Linear Regression Project

Thông tin

- **Họ và tên:** Bùi Vũ Hiếu Phụng
- **MSSV:** 18127185

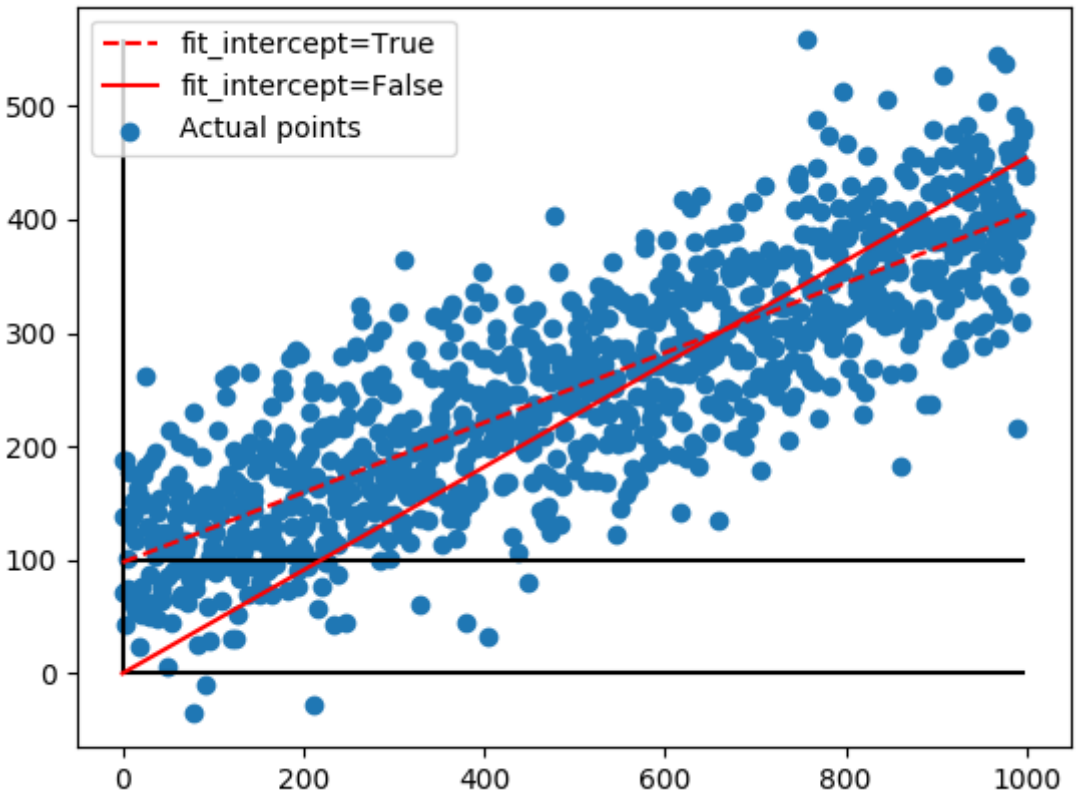
Chức năng	Tiến độ
Đọc dữ liệu và tách dữ liệu	100%
Cài đặt Linear Regression	100%
Cài đặt Cross Validation	100%
Câu a: Xây dựng mô hình trên 11 tính chất	100%
Câu b: Chọn tính chất tốt nhất và dựng mô hình	100%
Câu c: Xây dựng mô hình riêng (Chọn top x tính chất tốt nhất)	100%
Báo cáo	100%

Mô tả

Linear Regression

- Xây dựng mô hình đơn giản theo công thức: $Ax = b$. Trong đó
  - A là ma trận dữ liệu (số liệu) các tính chất rượu
  - b là label của mỗi dòng dữ liệu trong ma trận A đó, được thể hiện dưới dạng đánh giá (xếp loại) rượu
- Với mô hình trên, ta có thể tìm mô hình theo công thức: $\hat{x} = A^{\dagger} \cdot b$. Khi đó, mô hình sẽ đi qua gốc tọa độ của đồ thị $\rightarrow$ bị hạn chế
  $\Rightarrow$ Chọn mô hình theo công thức sau $Ax + b_0 = b$, mô hình khi ấy có thể tịnh tiến được trên trục tung để tăng sự linh hoạt

  Hình minh họa: *Ref: [0]*



- Bộ thư viện `sklearn` có công cụ để dựng mô hình hồi quy tuyến tính với công thức đã chọn, đó là `sklearn.linear_model.LinearRegression`. Sau khi fit các tham số bao gồm A và b vào mô hình, rất nhiều thuộc tính của mô hình Linear Regression được sinh ra, trong đó có:
  - `coef__`: $\hat{x}$
  - `intercept__` : b_0

  Gán chúng vào công thức trên, ta được mô hình hồi quy tuyến tính cần tìm.

Cross Validation

- Ở đây dùng K-Fold Cross-validation
  - Xáo trộn dữ liệu (*optional*)
  - Chia dataset thành **k** nhóm
  - Với mỗi nhóm:
    - Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
    - Các nhóm còn lại được sử dụng để huấn luyện mô hình
    - Huấn luyện mô hình
    - Đánh giá
  - Tổng hợp hiệu quả qua các đánh giá
- Bộ thư viện `sklearn` có 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 được $\hat{x}$, b_0
  - Áp dụng mô hình đó lên tập test: $A_{test} \cdot \hat{x} = b'$
  - Tính sai số so với label của tập test: $|b' - b_{test}|$. Kết quả tìm được sẽ là một ma trận có shape giống b_{test}, Khi đó ta tính trung bình của ma trận này, ta được sai số của mô hình trên tập train/test đó
  - Chạy hết tất cả các tập train/test được split ra ở trên, tính trung bình các sai số này ta được sai số trung bình của mô hình dựa trên phương pháp Cross Validation
- Số k được chọn cho K-Fold thường là **10** *Ref: [1]*

Câu a. Dựng mô hình trên 11 tính chất

- Tách dữ liệu đọc từ `wine.csv` thành bộ dữ liệu và bộ label
- Dùng `LinearRegression()` đã cài đặt để tìm model trên tất cả các tính chất

- Kết quả chạy được:
  - $\hat{x}$

```
array([ 4.75247531e-02, -1.06874258e+00, -2.68710829e-01,  3.49742662e-02,
       -1.59729560e+00,  3.48788138e-03, -3.79835506e-03, -3.94690810e+01,
       -2.45575908e-01,  7.73840794e-01,  2.69377496e-01])
```

  - b_0

```
42.91716245147436
```

  - Mô hình:

```
Model: A[ 4.75247531e-02 -1.06874258e+00 -2.68710829e-01  3.49742662e-02
 -1.59729560e+00  3.48788138e-03 -3.79835506e-03 -3.94690810e+01
 -2.45575908e-01  7.73840794e-01  2.69377496e-01] + 42.91716245147436 = b
```

Câu b. Chọn tính chất tốt nhất

- Đối với từng cột (tính chất), chạy `CrossValidation()` để tìm sai số của mô hình dựa trên mỗi tính chất
- Tìm tính chất có sai số bé nhất $\rightarrow$ Tính chất tốt nhất
- Chạy `LinearRegression()` để tìm model dựa trên tính chất này

- Kết quả chạy được:
  - Tính chất tốt nhất + Mô hình

```
Best property is alcohol
Model: A[0.37471047] + 1.7740758844499194 = b
CV error: 0.5689435008365984
```

  - Bảng sai số

Tính chất	Sai số
fixed acidity	0.6825140679808375
volatile acidity	0.6109359668341867
citric acid	0.6651332648312804
residual sugar	0.6976293712055988
chlorides	0.6885534089341065
free sulfur dioxide	0.6940545770844017
total sulfur dioxide	0.6480698378955821
density	0.6808205510681493
pH	0.6949644236500062
sulphates	0.6755085490249927
alcohol	0.5689435008365984

Câu c. Xây dựng mô hình riêng

- Chọn ra n tính chất tốt nhất (cho n chạy từ 2 đến 10) rồi chạy `CrossValidation()`
- Chọn ra bộ tính chất tốt nhất (sai số thấp nhất) → Các tính chất sẽ được chọn để dựng mô hình
- Chạy `LinearRegression()` để tìm model dựa trên các tính chất này

- Kết quả chạy được:
  - Các tính chất tốt nhất + Mô hình:

```
Best properties are: ['alcohol', 'volatile acidity', 'total sulfur dioxide', 'citric acid', 'sulphates', 'density', 'fixed acidity',
  'chlorides', 'free sulfur dioxide']
Model: A[ 2.79228621e-01 -1.08519171e+00 -3.27609434e-03 -2.50067753e-01
  7.55341209e-01 -3.10768383e+01  5.89618605e-02 -1.44151674e+00
  2.85670572e-03] + 33.61492533087279 = b
```

  - Bảng sai số

Các tính chất (index)	Sai số
10, 1	0.5309400304379519
10, 1, 6	0.521684703694745
10, 1, 6, 2	0.5217960880728357
10, 1, 6, 2, 9	0.5138102252985199
10, 1, 6, 2, 9, 7	0.5151620991110236
10, 1, 6, 2, 9, 7, 0	0.5126608137929859
10, 1, 6, 2, 9, 7, 0, 4	0.5107858526794741
10, 1, 6, 2, 9, 7, 0, 4, 5 (chọn)	0.5104161230846668
10, 1, 6, 2, 9, 7, 0, 4, 5, 8	0.5104894102915315

- Sai số của mô hình tự xây dựng và của mô hình 11 tính chất:
  - Mô hình tự xây dựng: 0.5104161230846668
  - Mô hình 11 tính chất: 0.5094507964775307

References

- [0] https://stackoverflow.com/questions/46779605/in-the-linearregression-method-in-sklearn-what-exactly-is-the-fit-intercept-par
- [1] https://stats.stackexchange.com/questions/27730/choice-of-k-in-k-fold-cross-validation?fbclid=IwAR02IS7qEz1KGSTgtUObQYA14JJa-k1FenGUuKOBXkoAXwEMDm6L3Pa9O20