Đại học Khoa học Tự nhiên – ĐHQG TPHCM

Môn: Toán ứng dụng và thống kê

BÁO CÁO ĐỒ ÁN 3: DATA FITTING

Họ tên: Phan Xuân Nam

Mã số sinh viên: 20127247

Thông tin đồ án

Cho dataset *wine.csv*, xây dựng mô hình đánh giá chất lượng rượu sử dụng phương pháp quy hồi tuyến tính.

Trong dataset, có tất cả 11 tính chất khác nhau:

Tính chất	Nội dung
fixed acidity	Nồng độ bay hơi
volatile acidity	Nồng độ axit axetic
citric acid	Nồng độ axit citric
residual sugar	Nồng độ đường còn dư
chlorides	Nồng độ muối
free sulfur dioxide	Nồng độ SO_2 tự do
total sulfur dioxide	Nồng độ SO_2 dạng tự do và liên kết
density	Độ đặc / Khối lượng riêng
pH	Nồng độ pH
sulphates	SO_4^{2-}
alcohol	Nồng độ rượu

Dựa vào các đặc trưng trên, ta so với "quality" để đánh giá chất lượng rượu

Các nội dung thực hiện

Ta sẽ sử dụng các thư viện sau:

- pandas: thực hiện các thao tác và đọc file .csv
- numpy: lưu các số liệu ở dạng array và thực hiện các thao tác trên đó
- copy: lưu các dữ liệu ở một biến, tránh bị tình trạng biến bị lưu thay đổi

Đọc dataset từ wine.csv và lưu vào biến **df**

```
df = pd.read_csv('wine.csv', sep = ";")
```

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

- Ta sử dụng công thức: $y = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{11} x_{11}$
- Lưu cột "quality" vào ma trận Y, các cột còn lại vào ma trận X
- Với ma trận X, ta sẽ tạo thêm ma trận gồm các số 1 và gán thêm vào cột đầu tiên.
- Định dạng X và Y về numpy.array để thực hiện các thao tác ma trận
- Sau khi thực hiện xong ta sẽ có kết quả như sau:
    - Với X:

```
array([[ 1. , 7.4 , 0.7 , ..., 3.51, 0.56, 9.4 ],
       [ 1. , 7.8 , 0.88, ..., 3.2 , 0.68, 9.8 ],
       [ 1. , 7.8 , 0.76, ..., 3.26, 0.65, 9.8 ], ...,
       [ 1. , 7.9 , 0.58, ..., 3.21, 0.58, 9.5 ],
       [ 1. , 7.7 , 0.57, ..., 3.16, 0.54, 9.8 ],
       [ 1. , 7.7 , 0.26, ..., 3.15, 0.79, 10.9 ]])
```

    - Với Y:

```
array([5, 5, 5, ..., 6, 6, 6], dtype=int64)
```

- Sử dụng hàm getAb(a, b) với tham số truyền vào là ma trận X và ma trận Y để biến đổi ma trận X về dạng chuyển vị (ở đây đã có sẵn nên không cần) và dạng cột của ma trận Y

```
def getAb(a, b):
    matrixA = a
    matrixB = b.reshape(len(b), 1)
    return matrixA, matrixB
```

- Tính giá trị của θ thông qua công thức $(A^T \cdot A)^{-1}(A^T b)$

```
thetaMatrix = np.dot(np.dot(np.linalg.inv(np.dot(a.T, a)), a.T), b)
```

- Tính R thông qua công thức $\|r\| = \|A\theta - b\|$

```
r = np.linalg.norm(a@thetaMatrix-b)
```

- Sau khi tính, ta sẽ in đáp án ra:

Value of R is: 22.09472

B. K-fold Cross Validation

Ý tưởng của phương pháp bắt đầu từ việc xét riêng từng đặc trưng và với mỗi đặc trưng, ta sẽ chia ra làm các k nhóm để test và train.

Ở dataset này, mỗi đặc trưng có tổng cộng 1199 dòng. Ta sẽ lựa k = 11 vì số lượng nhóm chia ra sẽ đều, tránh trường hợp một nhóm thiếu. Quy trình thực hiện chung của phương pháp Cross Validation là:

- (tùy chọn): sắp xếp ngẫu nhiên các dữ liệu
- Chia dữ liệu ra thành k nhóm
- Với mỗi nhóm trong trường dữ liệu:
    - Lấy 1 nhóm làm test và đưa ra model chung cho nhóm
    - k - 1 nhóm còn lại làm nhóm training
    - Với mỗi k - 1 nhóm còn lại, đem vào model đã test và đưa ra điểm số đánh giá r
    - Khi đã đánh giá hết k - 1 nhóm, ta sẽ lấy trung bình của tổng điểm đó và lưu vào một list lưu điểm trung bình
- Sau khi thực hiện hết k nhóm, ta sẽ lấy list điểm trung bình và lấy ra điểm có giá trị r cao nhất
- Với mỗi đặc trưng sẽ cho ra một giá trị r khác nhau. Để lựa chọn đặc trưng tốt nhất, ta sẽ lấy r nhỏ nhất.

Lý do chúng ta chọn r nhỏ nhất vì theo hệ số R^2, tức là hệ số dùng để đo mức độ phù hợp của mô hình nghiên cứu, có công thức tính là:

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

RSS là hệ số tổng bình phương phần dư

TSS là hệ số bình phương độ chênh lệch với trung bình

Nếu $\frac{RSS}{TSS}$ càng lớn thì R^2 sẽ càng bé, thì độ chính xác càng thấp và ngược lại. Trong môn học này, do chưa được giảng tới hệ số R^2 nên ta chỉ sử dụng hệ số $||r||$ để xét. Sau khi tính toán, ta có đáp số như sau:

```
Property: fixed acidity - 14.56273

Property: volatile acidity - 24.76224

Property: citric acid - 33.13839

Property: residual sugar - 27.83790

Property: chlorides - 34.23822

Property: free sulfur dioxide - 34.90090

Property: total sulfur dioxide - 38.66093

Property: density - 9.47490

Property: pH - 10.26924

Property: sulphates - 19.43794

Property: alcohol - 7.88443

Best property according to R_score is: alcohol with 7.88443
```

Theo như console thì đặc trưng tốt nhất là alcohol với số r = 7.8843.

C. Một mô hình mà mình cho là tốt nhất

Theo em, mô hình tốt nhất mà mình được học chính là sử dụng Log – Lin vì một lượng X tăng thì sẽ dẫn đến một lượng $\log(y) \times \hat{\theta}$ lần. Nghĩa là, giá trị của Y sẽ được nhân với e^{Θ}. Chung quy lại, sử dụng mô hình này sẽ cho ra kết quả $||r||$ nhỏ hơn ở **câu A**.

Cách thực hiện:

- Biến đổi các giá trị ở ma trận Y về dạng log tự nhiên (ln)
- Thực hiện tương tự như câu A, sẽ cho ra đáp án

```
Value of R using Log-Linear model is: 3.99213
```