MÔ HÌNH HỒI QUY

Tiết 2: Một số mô hình hồi quy

Ths Phạm Việt Anh

Viện Công nghệ HaUI
Trường Đại học Công nghiệp Hà Nội

Ngày 12 tháng 3 năm 2024

Nội dung

1. Một số vấn đề trong việc xây dựng mô hình
2. Mô hình hồi quy Ridge và Lasso
3. Tổng kết

Nội dung

1. Một số vấn đề trong việc xây dựng mô hình
2. Mô hình hồi quy Ridge và Lasso
3. Tổng kết

- Sự chênh lệch về miền giá trị của mỗi thuộc tính trong dữ liệu

- Sự chênh lệch về miền giá trị của mỗi thuộc tính trong dữ liệu

#	Sex	Age	Weight (kg)	Height (m)	Smoker	Religion	Social Class
1	F	32	94.87	1.72	Y	Christian	C
2	F	34	99.39	1.63	N	Christian	D
3	F	33	124.15	1.66	N	Hindu	C
4	M	52	49.77	1.71	Y	Christian	E
5	F	57	65.13	1.80	N	Hindu	C
6	F	39	58.71	1.74	N	Buddhist	E
7	F	39	67.41	1.56	N	Christian	C
8	F	47	67.19	1.79	Y	Christian	B
9	M	58	42.95	1.48	N	Christian	A
10	M	17	109.52	1.62	N	Christian	C
11	F	42	91.12	1.76	N	Buddhist	D
12	F	48	58.07	1.50	N	Islamist	D
13	M	43	46.69	1.61	N	Hindu	B
14	M	55	85.38	1.54	N	Islamist	C
15	M	34	39.77	1.70	N	Christian	B
16	M	34	83.90	1.74	N	Islamist	D
17	M	51	55.72	1.93	Y	Islamist	B
18	F	47	57.10	1.51	N	Christian	C
19	M	38	54.01	1.85	Y	Islamist	C
20	M	45	73.10	1.59	N	Islamist	C

Ảnh hưởng: Tác động tới tính hiệu quả của nhiều thuật toán, ví dụ thời gian thực hiện, quá trình hội tụ, hay thậm chí ảnh hưởng cả tới độ chính xác của thuật toán.

- Sự chênh lệch về miền giá trị của mỗi thuộc tính trong dữ liệu

#	Sex	Age	Weight (kg)	Height (m)	Smoker	Religion	Social Class
1	F	32	94.87	1.72	Y	Christian	C
2	F	34	99.39	1.63	N	Christian	D
3	F	33	124.15	1.66	N	Hindu	C
4	M	52	49.77	1.71	Y	Christian	E
5	F	57	65.13	1.80	N	Hindu	C
6	F	39	58.71	1.74	N	Buddhist	E
7	F	39	67.41	1.56	N	Christian	C
8	F	47	67.19	1.79	Y	Christian	B
9	M	58	42.95	1.48	N	Christian	A
10	M	17	109.52	1.62	N	Christian	C
11	F	42	91.12	1.76	N	Buddhist	D
12	F	48	58.07	1.50	N	Islamist	D
13	M	43	46.69	1.61	N	Hindu	B
14	M	55	85.38	1.54	N	Islamist	C
15	M	34	39.77	1.70	N	Christian	B
16	M	34	83.90	1.74	N	Islamist	D
17	M	51	55.72	1.93	Y	Islamist	B
18	F	47	57.10	1.51	N	Christian	C
19	M	38	54.01	1.85	Y	Islamist	C
20	M	45	73.10	1.59	N	Islamist	C

Ảnh hưởng: Tác động tới tính hiệu quả của nhiều thuật toán, ví dụ thời gian thực hiện, quá trình hội tụ, hay thậm chí ảnh hưởng cả tới độ chính xác của thuật toán.

- Cần xây dựng một số kỹ thuật tiền xử lý dữ liệu

Một số kỹ thuật chuẩn hóa dữ liệu

Kỹ thuật chuẩn hóa dữ liệu

Là một bước quan trọng trong quá trình tiền xử lý dữ liệu, nhằm chuyển các thuộc tính về cùng một thang đo chung, giúp cho việc biểu diễn dữ liệu dễ dàng và các mô hình phân tích nhất là các mô hình học máy hoạt động hiệu quả hơn.

- Chuẩn hóa min-max

$$X_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

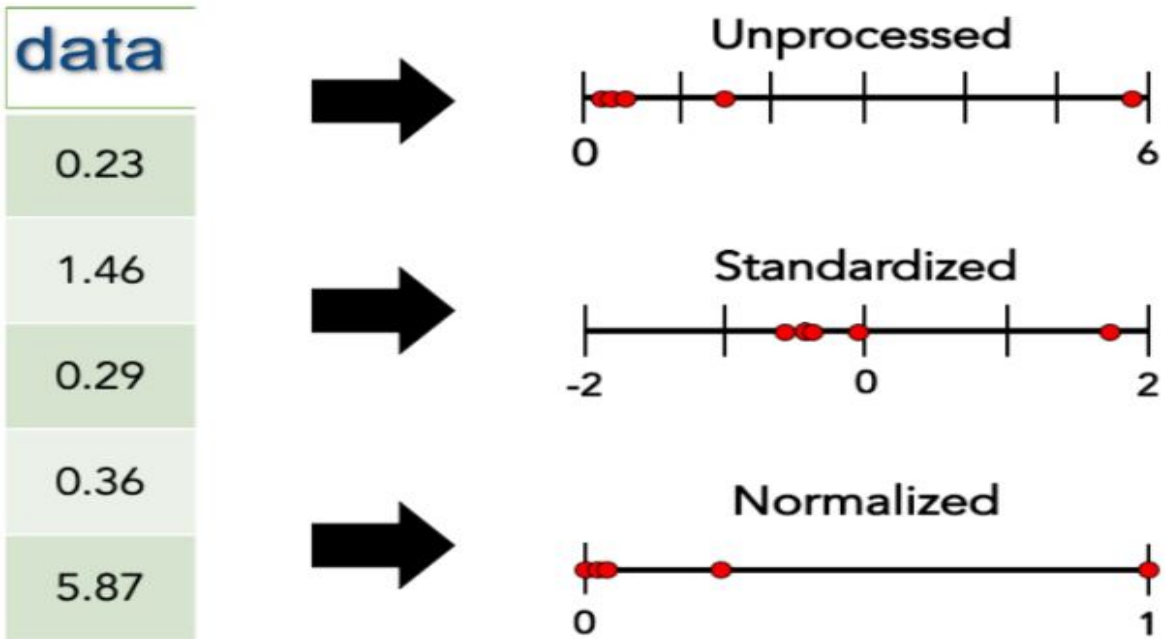
- Chuẩn hóa trung bình

$$X_{mean} = \frac{x - \bar{x}}{\max(x) - \min(x)} \tag{2}$$

Một số kỹ thuật chuẩn hóa dữ liệu

- Chuẩn hóa Z-score

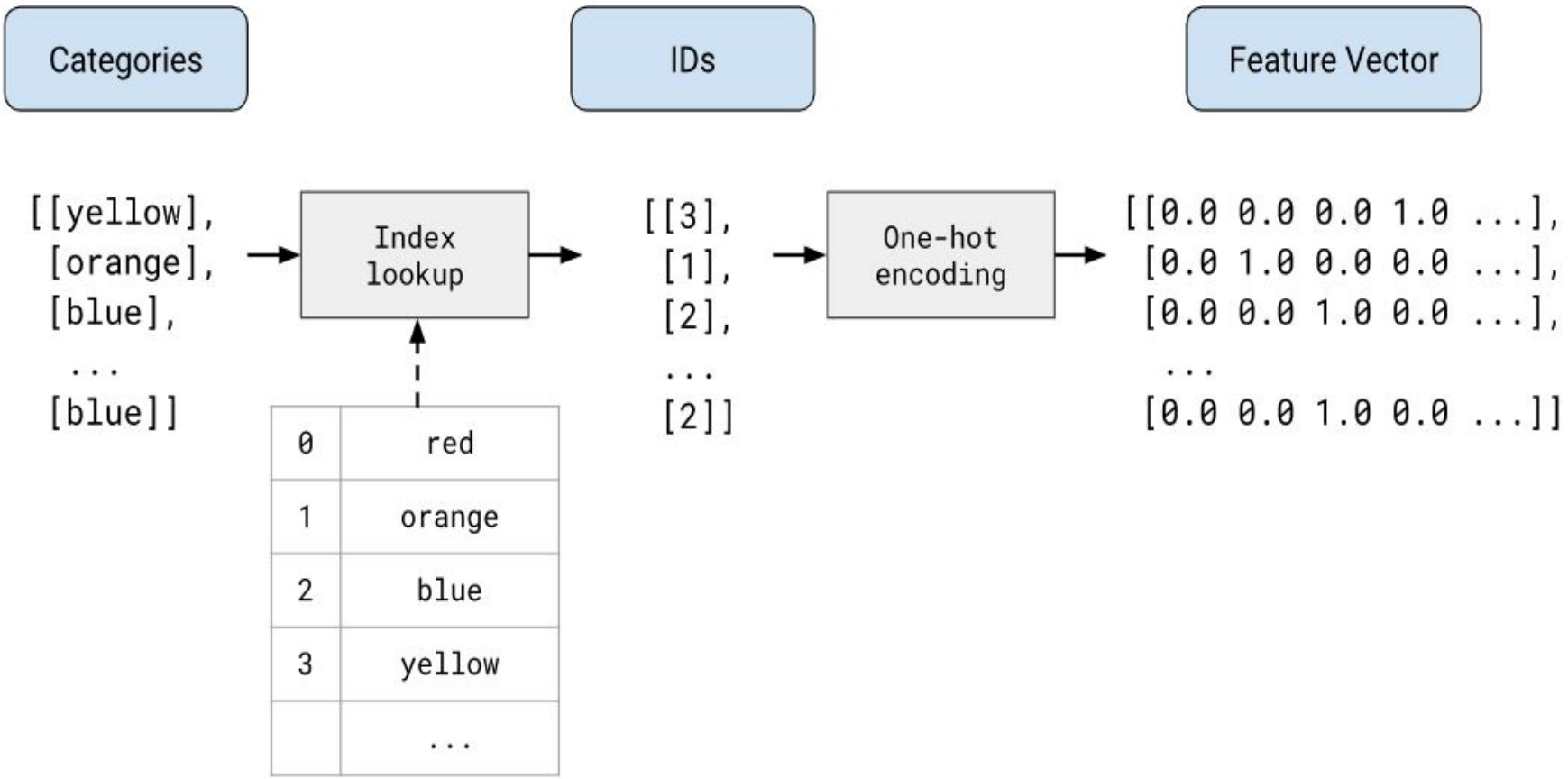
$$X_{stand} = \frac{x - \bar{x}}{\sigma_x} \tag{3}$$



Một số kỹ thuật chuẩn hóa dữ liệu

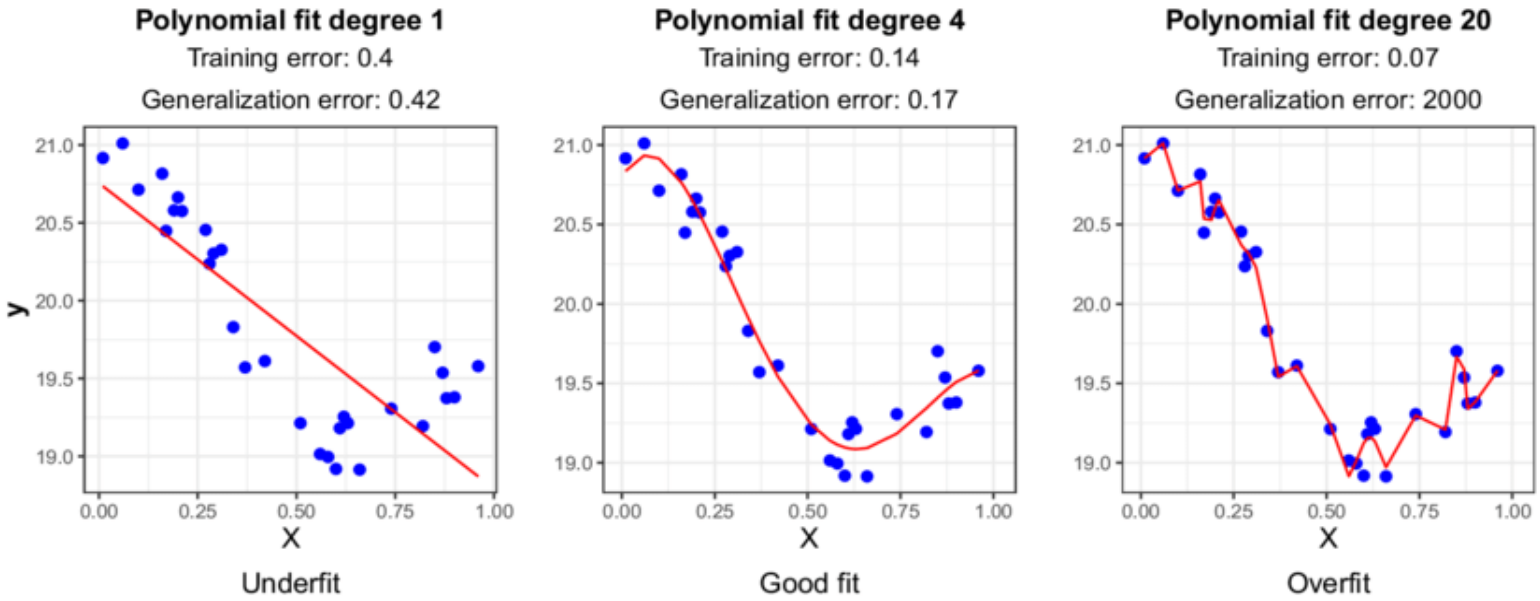
Kỹ thuật số hóa dữ liệu

Là kỹ thuật chuyển đổi các thuộc tính có giá trị dạng ký tự (symbol) về dạng số.

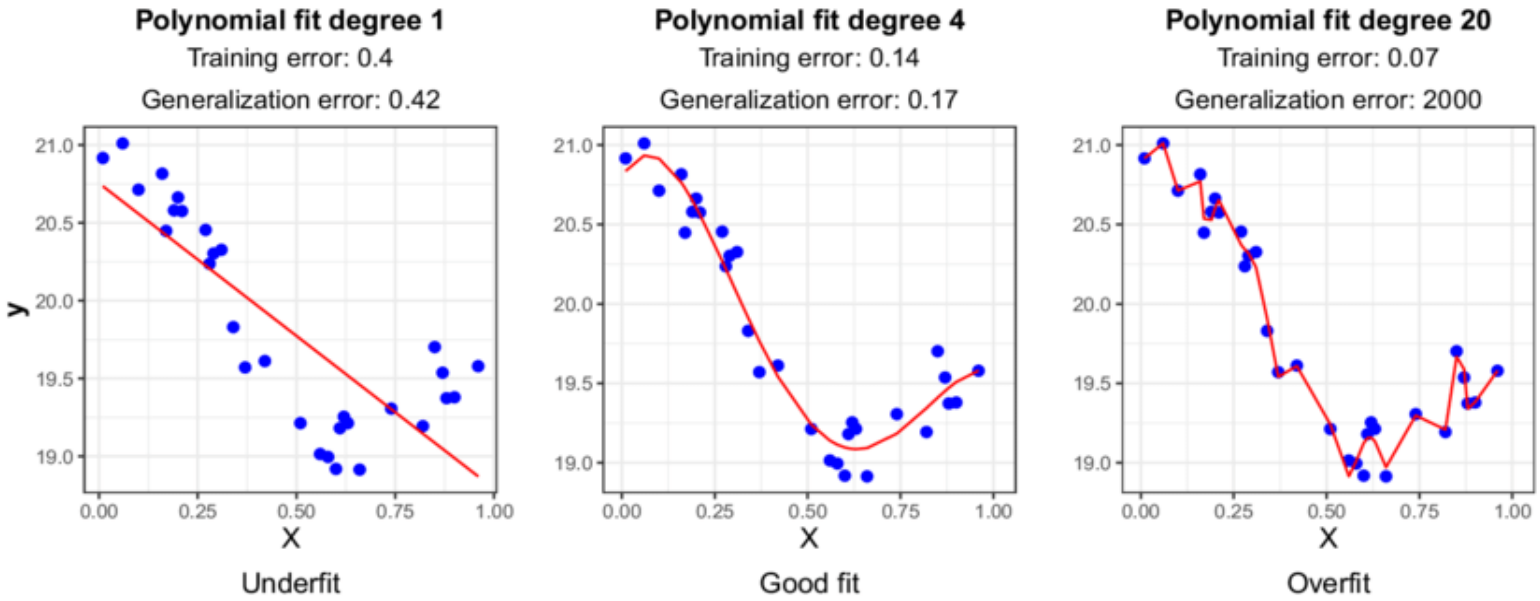


- Hiện tượng overfitting

Hiện tượng overfitting



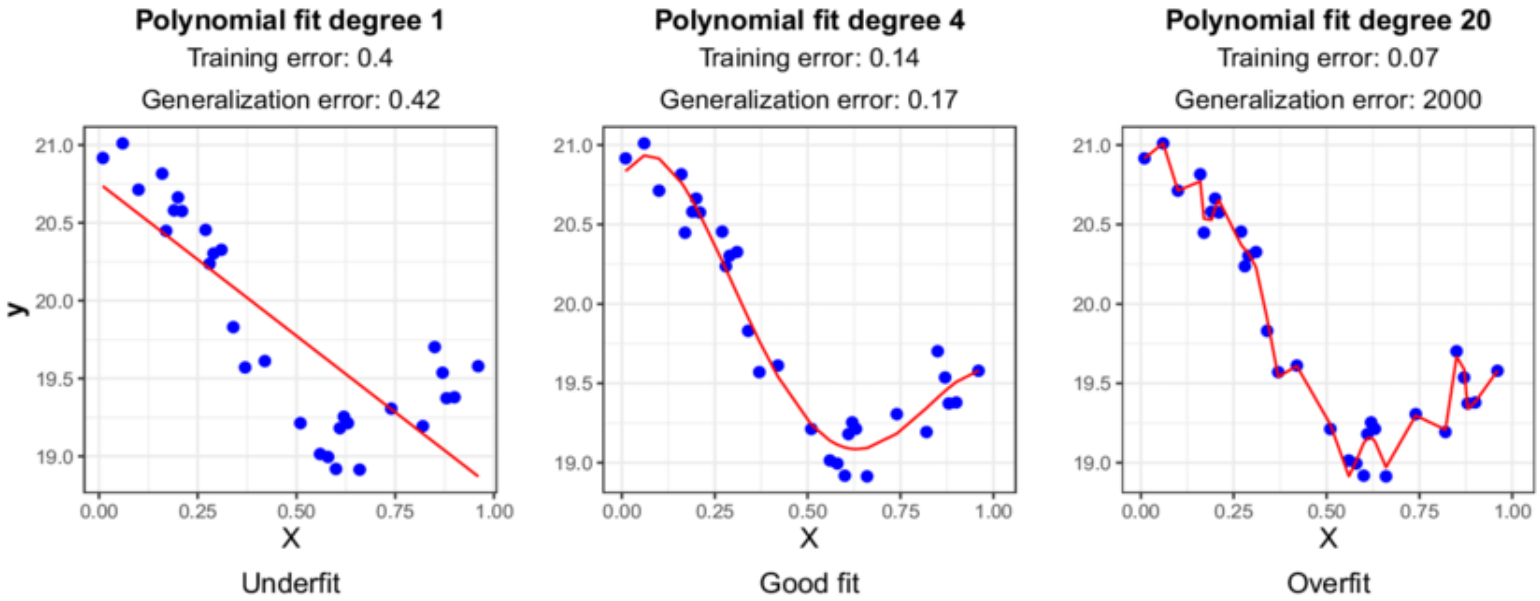
Hiện tượng overfitting



Overfitting

Là hiện tượng mô hình tìm được quá khớp với dữ liệu training nhưng lại sai lệch rất lớn với dữ liệu testing.

Hiện tượng overfitting

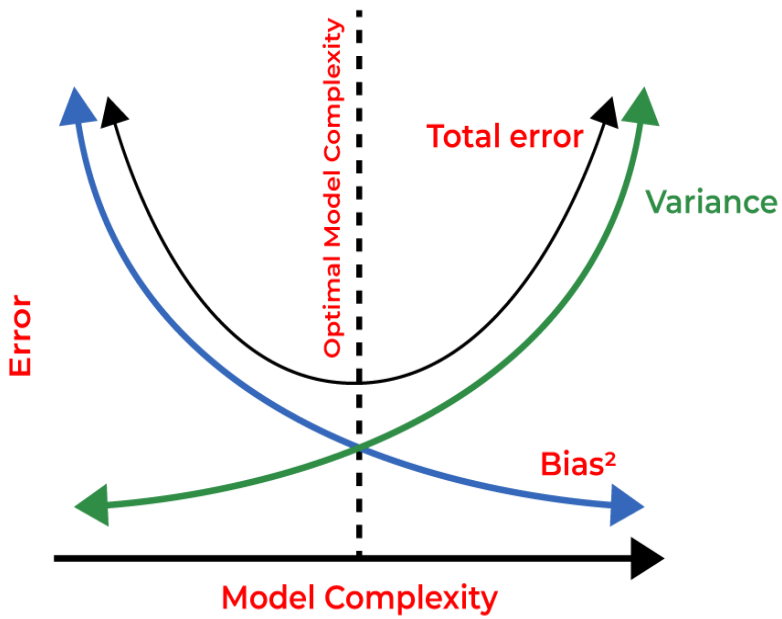


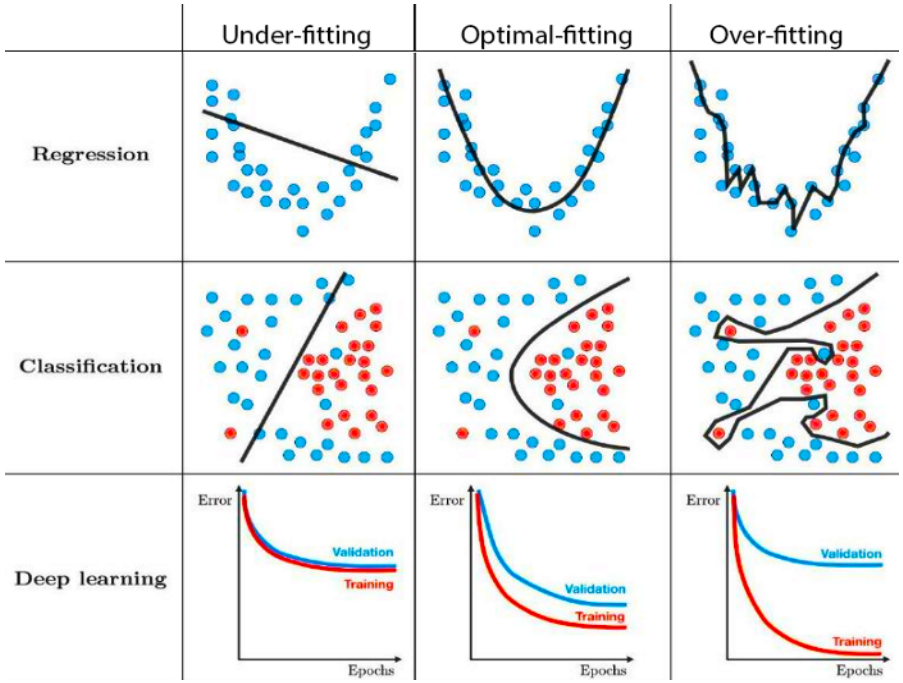
Overfitting

Là hiện tượng mô hình tìm được quá khớp với dữ liệu training nhưng lại sai lệch rất lớn với dữ liệu testing.

Một số biện pháp tránh overfitting

- Đưa mô hình từ phức tạp về dạng đơn giản

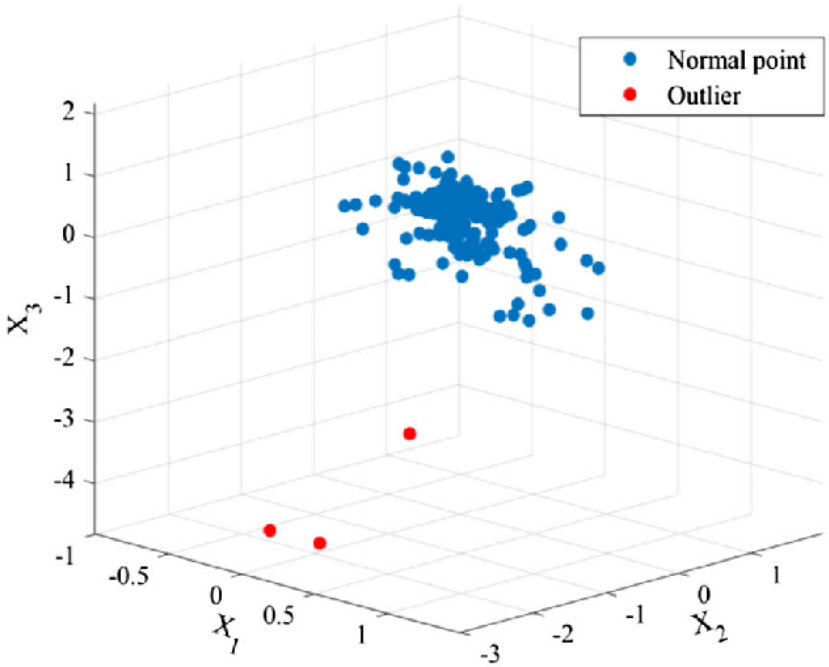


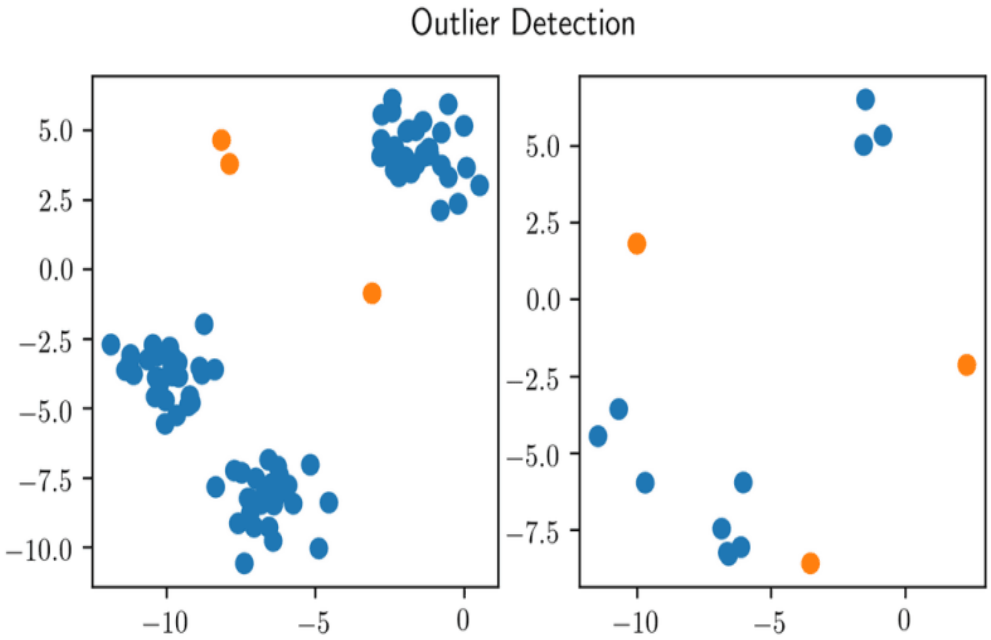


- Bổ sung thêm dữ liệu

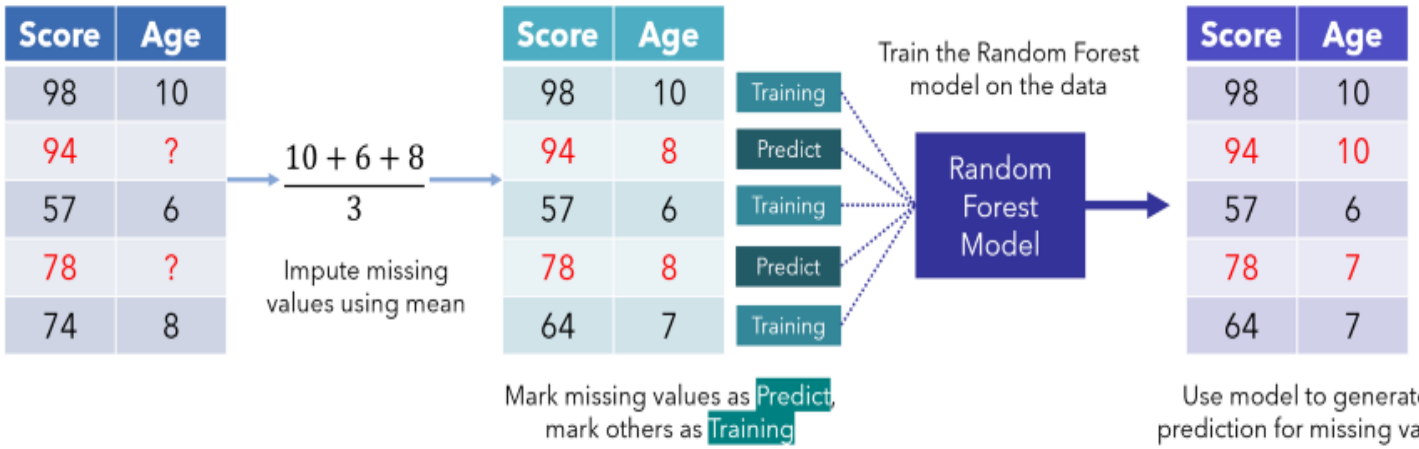
Lưu ý: Không sử dụng các mô hình có dạng phức tạp đối với tập dữ liệu có só chiều nhỏ.

- Loại bỏ ngoại lai (Remove outlier)

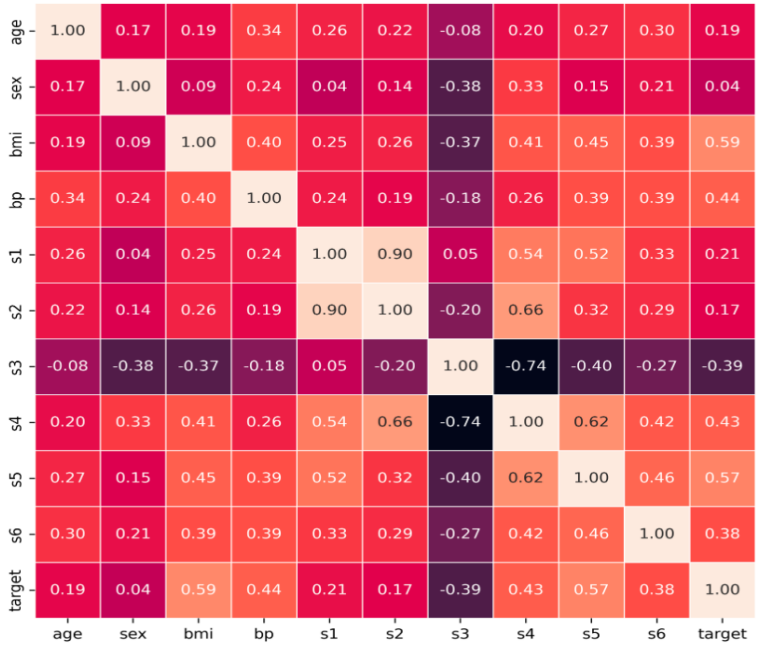


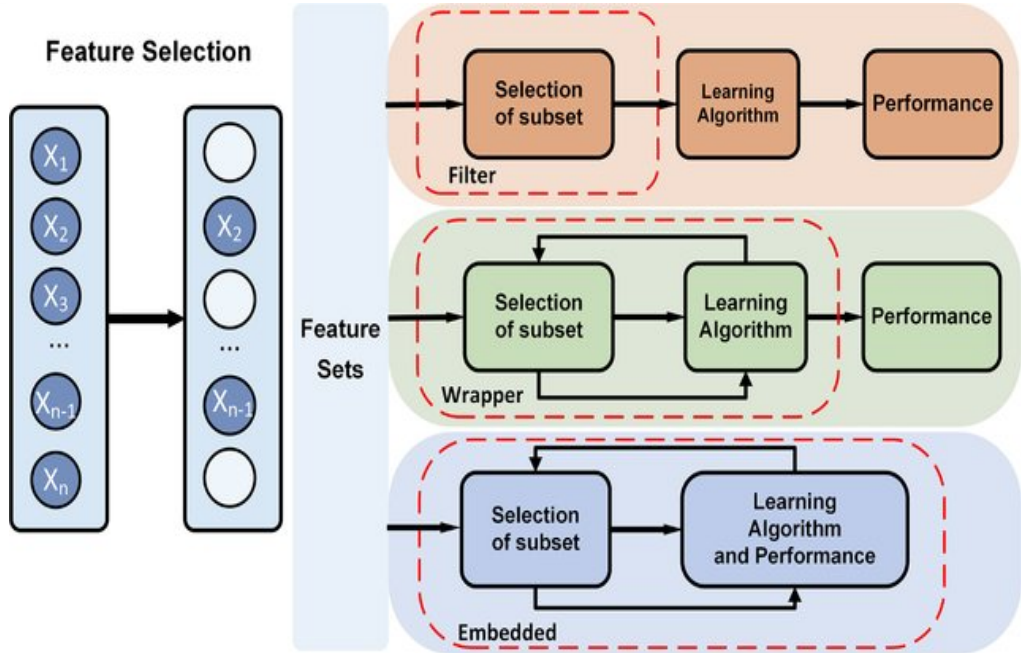


- Xử lý các trường hợp dữ liệu bị khuyết.

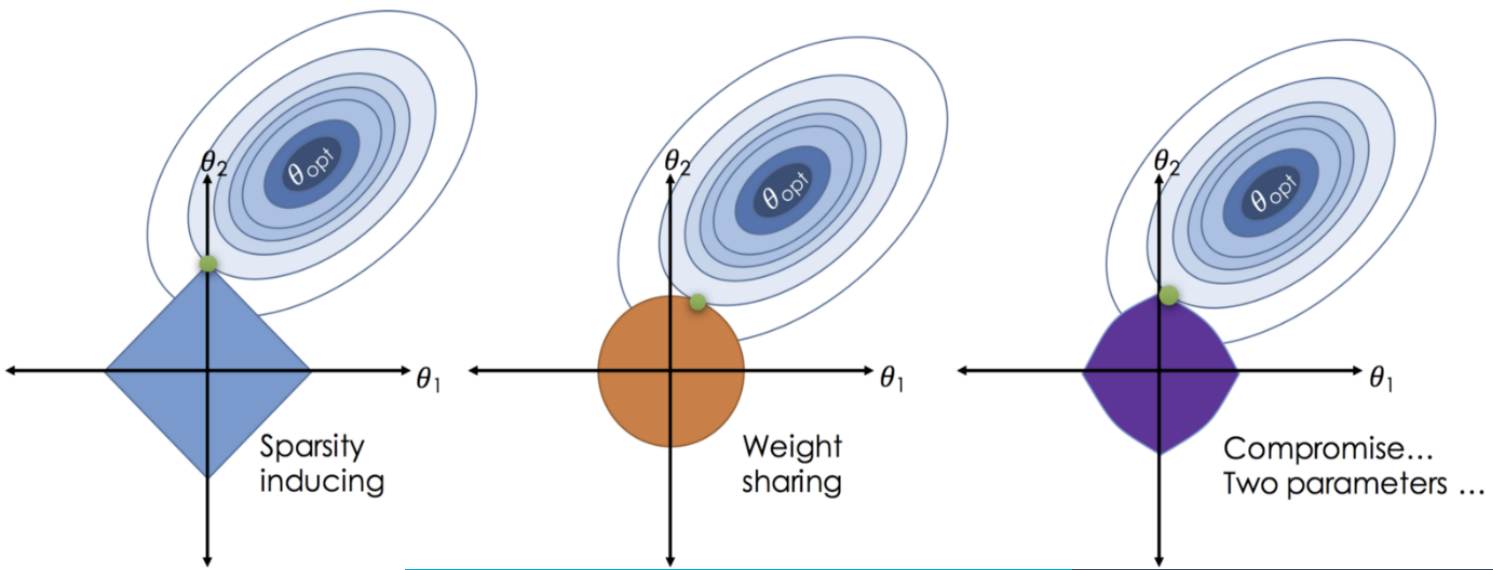


- Các thuộc tính có ảnh hưởng lớn tới nhau (Highly correlated features)





- Thay đổi hàm mất mát

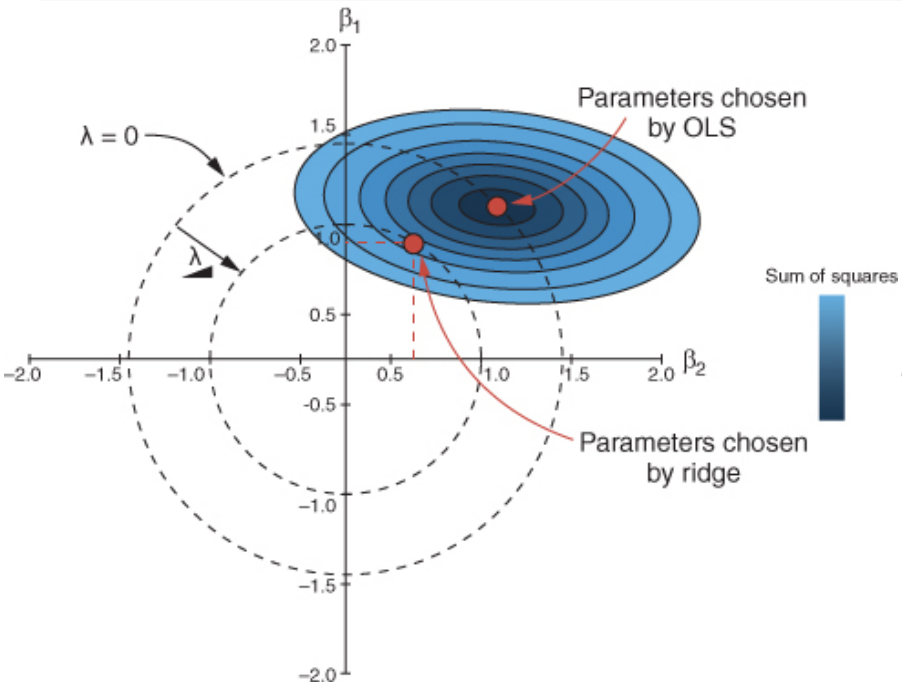


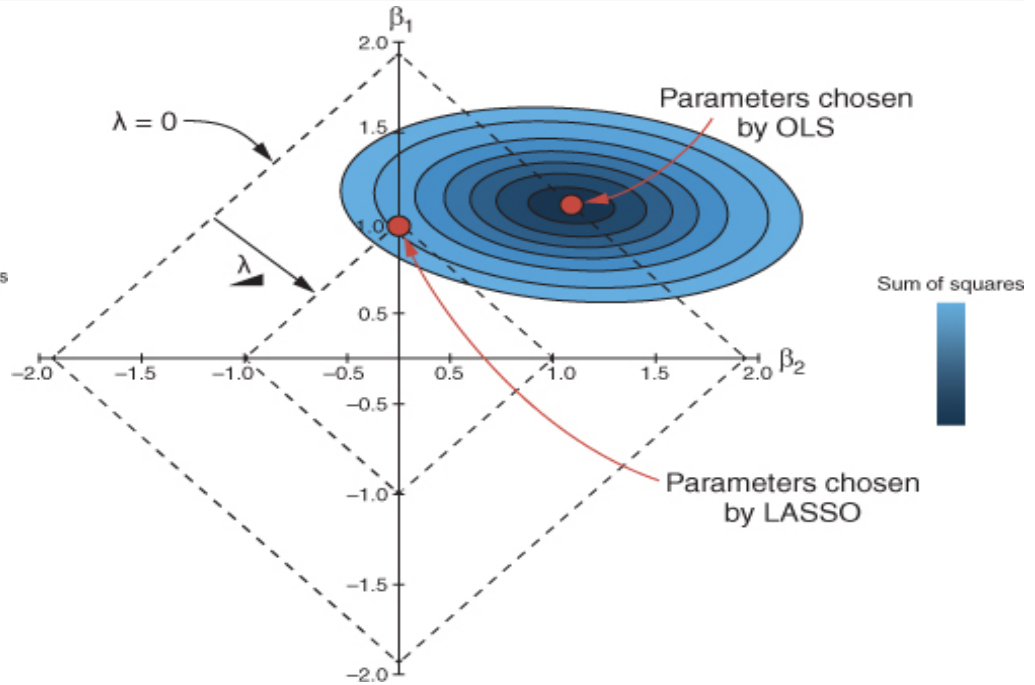
Nội dung

1. Một số vấn đề trong việc xây dựng mô hình
2. Mô hình hồi quy Ridge và Lasso
3. Tổng kết

Ridge và Lasso

Ridge và Lasso là những biến thể của hồi qui tuyến tính mà ở đó chúng ta thay đổi hàm mất mát để kiểm soát độ lớn của tham số huấn luyện nhằm giảm thiểu hiện tượng quá khớp trong các bài toán dự báo của học có giám sát.





Hồi Quy Ridge

Mục tiêu

Tương tự như phương pháp OLS, mô hình Ridge cũng tìm kiếm các trọng số của mô hình nhưng bổ sung thêm một đại lượng phạt để hạn chế sự biến đổi của các trọng số.

Chú ý

Đại lượng phạt chỉ được sử dụng trong quá trình tối ưu mô hình và tìm kiếm trọng số chứ không nằm trong phương trình dự đoán.

Hàm mất mát:

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

trong đó, $\alpha\left\|\mathbf{w}\right\|_2^2$ là đại lượng phạt (regularization term)

Chú ý

- Trường hợp $\alpha = 0$, bài toán trở về hồi quy tuyến tính.
- Trường hợp α nhỏ thì khả năng điều chỉnh overfitting thấp.
- Trường hợp α cao thì khả năng gia tăng mức độ kiểm soát lên độ lớn của các hệ số ước lượng sẽ cao.

Xác định trọng số:

- Lấy đạo hàm theo $\mathbf{w}$.

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \bar{\mathbf{X}}\left(\bar{\mathbf{X}}\mathbf{w} - \mathbf{y}\right) + 2\alpha\mathbf{w} = 0 \tag{5}$$

$$\Leftrightarrow \bar{\mathbf{X}}^T\bar{\mathbf{X}}\mathbf{w} - \bar{\mathbf{X}}\mathbf{y} + 2\alpha\mathbf{w} = 0 \Leftrightarrow \mathbf{w}\left(\bar{\mathbf{X}}^T\bar{\mathbf{X}} + 2\alpha\mathbf{I}\right) = \bar{\mathbf{X}}\mathbf{y}$$

$$\Rightarrow \boxed{\mathbf{w} = \left(\bar{\mathbf{X}}^T\bar{\mathbf{X}} + 2\alpha\mathbf{I}\right)^{-1}\bar{\mathbf{X}}\mathbf{y}}$$

- Bài toán tối ưu hàm mất mát của hồi qui Ridge về bản chất là tối ưu tổng bình phương phần dư và đại lượng phạt.

Hồi quy Ridge với thư viện Scikitlearn:

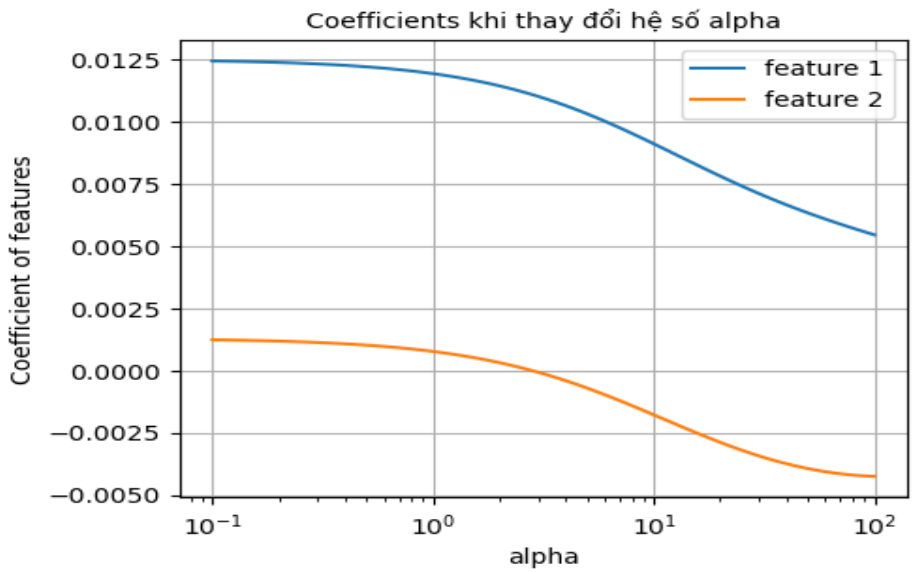
In [12]:
```
# import thư viện sklearn
from sklearn.linear_model import Ridge

# Khởi tạo bộ tham số alpha để thử nghiệm
n_alphas = 200
alphas = 1/np.logspace(1, -2, n_alphas)

# Khởi tạo các mảng tham số
coefs, inters, scores = [], [], []

# Duyệt alpha để tìm giá trị tốt nhất
for alpha in alphas:

    # Khởi tạo model
    ridge = Ridge(alpha=alpha, fit_intercept=True)

    # Huấn luyện model
    ridge.fit(X, y)

    # Đưa giá trị tham số vào mảng
    coefs.append(ridge.coef_)
    inters.append(ridge.intercept_)
    scores.append(ridge.score(X, y))

# Lấy index mà có giá trị score cao nhất
index_max = np.argmax(scores)

print( 'Best alpha: ', alphas[index_max])
print( 'Coefficient : ', coefs[index_max] )
print( 'Interception  : ', inters[index_max] )
print('Score of model : ', scores[index_max] )
```

```
Best alpha:  0.1
Coefficient :  [[0.01246261 0.00124324]]
Interception  :  [0.54241654]
Score of model :  0.9952984212286509
```

Duyệt giá trị α cho mô hình.

In [11]:
```
# import thư viện sklearn
from sklearn.linear_model import Ridge
# Khởi tạo model
reg_ridge = Ridge(alpha = 1.0)
# Huấn luyện model
reg_ridge.fit(X, y)
# Bộ tham số model
print( 'Coefficient : ', reg_ridge.coef_ )
print( 'Interception  : ', reg_ridge.intercept_ )
# Mức sai số của model
print('Score of model : ', reg_ridge.score(X, y))
```

```
Coefficient :  [[0.01195093 0.00077223]]
Interception  :  [0.59023779]
Score of model :  0.9951590382690014
```



Hồi quy Lasso

Mục tiêu

Thay vì sử dụng đại lượng phạt là norm chuẩn bậc hai thì mô hình Lasso sử dụng norm chuẩn bậc 1 để hạn chế sự biến đổi của các trọng số.

Hàm mất mát:

$$\mathcal{L}_{Lasso}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\right\|_2^2 + \alpha\left\|\mathbf{w}\right\|_1 \tag{6}$$

trong đó, $\left\|\mathbf{w}\right\|_1 < C$, $C > 0$.

Chú ý

Mô hình Lasso thường tạo ra nghiệm thưa, do đó nhiều thành phần trọng số có giá trị bằng 0. Đây là một ưu điểm của mô hình Lasso nên có thể được ứng dụng trong việc lựa chọn các feature.

Hồi quy Lasso với thư viện Scikitlearn:

```python
# Import các thư viện
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.linear_model import Lasso
from sklearn.model_selection import GridSearchCV

"""
Khởi tạo pipeline gồm 2 bước:
- 'scaler' để chuẩn hoá đầu vào với kỹ thuật chuẩn hóa min-max
- 'model' là tham số chọn mô hình huấn luyện với Lasso
"""
pipeline = Pipeline([
                    ('scaler', StandardScaler()),
                    ('model', Lasso())])

# GridSearch mô hình trên không gian tham số alpha
search = GridSearchCV(pipeline,
                      # Tham số alpha từ 0.01->0.1 huấn luyện mô hình
                      {'model__alpha':np.arange(0.01, 0.1, 0.05)},
                      cv = 3, # số validation trên tập kiểm tra
                      # trung bình tổng bình phương phần dư
                      scoring="neg_mean_squared_error",
                      verbose = 3
                     )
# Huấn luyện mô hình với tham số được chọn
search.fit(X, y)
print(search.best_estimator_)
print('Best core: ', search.best_score_)
```

```
Fitting 3 folds for each of 2 candidates, totalling 6 fits
[CV 1/3] END .................model__alpha=0.01;, score=-0.003 total time=   0.0s
[CV 2/3] END .................model__alpha=0.01;, score=-0.000 total time=   0.0s
[CV 3/3] END .................model__alpha=0.01;, score=-0.000 total time=   0.0s
[CV 1/3] END model__alpha=0.060000000000000005;, score=-0.014 total time=   0.0s
[CV 2/3] END model__alpha=0.060000000000000005;, score=-0.001 total time=   0.0s
[CV 3/3] END model__alpha=0.060000000000000005;, score=-0.010 total time=   0.0s
Pipeline(steps=[('scaler', StandardScaler()), ('model', Lasso(alpha=0.01))])
Best core:  -0.001067384584706267
```

Duyệt giá trị α cho mô hình Lasso bằng Grid.

	0	1
0	1.000000	-1.000000
1	-1.000000	1.000000

```python
print(search.best_estimator_.named_steps.model.coef_)
print(search.best_estimator_.named_steps.model.intercept_)
```

```
[ 0.05275291 -0.        ]
[1.59307692]
```

Feature thứ hai bị triệt tiêu.

Một số đánh giá

MÔ HÌNH HỒI QUY RIDGE VÀ LASSO	
Ưu điểm	**Nhược điểm**
Sử dụng với bảng chứa các thuộc tính có độ tương quan cao	Làm tăng giá trị của tham số Bias trong mô hình
Tránh xảy ra hiện tượng overfitting	Phải tìm kiếm giá trị tham số α
Tự động tìm kiếm các thuộc tính quan trọng	

Nội dung

1. Một số vấn đề trong việc xây dựng mô hình
2. Mô hình hồi quy Ridge và Lasso
3. Tổng kết

Kiến thức quan trọng

- Nắm rõ các khái niệm cơ bản
  - Mô hình hồi quy Ridge và Lasso.
  - Hiểu được ý nghĩa và mục đích của các mô hình hồi quy Ridge và Lasso.
- Hiểu rõ quá trình xây dựng mô hình
  - Các kiến thức liên quan tới hàm mất mát và việc tìm nghiệm tối ưu của mô hình Ridge và Lasso.
  - Nắm rõ ưu nhược điểm của mô hình Ridge và Lasso.
- 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.**
- **Tham khảo về mô hình Elastic Net và đọc trước tài liệu về mô hình hồi quy Logistic.**
- **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