Vietnam National University – Ho Chi Minh city

University of Science

Faculty of Information Technology

BÁO CÁO ĐỒ ÁN TOÁN ỨNG DỤNG VÀ THỐNG KÊ

APPLIED MATHEMATICS AND STATISTICS

Ho Chi Minh, 2021

Vietnam National University – Ho Chi Minh city

University of Science

Faculty of Information Technology

BÁO CÁO ĐỒ ÁN 3

2020 – 2021

APPLIED MATHEMATICS AND STATISTICS

Lớp: 19CLC7
Giáo viên hướng dẫn: Phan Thị Phương Uyên

STT	MSSV	Họ tên	Email
1	19127017	Trương Gia Đạt	19127017@student.hcmus.edu.vn

Ho Chi Minh, 2021

Table of Contents

Giới thiệu – Linear Regression

1. Tập dữ liệu sử dụng: $\boldsymbol{wine.csv}$

2. Mô hình

- Xây dựng mô hình hồi quy tuyến tính theo công thức $\boldsymbol{Ax} = \boldsymbol{b}$ với:
  - A là ma trận dữ liệu các tính chất rượu có kích thước $m \times n (m > n)$.
  - b là dữ liệu vector cột có kích thước m.

- Với mô hình trên, ta tìm nghiệm $\boldsymbol{x}$ của mô hình theo công thức $\hat{\boldsymbol{x}} = \boldsymbol{A}^{\dagger} \cdot \boldsymbol{b}$. Khi đó, mô hình sẽ đi qua gốc tọa độ → bị hạn chế.

⇒ Để cho mô hình có thể linh hoạt ta dựng mô hình theo công thức $Ax + b_0 = b$ hay $y = Ax + b_0$ để mô hình có thể tịnh tiến trên trục đồ thị.

Câu a: Dựng mô hình trên 11 tính chất

1. Thư viện sklearn

- Trong thư viện này, ta sử dụng sklearn.linear_model.LinearRegression để fit các tham số A và b vào mô hình và ta lấy ra các thuộc tính cần thiết của mô hình Linear Regression gồm:
  - ❖ coef__ : $\hat{x}$
  - ❖ intercept__ : b_0

2. Kết quả Demo
   - $\hat{x}$

```
[ 4.79658267e-02 -1.06797380e+00 -2.68453927e-01  3.50267451e-02
 -1.59557504e+00  3.47539059e-03 -3.79299466e-03 -3.98102920e+01
 -2.40172280e-01  7.74368364e-01  2.69212248e-01]
```

   - b_0

```
43.2363757146901
```

   - $y = b_0 + Ax$

```
Model: y = 43.2363757146901 + [ 4.79658267e-02 -1.06797380e+00 -2.68453927e-01  3.50267451e-02
 -1.59557504e+00  3.47539059e-03 -3.79299466e-03 -3.98102920e+01
 -2.40172280e-01  7.74368364e-01  2.69212248e-01]*x
```

Câu b: Chọn tính chất tốt nhất – Cross Validation

- K-Fold Cross-Validation:
  - Xáo trộn dữ liệu (Shuffle).
  - Chia dataset thành **k** nhóm.
  - Huấn luyện mô hình và kiểm tra, đánh giá.
  - Tổng hợp hiệu quả qua các đánh giá.

- Trong thư viên sklearn, ta sử dụng hàm sklearn.model_selection.KFold để tự động chia tập dữ liệu ra làm k nhóm và split ra làm k bộ dữ liệu với bộ train/test khác nhau.
  - Xây dựng mô hình trên tập train, ta thu được $\hat{x}, b_0$.
  - Áp dụng mô hình đó lên tập test $A_{test} \cdot \hat{x} = b'$.
  - Tính sai số so với tập test $|b_{test} - b'|$. Kết quả trả về sẽ ra một ma trận có kích thước giống b_{test} và ta tính trung bình trên ma trận này để có được sai số của mô hình trên tập train/test tương ứng.
  - Chạy hết các tập train/test được split ra, tính trung bình các sai số này ra được sai số trung bình của mô hình dựa trên phương pháp Cross Validation.

- Kết quả Demo
  - Tính chất tốt nhất

```
Best attribute is alcohol
Attribute model: y = 1.7807151719965795 + [0.37403439]*x
Attribute error: 0.5677816514839591
```

  - Bảng sai số của các tính chất

fixed acidity	0.6792699030016914
volatile acidity	0.6095646479692426
citric acid	0.6629423418805286
residual sugar	0.6946287033554557
chlorides	0.6856326933421388
free sulfur dioxide	0.6905617876746643
total sulfur dioxide	0.6454500279547847
density	0.676138773050523
pH	0.6924633985999818
sulphates	0.6714887886705413
alcohol	0.5677816514839591

Câu c: Xây dựng mô hình riêng

- Chọn ra n tính chất tốt nhất và chạy `CrossValidation()`
- Chọn ra nhóm tính chất tốt nhất (sai số thấp nhất) → Xây dựng mô hình dựa trên nhóm tính chất này.
- Sử dụng `LinearRegression()` để tìm model của nhóm tính chất.
- Kết quả Demo
  - Nhóm tính chất tốt nhất

```
Best attributes are ['alcohol', 'volatile acidity', 'total sulfur dioxide', 'citric acid', 'sulphates', 'density', '
fixed acidity', 'chlorides', 'free sulfur dioxide']
Attributes model: y = 33.60217206120657 + [ 2.79189114e-01 -1.08542522e+00 -3.27611003e-03 -2.50107046e-01
  7.55280875e-01 -3.10636950e+01  5.89809754e-02 -1.44110179e+00
  2.85695898e-03]*x
Attributes error: 0.5096349244682594
```

  - Bảng sai số của từng nhóm tính chất (chỉ số index)

```
[10  1]                          0.5301898589914902
[10  1  6]                       0.5205563462625263
[10  1  6  2]                    0.5207943814011727
[10  1  6  2  9]                 0.5131897805883002
[10  1  6  2  9  7]              0.5137641119641264
[10  1  6  2  9  7  0]           0.5113084462973727
[10  1  6  2  9  7  0  4]        0.5100202164533685
[10  1  6  2  9  7  0  4  5]     0.5096349244682594
[10  1  6  2  9  7  0  4  5  8]  0.5097843456000299
```

References

- [0] https://scikit-learn.org/stable/modules/linear_model.html
- [1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html