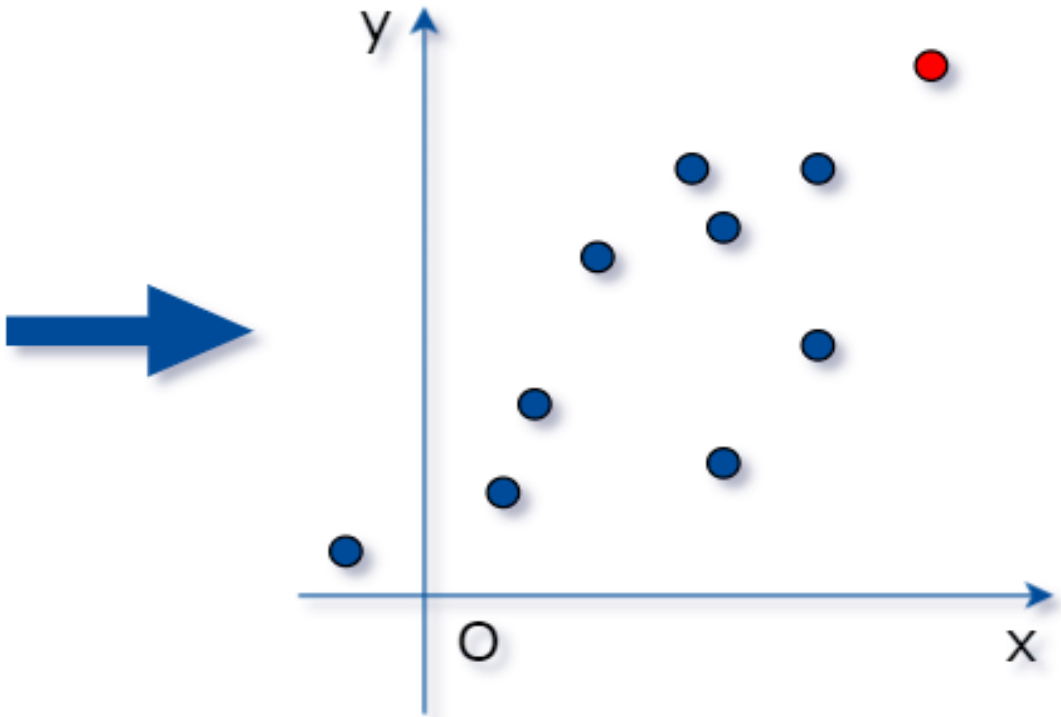




Mục đích

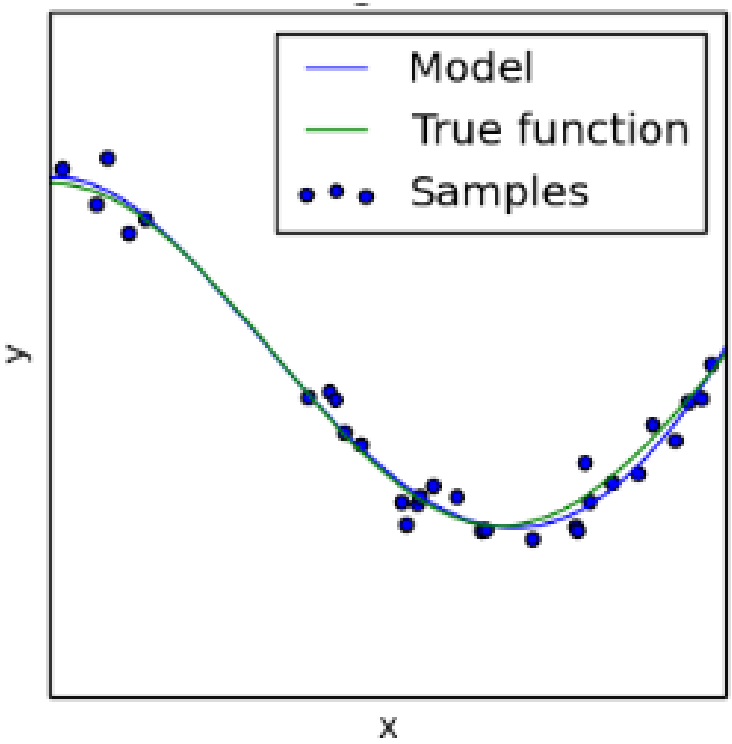
- Xét một ví dụ sau:

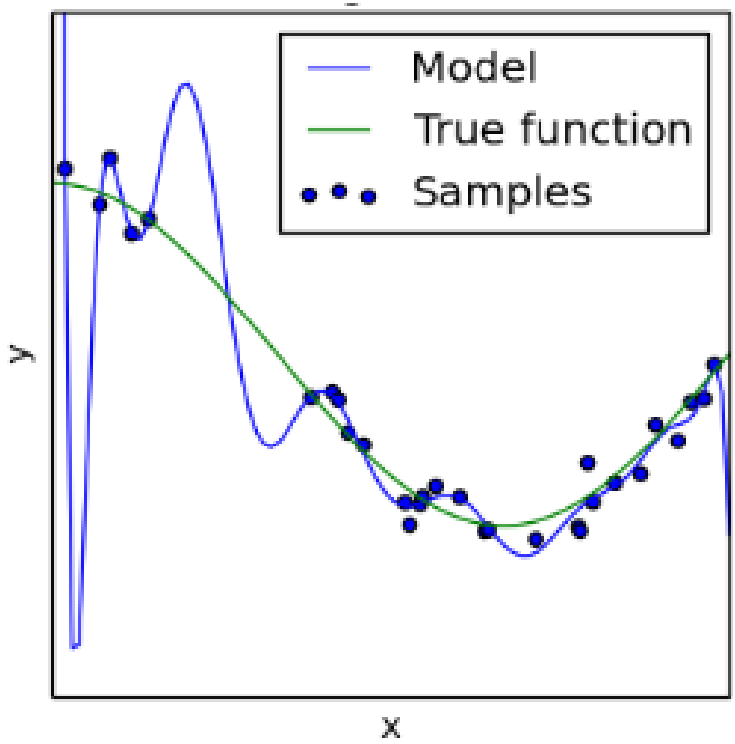
	x_1	y
$\boldsymbol{x}_1$	0.13	-0.91
$\boldsymbol{x}_2$	1.02	-0.17
$\boldsymbol{x}_3$	3.17	1.61
$\boldsymbol{x}_4$	-2.76	-3.31
$\boldsymbol{x}_5$	1.44	0.18
$\boldsymbol{x}_6$	5.28	3.36
$\boldsymbol{x}_7$	-1.74	-2.46
$\boldsymbol{x}_8$	7.93	5.56
...	...	...



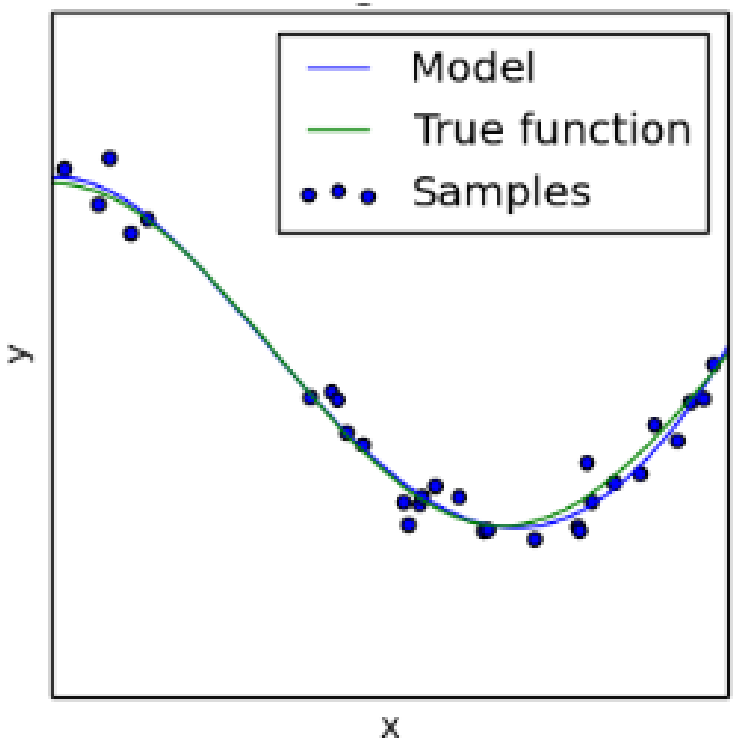
- Xác định một hàm $f(\boldsymbol{x})$ sao cho đi qua hết các điểm dữ liệu.
- Hàm $f(\boldsymbol{x})$ sử dụng để dự đoán cho các quan sát tiếp theo.

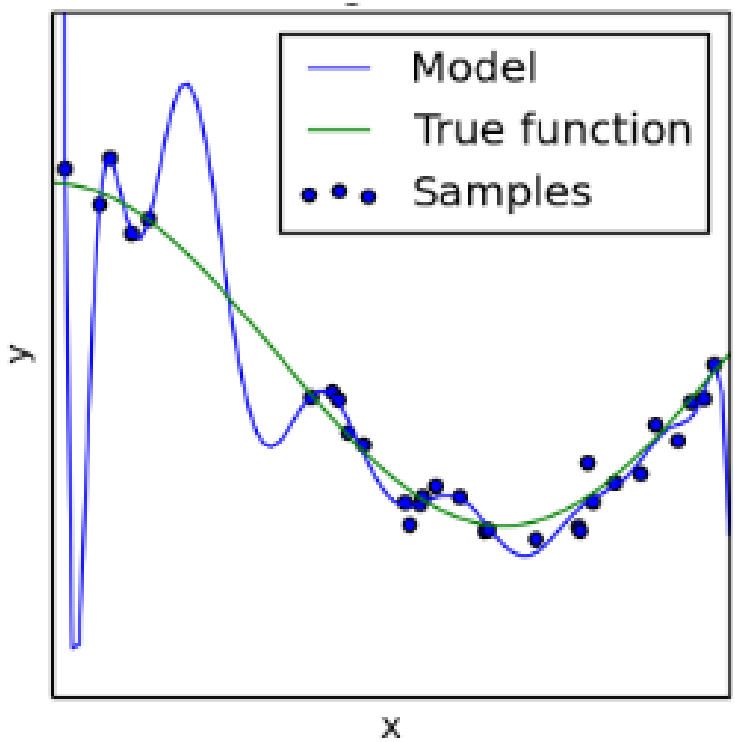
Các hàm có tính phức tạp (đa thức)





- Các hàm có tính phức tạp (đa thức)





Nhận xét:

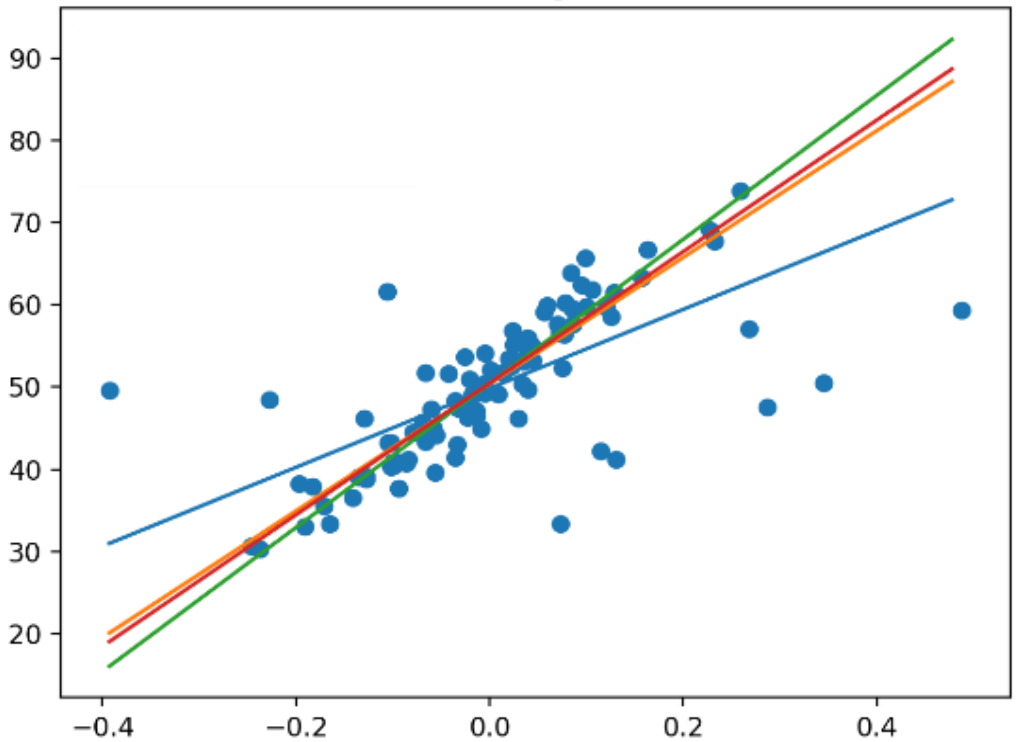
- Khó khăn trong việc xác định.
- Ảnh hưởng tốc độ tính toán với các dữ liệu tương lai.
- Không đảm bảo hiệu quả khi dự đoán dữ liệu tương lai.

- Xác định một hàm tuyến tính đảm bảo sai số dự đoán là ít nhất, dễ dàng tính toán với các dữ liệu tương lai và dễ dàng xây dựng.

- Mô hình HQTT tổng quát:

$$f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$$

  - Xác định/ tìm kiếm $f(\boldsymbol{x})$?
  - Tổng quát hóa là tốt nhất
  - $f(\boldsymbol{x})$ tốt hơn so với $g(\boldsymbol{x})$?

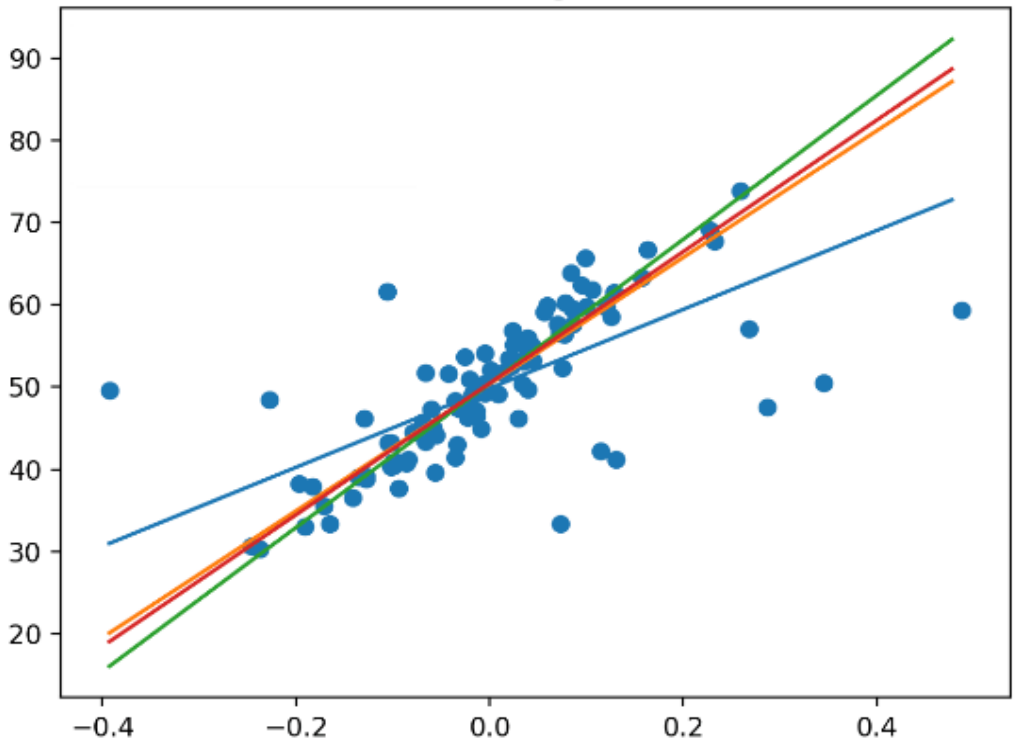


- Xác định một hàm tuyến tính đảm bảo sai số dự đoán là ít nhất, dễ dàng tính toán với các dữ liệu tương lai và dễ dàng xây dựng.

- Mô hình HQTT tổng quát:

$$f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$$

  - Xác định/ tìm kiếm $f(\boldsymbol{x})$?
  - Tổng quát hóa là tốt nhất
  - $f(\boldsymbol{x})$ tốt hơn so với $g(\boldsymbol{x})$?



Nhận xét: **Hệ số b được đưa vào mô hình để mang lại tính tổng quát hóa. Việc xác định/ tìm $f(x)$ là quá trình xây dựng mô hình hồi quy tuyến tính.**

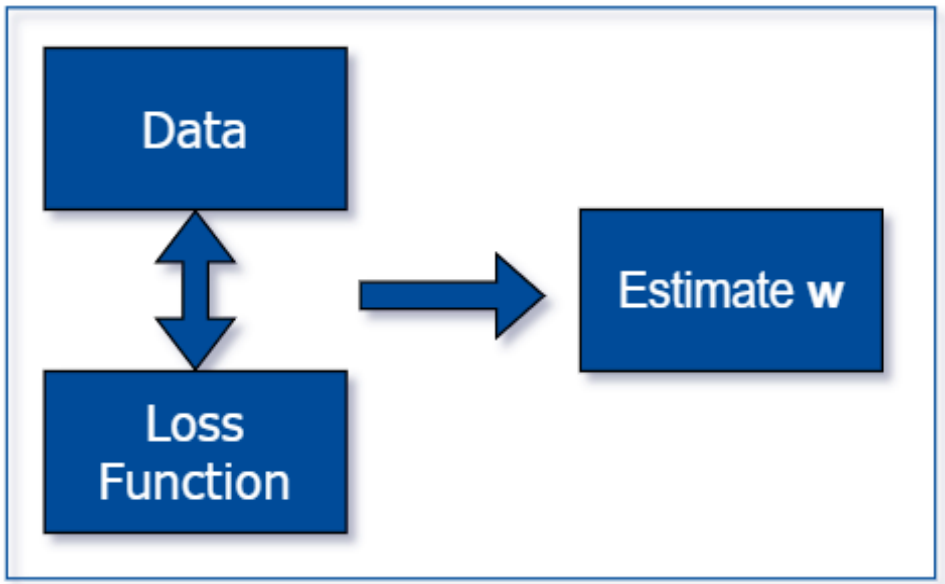
Nội dung

Định nghĩa

Xây dựng mô hình

Là việc xác định các tham số của mô hình dựa trên các dữ liệu đầu vào thông qua một độ đo tiêu chuẩn nào đó.

- Độ đo:
  - Dựa trên các dữ liệu đầu vào
  - Đánh giá mức độ chênh lệch giữa giá trị đầu ra thực tế và giá trị đầu ra được dự đoán.

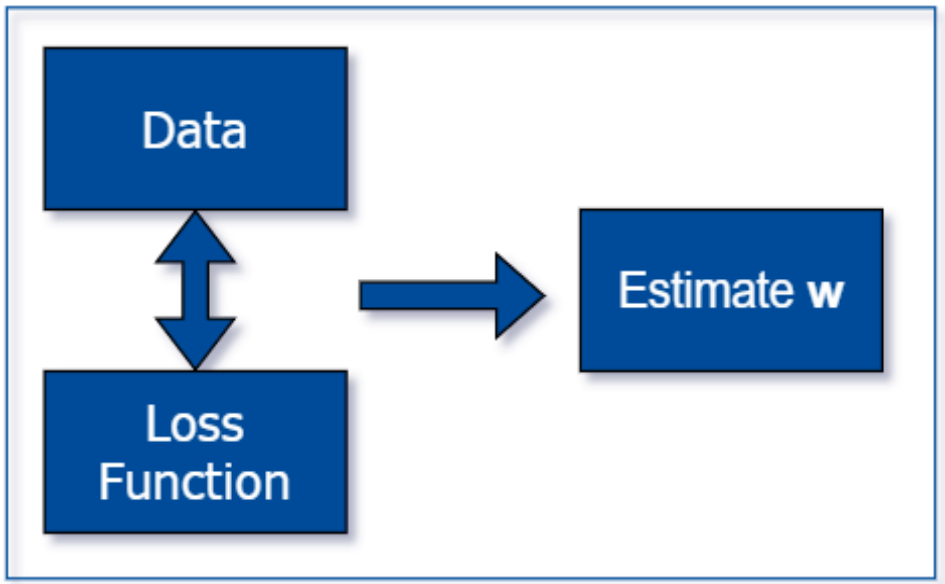


Định nghĩa

Xây dựng mô hình

Là việc xác định các tham số của mô hình dựa trên các dữ liệu đầu vào thông qua một độ đo tiêu chuẩn nào đó.

- Độ đo:
  - Dựa trên các dữ liệu đầu vào
  - Đánh giá mức độ chênh lệch giữa giá trị đầu ra thực tế và giá trị đầu ra được dự đoán.



Nhận xét: Hàm mất mát/lỗi thường được sử dụng. Quá trình tối ưu hàm mất mát được gọi là quá trình học của mô hình học máy.

Hàm mất mát

- Xét mỗi điểm dữ liệu $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, ta cần xây dựng hàm $f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$ sao cho $y \cong f(\boldsymbol{x})$.

Hàm mất mát

- Xét mỗi điểm dữ liệu $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, ta cần xây dựng hàm $f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$ sao cho $y \cong f(\boldsymbol{x})$.
  - Mỗi điểm dữ liệu bổ sung thêm thuộc tính $x_0 = 1$ và đặt $w_0 = b x_0$.

Hàm mất mát

- Xét mỗi điểm dữ liệu $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, ta cần xây dựng hàm $f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$ sao cho $y \cong f(\boldsymbol{x})$.
  - Mỗi điểm dữ liệu bổ sung thêm thuộc tính $x_0 = 1$ và đặt $w_0 = b x_0$.
  - Với mỗi điểm dữ liệu $\bar{\boldsymbol{x}} = (x_0, x_1, x_2, \ldots, x_m)$, ta cần xác định một giá trị dự đoán $\hat{y}$ sao cho phương trình sau đây đạt giá trị nhỏ nhất:

$$\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - w_0 x_0 - w_1 x_1 - \ldots - w_m x_m)^2 \qquad (1)$$

Hàm mất mát

- Xét mỗi điểm dữ liệu $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, ta cần xây dựng hàm $f(\boldsymbol{x}) = b + w_1 x_1 + \ldots + w_m x_m$ sao cho $y \cong f(\boldsymbol{x})$.
  - Mỗi điểm dữ liệu bổ sung thêm thuộc tính $x_0 = 1$ và đặt $w_0 = b x_0$.
  - Với mỗi điểm dữ liệu $\bar{\boldsymbol{x}} = (x_0, x_1, x_2, \ldots, x_m)$, ta cần xác định một giá trị dự đoán $\hat{y}$ sao cho phương trình sau đây đạt giá trị nhỏ nhất:

$$\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - w_0 x_0 - w_1 x_1 - \ldots - w_m x_m)^2 \qquad (1)$$

  - Nếu đặt $\boldsymbol{w} = (w_0, \ldots, w_m)^T$ là một vector hệ số cần tối ưu (dạng cột), khi đó phương trình (1) trở thành:

$$\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \bar{\boldsymbol{x}}\boldsymbol{w})^2 \qquad (2)$$

- Nếu xét trên các cặp dữ liệu đã biết $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots n$, với n là số lượng dữ liệu quan sát được. Điều mong muốn là tổng sai số phải là nhỏ nhất. Điều này dẫn tới bài toán tìm $\boldsymbol{w}$ sao cho phương trình sau là nhỏ nhất:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \bar{\boldsymbol{x}}_i \boldsymbol{w})^2 \tag{3}$$

- Nếu xét trên các cặp dữ liệu đã biết $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots n$, với n là số lượng dữ liệu quan sát được. Điều mong muốn là tổng sai số phải là nhỏ nhất. Điều này dẫn tới bài toán tìm $\boldsymbol{w}$ sao cho phương trình sau là nhỏ nhất:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \bar{\boldsymbol{x}}_i \boldsymbol{w})^2 \tag{3}$$

- $\mathcal{L}(\boldsymbol{w})$ được gọi là hàm mất mát của bài toán hồi quy tuyến tính và được sử dụng để xác định các tham số cho mô hình học máy.

- Nếu xét trên các cặp dữ liệu đã biết $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots n$, với n là số lượng dữ liệu quan sát được. Điều mong muốn là tổng sai số phải là nhỏ nhất. Điều này dẫn tới bài toán tìm $\boldsymbol{w}$ sao cho phương trình sau là nhỏ nhất:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \bar{\boldsymbol{x}}_i \boldsymbol{w})^2 \tag{3}$$

- $\mathcal{L}(\boldsymbol{w})$ được gọi là hàm mất mát của bài toán hồi quy tuyến tính và được sử dụng để xác định các tham số cho mô hình học máy.
- Đặt $\boldsymbol{w}^*$ là giá trị của $\boldsymbol{w}$ sao cho $\mathcal{L}(\boldsymbol{w})$ đạt giá trị nhỏ nhất. Khi đó:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$$

- Nếu xét trên các cặp dữ liệu đã biết $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots n$, với n là số lượng dữ liệu quan sát được. Điều mong muốn là tổng sai số phải là nhỏ nhất. Điều này dẫn tới bài toán tìm $\boldsymbol{w}$ sao cho phương trình sau là nhỏ nhất:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \bar{\boldsymbol{x}}_i \boldsymbol{w})^2 \tag{3}$$

- $\mathcal{L}(\boldsymbol{w})$ được gọi là hàm mất mát của bài toán hồi quy tuyến tính và được sử dụng để xác định các tham số cho mô hình học máy.
- Đặt $\boldsymbol{w}^*$ là giá trị của $\boldsymbol{w}$ sao cho $\mathcal{L}(\boldsymbol{w})$ đạt giá trị nhỏ nhất. Khi đó:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$$

- $\boldsymbol{w}^*$ được gọi là điểm tối ưu của mô hình.

Xác định tham số

- Xác định vector hệ số $\boldsymbol{w}$.
  - Có rất nhiều phương pháp để xác định $\boldsymbol{w}^*$. Phương pháp tổng bình phương nhỏ nhất (OLS-Odinary Least Squares) thường được sử dụng.

Xác định tham số

- Xác định vector hệ số $\boldsymbol{w}$.

- Có rất nhiều phương pháp để xác định $\boldsymbol{w}^*$. Phương pháp tổng bình phương nhỏ nhất (OLS-Odinary Least Squares) thường được sử dụng.
- Đặt $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_n)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ là ma trận chứa các vector hàng của dữ liệu đầu vào và vector cột chứa các dữ liệu đầu ra.

Xác định tham số

- Xác định vector hệ số $\boldsymbol{w}$.

- Có rất nhiều phương pháp để xác định $\boldsymbol{w}^*$. Phương pháp tổng bình phương nhỏ nhất (OLS-Odinary Least Squares) thường được sử dụng.
- Đặt $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_n)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ là ma trận chứa các vector hàng của dữ liệu đầu vào và vector cột chứa các dữ liệu đầu ra.
- Phương trình (3) trở thành:

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2} \left\| \boldsymbol{y} - \bar{\boldsymbol{X}} \boldsymbol{w} \right\|_2^2 \tag{4}$$

Xác định tham số

- Xác định vector hệ số $\boldsymbol{w}$.

- Có rất nhiều phương pháp để xác định $\boldsymbol{w}^*$. Phương pháp tổng bình phương nhỏ nhất (OLS-Odinary Least Squares) thường được sử dụng.
- Đặt $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_n)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ là ma trận chứa các vector hàng của dữ liệu đầu vào và vector cột chứa các dữ liệu đầu ra.
- Phương trình (3) trở thành:

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2} \left\| \boldsymbol{y} - \bar{\boldsymbol{X}} \boldsymbol{w} \right\|_2^2 \tag{4}$$

- $\|\boldsymbol{z}\|_2^2$ là tổng bình phương mỗi phần tử của $\boldsymbol{z}$.

- Lấy đạo hàm theo $\boldsymbol{w}$.

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) \tag{5}$$

- Lấy đạo hàm theo $\boldsymbol{w}$.

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) \tag{5}$$

- $\bar{\boldsymbol{X}} = n \times m$, $\boldsymbol{w} = m \times 1$, $\boldsymbol{y} = n \times 1 \Rightarrow \left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) = n \times 1$

- Lấy đạo hàm theo $\boldsymbol{w}$.

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) \tag{5}$$

- $\bar{\boldsymbol{X}} = n \times m$, $\boldsymbol{w} = m \times 1$, $\boldsymbol{y} = n \times 1 \Rightarrow \left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) = n \times 1$
- Để thỏa mãn phương trình đạo hàm cần chuyển thành:

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}^T\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) = 0 \Rightarrow \boxed{\boldsymbol{w} = \left(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\right)^{-1}\bar{\boldsymbol{X}}^T\boldsymbol{y}} \tag{6}$$

- Lấy đạo hàm theo $\boldsymbol{w}$.

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) \tag{5}$$

- $\bar{\boldsymbol{X}} = n \times m$, $\boldsymbol{w} = m \times 1$, $\boldsymbol{y} = n \times 1 \Rightarrow \left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) = n \times 1$
- Để thỏa mãn phương trình đạo hàm cần chuyển thành:

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \bar{\boldsymbol{X}}^T\left(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}\right) = 0 \Rightarrow \boxed{\boldsymbol{w} = \left(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\right)^{-1}\bar{\boldsymbol{X}}^T\boldsymbol{y}} \tag{6}$$

Nhận xét:

- **$\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ phải là một ma trận khả nghịch.**
- **Trường hợp $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ không khả nghịch thì phương pháp OLS không thể sử dụng.**

Nội dung

Ví dụ minh họa

Với 13 quan sát cho biết diện tích và giá của 13 căn nhà. Xây dựng một mô hình hồi quy tuyến tính thể hiện mối quan hệ giữa diện tích và giá nhà.

O	S (m2)	p (Tỷ VNĐ)
x_1	73.5	1.49
x_2	75.0	1.50
x_3	76.5	1.51
x_4	79.0	1.54
x_5	81.5	1.58
x_6	82.5	1.59
x_7	84	1.60
x_8	85	1.62
x_9	86.5	1.63
x_{10}	87.5	1.64
x_{11}	89	1.66
x_{12}	90	1.67
x_{13}	91.5	1.68

Công cụ triển khai		
Ngôn ngữ	Python	3.7
Môi trường	Anaconda	Latest
Thư viện	Numpy, matplotlib	Latest
Soạn thảo	Jupyter Notebook	Latest

Khuyến nghị: Sử dụng Google Colab

- Truy cập vào link sau:

`https://colab.research.google.com/`

Triển khai mô hình

- Thêm thư viện, chuẩn bị dữ liệu đầu vào và xây dựng hàm trực quan hóa dữ liệu

```
[1]  import matplotlib.pyplot as plt
     import numpy as np
```

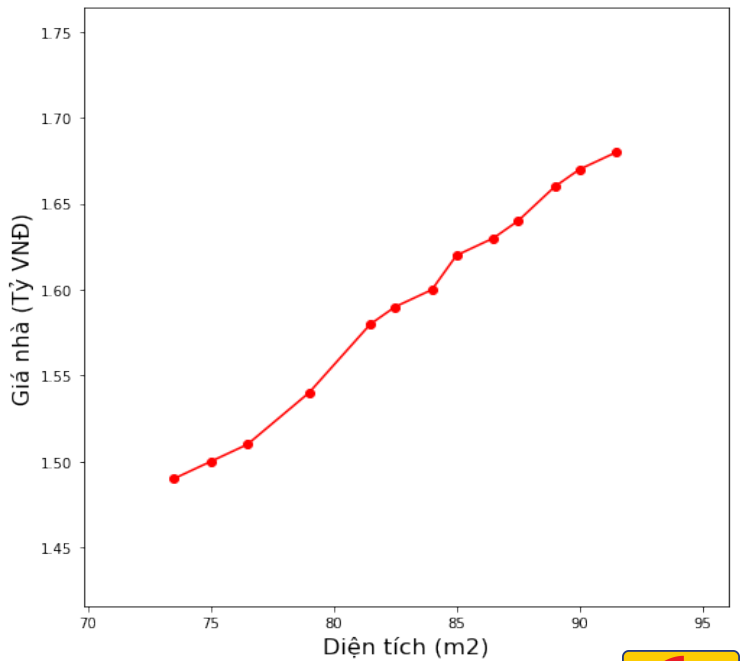
```
[2]  # Tiền xử lý dữ liệu
     # X (nxm) , y (nx1)
     X = np.array([[73.5,75.,76.5,79.,81.5,82.5,84.,85.,86.5,87.5,89.,90.,91.5]]).T
     y = np.array([[1.49,1.50,1.51,1.54,1.58,1.59,1.60,1.62,1.63,1.64,1.66,1.67,1.68]]).T
```

```
[3]  """
     Xây dựng hàm trực quan hóa dữ liệu
     """
     def plotData(X, y, title="", xlabel="", ylabel=""):
       # Tạo khung của đồ thị
       plt.figure(figsize=(8,8))
       plt.plot(X, y, 'r-o', label="price")

       # Xác định các giá trị max và min của dữ liệu
       X_min = np.min(X)
       X_max = np.max(X)
       y_min = np.min(y)
       y_max = np.max(y)

       plt.axis([X_min*0.95, X_max*1.05, y_min*0.95, y_max*1.05])
       plt.xlabel(xlabel, fontsize = 16)
       plt.ylabel(ylabel, fontsize = 16)
       plt.show()
```

```
[4]  plotData(X, y, title="Giá nhà theo diện tích", xlabel="Diện tích (m2)", ylabel="Giá nhà (Tỷ VNĐ)")
```



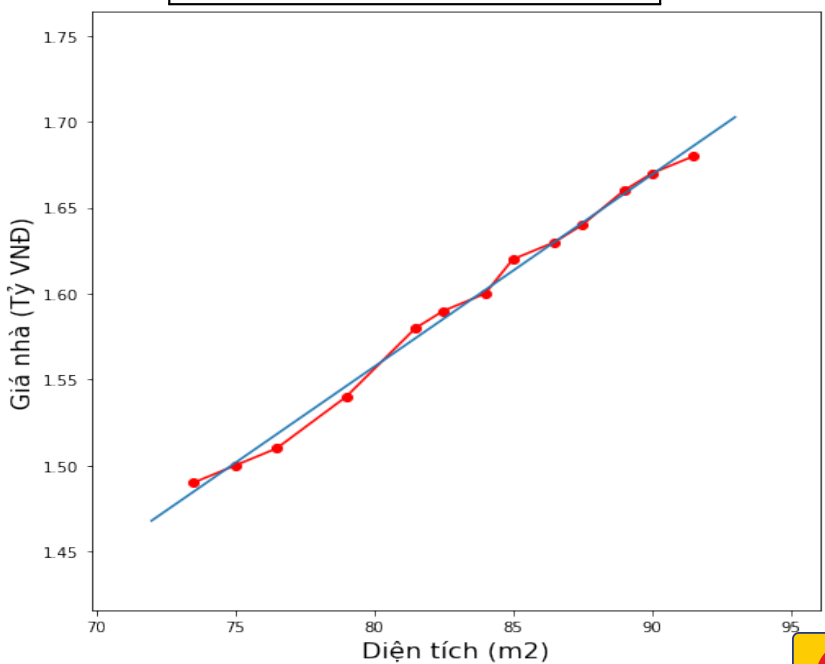
Xây dựng hàm tính toán tham số

- Tạo ma trận hệ số điều chỉnh sử dụng np.ones
- Ghép ma trận hệ số điều chỉnh và ma trận dữ liệu theo cột sử dụng np.concatenate
- Để dễ dàng tính toán, đặt $\boldsymbol{A} = \bar{\boldsymbol{X}}^T \bar{\boldsymbol{X}}$, $\boldsymbol{b} = \bar{\boldsymbol{X}}^T \boldsymbol{y}$

```
[5]  """
     Xác định tham số mô hình w1 và w0
     """
     def model(X, y):
       # Tạo ma trận hệ số điều chỉnh
       one = np.ones((X.shape[0],1))
       # Ghép theo cột với ma trận X (ma trận dữ liệu)
       Xbar = np.concatenate((one, X), axis = 1)
       A = np.dot(Xbar.T, Xbar)
       b = np.dot(Xbar.T, y)

       # Tính nghiệm
       w = np.dot(np.linalg.pinv(A),b)
       w_0 = w[0][0]
       w_1 = w[1][0]
       #print(w)
       print("Mô hình:" , w_0 , "+" , w_1 , "*x_1")
       return w_0, w_1
```

$$y = 0.66 + 0.01x_1$$



Đánh giá mô hình

- Trong mô hình hồi quy tuyến tính, người ta thường sử dụng chỉ số R-squared để đánh giá chất lượng của mô hình.
- R-squared cho ta biết mức độ các biến đầu vào (biến đầu vào) sẽ giải thích được bao nhiêu phần trăm các biến mục tiêu.
- R-squared càng lớn thì mô hình càng tốt.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{9}$$

Đánh giá mô hình

Trong đó TSS là tổng bình phương sai số toàn bộ mô hình (Total Sum Squared), RSS là tổng bình phương sai số ngẫu nhiên (Residual Sum Squared), ESS là tổng bình phương sai số được giải thích bởi mô hình (Explained Sum Squared).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{10}$$

Đánh giá mô hình

PHƯƠNG PHÁP OLS	
Ưu điểm	**Nhược điểm**
Có khả năng tổng quát hóa mô hình	Rất nhạy cảm với các dữ liệu nhiễu, không nhất quán
Thời gian xây dựng và tính toán nhanh	Mô hình không làm việc khi ma trận $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ là không khả nghịch
Có khả năng dự đoán tốt với các dữ liệu có quy luật	Không hiệu quả với các dữ liệu phân phối không bình thường

Đánh giá mô hình

PHƯƠNG PHÁP OLS	
Ưu điểm	**Nhược điểm**
Có khả năng tổng quát hóa mô hình	Rất nhạy cảm với các dữ liệu nhiễu, không nhất quán
Thời gian xây dựng và tính toán nhanh	Mô hình không làm việc khi ma trận $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ là không khả nghịch
Có khả năng dự đoán tốt với các dữ liệu có quy luật	Không hiệu quả với các dữ liệu phân phối không bình thường

Nhận xét: Có thể sử dụng các mô hình hồi quy biến thể khác để hạn chế một số nhược điểm trên:

- **Sử dụng mô hình hồi quy Ridge, Lasso**
- **Sử dụng phương pháp đạo hàm ngược**
- **Sử dụng mô hình hồi quy dạng phi tuyến**

Nội dung

Kiến thức quan trọng

- Nắm rõ các khái niệm cơ bản
  - Mô hình hồi quy, mô hình hồi quy tuyến tính.
  - Hiểu được ý nghĩa và mục đích của các mô hình hồi quy và mô hình hồi quy tuyến tính.
- Hiểu rõ quá trình xây dựng mô hình
  - Các kiến thức liên quan tới quá trình học, hàm mất mát.
  - Nắm rõ phương pháp OLS và cách biến đổi nghiệm tối ưu của mô hình.
- Biết cách triển khai mô hình trên các dữ liệu thực tế

Bài tập về nhà

Yêu cầu:

- **Xem lại mã nguồn và Slide bài giảng trên lớp sau đó tiến hành lựa chọn một bộ dữ liệu mẫu để thực hành.**
- **Đọc trước tài liệu về mô hình hồi quy Lasso, mô hình hồi quy Ridge và cách đánh giá hiệu quả của mô hình hồi quy tuyến tính.**
- **Truy cập và đường dẫn dưới đây để tham khảo và đọc trước các tài liệu yêu cầu.**

Tài liệu tham khảo: https://bit.ly/tltkhm

Mã Nguồn: https://bit.ly/sourceb1

THANK YOU