Lab04: Linear regression

Courses `MTH00051`: Toán ứng dụng và thống kê
`18CLC6`, `FIT - HCMUS`.
`04/09/2020`

Đây là đồ án cá nhân, do một thành viên thực hiện:

- `18127231`: Đoàn Đình Toàn (GitHub: @t3bol90)

Về đồ án này:

Đề bài ở::
https://courses.ctda.hcmus.edu.vn/mod/resource/view.php?id=22047

Giới thiệu:

File **"wine.csv"** là cơ sở dữ liệu đánh giá chất lượng của 1200 chai rượu vang theo thang điểm 1 - 10 dựa trên 11 tính chất khác nhau. (File đính kèm trong file zip).

Yêu cầu:

1. Xây dựng mô hình đánh giá chất lượng rượu sử dụng phương pháp hồi quy tuyến tính.

```
a. Sử dụng toàn bộ 11 đặc trưng đề bài cung cấp.

b. Sử dụng duy nhất 1 đặc trưng cho kết quả tốt nhất. (Gợi ý: Phương pháp
Cross Validation)

c. Xây dựng một mô hình của riêng bạn cho kết quả tốt nhất.
```

Môi trường thực hiện

`Python 3.7` với `Jupyter notebook` và chạy hoàn toàn ở `local`.

Mức độ hoàn thành:

Yêu cầu	Mức độ hoàn thành
Xây dựng mô hình hồi qui tuyến tính với 11 đặc trưng (features)	100%
Xây dựng mô hình hồi qui tuyến tính với 1 đặc trưng và chỉ ra feature có kết quả tốt nhất (sử dụng phương pháp Cross Validation)	100%
Xây dựng mô hình riêng, sử dụng nhiều phương pháp và đạt kết quả tốt nhất	100%

Ý tưởng thực hiện & mô tả:

Sau khi đọc file đầu vào, ta có thể thấy như sau:

Các features lần lượt là:fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol (11 features). Output mà ta cần predict là quality - có giá trị trong khoảng 0-10 (nhưng trong data có giá trị từ 3.0 tới 8.0).

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1194	7.0	0.745	0.12	1.8	0.114	15.0	64	0.99588	3.22	0.59	9.5	6
1195	6.2	0.430	0.22	1.8	0.078	21.0	56	0.99633	3.52	0.60	9.5	6
1196	7.9	0.580	0.23	2.3	0.076	23.0	94	0.99686	3.21	0.58	9.5	6
1197	7.7	0.570	0.21	1.5	0.069	4.0	9	0.99458	3.16	0.54	9.8	6
1198	7.7	0.260	0.26	2.0	0.052	19.0	77	0.99510	3.15	0.79	10.9	6

1199 rows × 12 columns

Input data có 1199 dòng và 12 cột (11 features và 1 label)

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000	1199.000000
mean	8.625271	0.519133	0.293353	2.564470	0.089266	15.242702	46.884070	0.997059	3.298582	0.665738	10.383069	5.664721
std	1.781795	0.179208	0.196751	1.264441	0.048310	10.210406	33.949177	0.001878	0.156161	0.175921	1.091891	0.809593
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.300000	0.390000	0.120000	1.900000	0.071000	7.000000	21.000000	0.996000	3.195000	0.560000	9.500000	5.000000
50%	8.300000	0.500000	0.290000	2.200000	0.080000	13.000000	38.000000	0.997020	3.300000	0.620000	10.000000	6.000000
75%	9.600000	0.630000	0.450000	2.700000	0.092000	21.000000	63.000000	0.998175	3.390000	0.735000	11.000000	6.000000
max	15.900000	1.330000	1.000000	15.500000	0.611000	68.000000	289.000000	1.003200	3.900000	2.000000	14.900000	8.000000

Mô tả data

```
RangeIndex: 1199 entries, 0 to 1198
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1199 non-null   float64
 1   volatile acidity      1199 non-null   float64
 2   citric acid           1199 non-null   float64
 3   residual sugar        1199 non-null   float64
 4   chlorides             1199 non-null   float64
 5   free sulfur dioxide   1199 non-null   float64
 6   total sulfur dioxide  1199 non-null   int64
 7   density               1199 non-null   float64
 8   pH                    1199 non-null   float64
 9   sulphates             1199 non-null   float64
 10  alcohol               1199 non-null   float64
 11  quality               1199 non-null   int64
dtypes: float64(10), int64(2)
memory usage: 112.5 KB
```

Kiểm tra có null không và kiểu dữ liệu của từng dòng

Trong bài làm, em sẽ sử dụng mô hình hồi qui tuyến tính, gọi thẳng từ thư viện `skitlearn`, theo công thức sau [1]:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_\rho X_\rho + \sigma(Y)$$

Với β_0 còn gọi là intercept và $\beta_1 \ldots \beta_n$ là hệ số tương quan tuyến tính (correlation coefficient).

Hàm loss/error được dụng là hàm `mean square error` được chuyển hóa từ `least square error` bằng cách lấy trung bình lại [2]:

$$\arg\min_\beta MSE = \arg\min_\beta \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y - X\beta)^2}_{MSE} = \arg\min_\beta \underbrace{\sum_{i=1}^{N}(y - X\beta)^2}_{\text{sum of least squares}}$$

Cũng chính là hàm `mean_squared_error` trong `skitlearn.metrics`.

a. Sử dụng toàn bộ 11 đặc trưng đề bài cung cấp.

Ở đây sau khi tách cột 'quality' ra khỏi data làm label thì ta có được phần còn lại. Đặt hết phần còn lại vào train thì ta sẽ có được mô hình hồi qui tuyến tính với 11 đặc trưng:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.4	0.700	0.00	1.9	0.076	11.0	34	0.99780	3.51	0.56	9.4
1	7.8	0.880	0.00	2.6	0.098	25.0	67	0.99680	3.20	0.68	9.8
2	7.8	0.760	0.04	2.3	0.092	15.0	54	0.99700	3.26	0.65	9.8
3	11.2	0.280	0.56	1.9	0.075	17.0	60	0.99800	3.16	0.58	9.8
4	7.4	0.700	0.00	1.9	0.076	11.0	34	0.99780	3.51	0.56	9.4
...	...	...	...	...	...	...	...	...	...	...	...
1194	7.0	0.745	0.12	1.8	0.114	15.0	64	0.99588	3.22	0.59	9.5
1195	6.2	0.430	0.22	1.8	0.078	21.0	56	0.99633	3.52	0.60	9.5
1196	7.9	0.580	0.23	2.3	0.076	23.0	94	0.99686	3.21	0.58	9.5
1197	7.7	0.570	0.21	1.5	0.069	4.0	9	0.99458	3.16	0.54	9.8
1198	7.7	0.260	0.26	2.0	0.052	19.0	77	0.99510	3.15	0.79	10.9

1199 rows × 11 columns

Data sau khi bỏ cột label

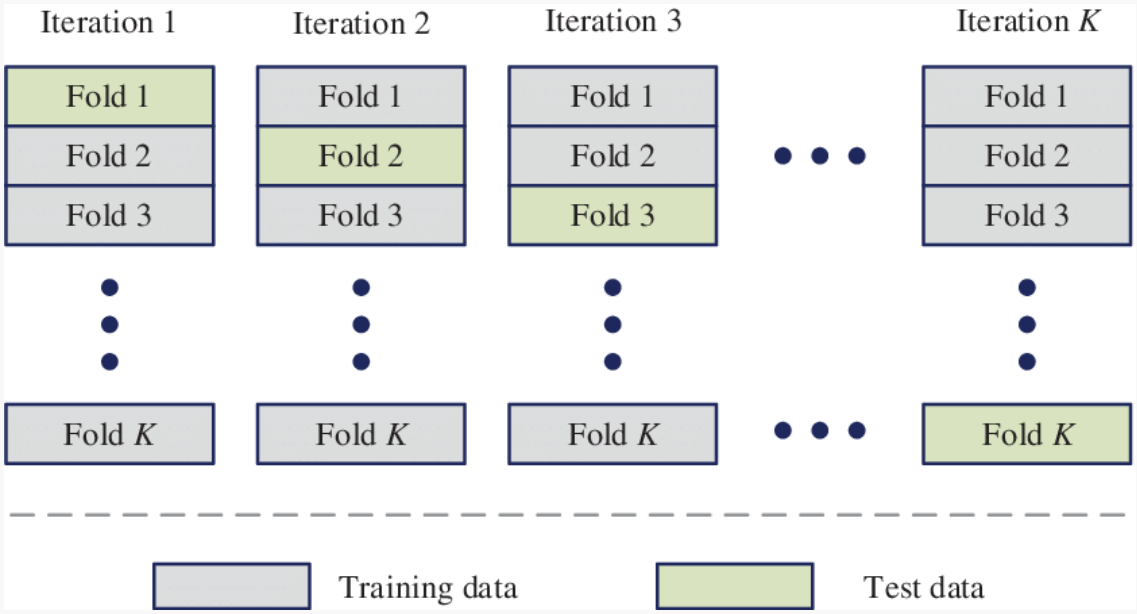
Với mô hình này, ta có được kết quả như phía dưới:

```
Mean squared error: 0.4146602076820373
      Actual  Predicted
554        5   6.086378
265        7   6.133711
962        5   5.383062
455        8   6.647937
805        7   6.889903
971        6   6.361269
555        5   6.086378
238        6   5.086802
909        6   6.246939
914        6   6.246939
338        6   6.125465
64         5   5.622773
81         5   5.098500
580        5   5.389785
130        5   4.717030
886        6   5.467497
452        6   5.474122
676        6   5.596847
496        6   5.131519
796        5   5.456793
1144       5   5.684587
142        6   7.104504
367        5   5.176109
193        5   5.404825
160        5   5.079523
```

`Model 11 features có sai số bình phương trung bình là 0.4146602076820373`

b. Sử dụng duy nhất 1 đặc trưng cho kết quả tốt nhất. (Gợi ý: Phương pháp Cross Validation)

Sử dụng phương pháp Cross validation (với hàm KFold trong thư viện), lần lượt chia các bộ train/test thành các phần và hoán đổi như bên dưới:



Mô phỏng phương pháp K Flod Cross Validation [3] (Trong bài làm em chọn K = 10)**

Sau khi lấy giá trị trung bình của các lần iteration, ta có được bảng error của từng model ứng với mỗi feature được chọn:

	Feature	Error
0	alcohol	0.500657
1	chlorides	0.665687
2	citric acid	0.637686
3	density	0.652420
4	fixed acidity	0.661292
5	free sulfur dioxide	0.668477
6	pH	0.671821
7	residual sugar	0.672908
8	sulphates	0.649935
9	total sulfur dioxide	0.642599
10	volatile acidity	0.575691

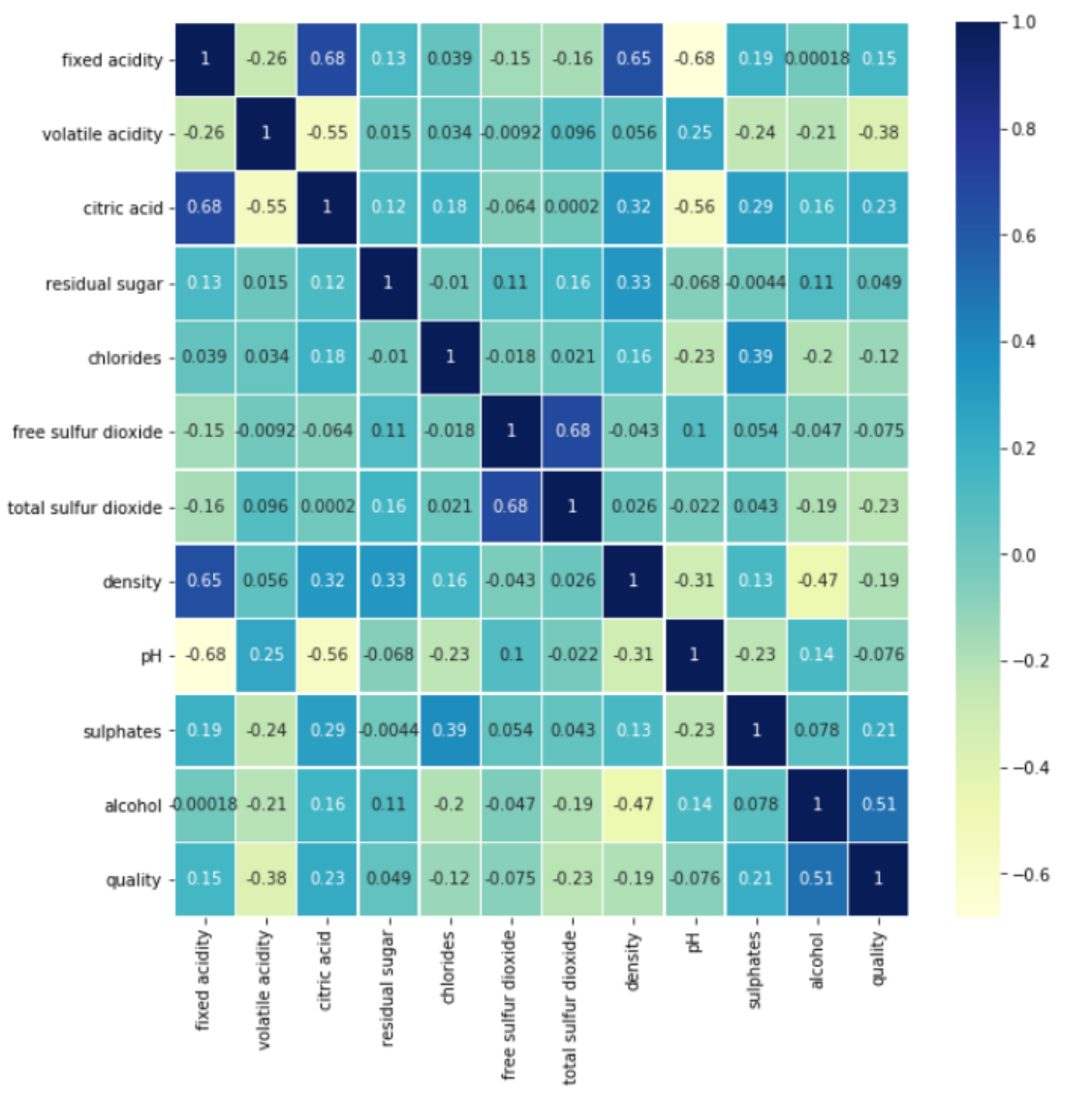
Mô hình ứng với từng feature và sai số

Ở đây ta có thể đưa ra kết luận rằng đối với môi hình đơn lẻ thì mô hình ứng với feature `alcohol` có kết quả tốt nhất với sai số thấp nhất.

c. Xây dựng một mô hình của riêng bạn cho kết quả tốt nhất.

Ở đây em sẽ sử dụng phương pháp Correlations (Tương quan) dựa trên coefficient. Correlation là một phương pháp của giải tích để khảo sát sự liên quan giữa hai trường/đặc điểm trên tập (ví dụ như chiều cao và cân nặng, giá vàng và gía gạo,...). Dựa vào sự hệ số tương quan, ta có thể đánh giá được sơ bộ rằng feature nào ảnh hưởng tới model của mình và ảnh hưởng nhiều hay ít. Hệ số tương quan của r với r là bằng 1 và các trường khác với nhau thì nhỏ hơn 1. Càng gần 1 thì càng liên quan còn càng gần -1 thì càng không liên quan [4].

Dưới đây là biểu đồ tương quan giữa các features:



Biểu đồ tương quan

Tiến hành chọn lọc ra những features có độ tương quan lớn hơn một ngưỡng $alpha$ nào đó, sau nhiều lần điều chỉnh thì em có được $alpha = 0.1$ có được kết quả tốt nhất và tốt hơn các model ở câu a và b.

```
correlations = df.corr()['quality'].drop('quality')
def get_features(correlation_threshold):
    abs_corrs = correlations.abs()
    high_correlations = abs_corrs[abs_corrs >
correlation_threshold].index.values.tolist()
    return high_correlations
```

Các features được chọn bao gồm: 'fixed acidity', 'volatile acidity', 'citric acid', 'chlorides', 'total sulfur dioxide', 'density', 'sulphates', 'alcohol'.

Các features trên có được thông qua việc khảo sát hai biến ngẫu nhiên với độ tin cậy 0.1 (sử dụng kiến thức xác suất thống kê học ở kỳ trước):

	coef	std err	t	P>|t|	[0.025	0.975]
fixed acidity	0.0355	0.016	2.250	0.025	0.005	0.066
volatile acidity	-1.1633	0.133	-8.725	0.000	-1.425	-0.902
citric acid	-0.2766	0.167	-1.660	0.097	-0.604	0.050
chlorides	-1.4498	0.454	-3.195	0.001	-2.340	-0.559
total sulfur dioxide	-0.0027	0.001	-4.604	0.000	-0.004	-0.002
density	2.6326	0.268	9.831	0.000	2.107	3.158
sulphates	0.7334	0.122	6.007	0.000	0.494	0.973
alcohol	0.3070	0.019	16.437	0.000	0.270	0.344

Bảng tổng kết model

Model này sau khi train thì có sai số nhỏ hơn: 0.36849063618742633.

```
# Tính loss
y_pred = model.predict(X_test)
rmse = mser(y_test, y_pred)
print("Error:", rmse)
```

```
Error: 0.36849063618742633
```

Kết quả của model tự tạo

Kết luận:

Qua bài lab lần này, em có được nhiều kiến thức hơn về việc ứng dụng những gì mình đã học về vào thực tế là xây dựng mô hình hồi qui tuyến tính để dự đoán trên tập dữ liệu.

Do có nhiều đồ án khác bủa vây xung quanh nên thời gian dành cho đồ án này cũng không đủ nhiều để phát triển sâu thêm. Trong giới hạn đồ án, em cũng đã thử các phương pháp có thể làm và tham khảo nhiều bài viết/tài liệu khác và đã có references bên dưới. Ngoài các hướng dẫn trong hướng dẫn đồ án, em cũng đã tìm hiểu thêm được một phương pháp khác để cải tiến và đánh giá mô hình là `Correlations`, ngoài ra còn thấy được sự vận dụng của kiến thức môn học kỳ

trước là môn Xác Suất Thống Kê (tổng kết 10.0 nhưng cũng không hiểu là ứng dụng ở đâu, nhưng giờ thì biết thêm một ứng dụng rồi).

Để hoàn thiện bài lab này, không thể kể đến các thầy cô giảng dạy cùng đội ngũ trợ giảng nhiệt tình, tận tâm với tụi em ở môn học này, chúc mọi người sức khỏe và niềm vui trong công việc giảng dạy tại HCMUS.

References:

Have a Great Day

---

1. https://towardsdatascience.com/linear-regression-and-a-quality-bottle-of-wine-b053ab768a53 ↩
2. https://stats.stackexchange.com/questions/297957/is-least-square-error-related-to-mean-square-error/297968 ↩
3. https://www.researchgate.net/figure/K-fold-cross-validation-method_fig2_331209203 ↩
4. https://www.surveysystem.com/correlation.htm ↩