ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

|ĐỒ ÁN 2|

Môn: Khai thác dữ liệu và ứng dụng

Chủ đề:

Triển khai mô hình phân loại bệnh ung thư vú

Lecturer: Nguyễn Ngọc Thành
TA: Nguyễn Thái Vũ

Thành phố Hồ Chí Minh, năm 2022.

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

|ĐỒ ÁN 2|

Môn: Khai thác dữ liệu và ứng dụng

Chủ đề:

Triển khai mô hình phân loại bệnh ung thư vú

|Danh sách thành viên|
Nguyễn Ngọc Uyên Trang - 19127074
Kiều Hải Đăng – 19127347
Lã Minh Hiếu – 19127400

Thành phố Hồ Chí Minh, năm 2022.

Mục lục

1. Chuẩn bị dữ liệu

a. Làm sạch dữ liệu:

- Theo như dữ liệu gốc ban đầu, ta thấy có một cột trống ở phía cuối vì thế ta cần phải bỏ đi cột này.

Unnamed: 32
NaN
NaN
NaN
NaN
NaN
...
NaN
NaN
NaN
NaN
NaN

Hình 1: Cột trống

- Với trường dữ liệu có thuộc tính là ID, thì đây là thuộc tính duy nhất cho nên ta sẽ không dùng vào mô hình phân lớp mà ta sẽ bỏ nó đi.

id
842302
842517
84300903
84348301
84358402
...
926424
926682
926954
927241
92751

Hình 2: Cột ID

b. Tiền xử lý dữ liệu

- Với cột dữ liệu là dạng categorical như cột "Diagnosis" với hai giá trị M (malignant) và B (benign).
- Đây là cột dữ liệu với ý nghĩa là chuẩn đoán xem bệnh nhân có bị ung thư vú lành tính (B) hay ác tính (M).
- Ở đây nhóm em sẽ chuyển dạng hiện tại của cột này về với dạng 0 và 1; 0 cho B và 1 cho M.

c. Phần code cho việc chuẩn bị dữ liệu

```
def Preprocessing(df):

    # Mã hóa biến nhị phân
    df['diagnosis'] = np.where(df['diagnosis'] == 'M', 1, 0)

    df = df.drop(['Unnamed: 32', 'id'], inplace=True, axis=1)
```

Hình 3: Mã nguồn tiền xử lý dữ liệu

2. Huấn luyện mô hình

a. Chia tập dữ liệu để huấn luyện, kiểm thử

- Về việc chia tập dữ liệu để chuẩn bị cho việc chạy mô hình, nhóm em sẽ chia theo tỉ lệ 80:20; với tập huấn luyện là 80 và việc kiểm thử sẽ là 20.
- Nhìn chung việc chọn tỉ lệ sao cho đúng cũng rất khó nói, đứng ở quan điểm và góc nhìn của nhóm em thì nhóm em nhận thấy rằng 80 - 20 là một tỉ lệ hợp lý bởi vì thực tế thông thường em nhận thấy các mô hình học máy hay sử dụng tỉ lệ 90 - 10. Đó là tỉ lệ tốt nhưng rủi ro cũng khá cao vì thực tế chúng ta không thể biết được rằng dữ liệu có bị phân bố lệch lên trên hay lệch xuống dưới không. Chính vì vậy nhóm em quyết định chọn tỉ lệ 80 - 20 để phù hợp 3 yếu tố: tránh overfitting, tránh underfitting, hạn chế tỉ lệ rủi ro. Tất nhiên câu chuyện ai đúng ai sai còn dựa vào dữ liệu thực tế, độ lớn của dữ liệu và nhiều yếu tố khác. Ở góc độ này nhóm em chỉ muốn hạn chế rủi ro cho mô hình.

b. Tham số của mô hình (KNN)

- Ở bài lần này, nhóm chúng em dùng mô hình KNN để phân lớp.
- Đối với hệ số k trong mô hình KNN, sau khi thử và cũng có một số hiểu biết nên nhóm chúng em chọn hệ số k cho mô hình là 20.
- Hệ số k là dựa vào những dòng dữ liệu gần nhất với k dòng dữ liệu để có thể phân lớp và nhận ra dòng dữ liệu mới.
  - Ở đây, nhóm em thử với 4 giá trị k và có nhận xét sau:
  - Ở giá trị k = 5: thì độ chính xác của mô hình là thấp nhất.

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95        72
           1       0.89      0.95      0.92        42

    accuracy                           0.94       114
   macro avg       0.93      0.94      0.93       114
weighted avg       0.94      0.94      0.94       114

Training Score:
 97.58241758241758
Valid Score:
 93.85964912280701
```

Hình 4: Độ chính xác với k = 5

  - Khi tăng k = 10 thì độ chính xác cao hơn hẳn mô hình với k = 5.

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.97      0.96      0.97        72
           1       0.93      0.95      0.94        42

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114

Training Score:
 96.92307692307692
Valid Score:
 95.6140350877193
```

Hình 5: Độ chính xác với k = 10

- Tuy nhiên khi tăng k = 15 độ chính xác của mô hình lại thấp hơn mô hình với k = 10. Nhưng đối với tập training độ chính xác thấp hơn mô hình với k = 5 và có độ chính xác ở tập kiểm thử bằng nhau.

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.96      0.94      0.95        72
           1       0.91      0.93      0.92        42

    accuracy                           0.94       114
   macro avg       0.93      0.94      0.93       114
weighted avg       0.94      0.94      0.94       114

Training Score:
 95.82417582417582
Valid Score:
 93.85964912280701
```

Hình 6: Độ chính xác với k = 15

- Dừng lại ở k = 20, thì nhóm em thấy mô hình đạt được độ chính xác ở tập kiểm thử bằng với mô hình với k = 10 nhưng ở tập train thì lại thấp hơn.

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.96      0.97      0.97        72
           1       0.95      0.93      0.94        42

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114

Training Score:
 95.6043956043956
Valid Score:
 95.6140350877193
```

Hình 7: Độ chính xác với k = 20

→ Từ 4 giá trị k nhóm chúng em đã thử thì nhóm chúng em quyết định chọn tham số với k = 10 để phù hợp với mô hình và dữ liệu.

3. Báo cáo kết quả, độ chính xác.

```
Classification Report is:
              precision    recall  f1-score   support

           0       0.97      0.96      0.97        72
           1       0.93      0.95      0.94        42

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114

Training Score:
 96.92307692307692
Valid Score:
 95.6140350877193
```

Hình 8: Kết quả chạy mô hình với k = 10

- Với bảng báo cáo về sự phân lớp của mô hình:
  - Ta thấy độ chính xác với mô hình ở tập train là 96.9%, và độ chính xác ở tập kiểm thử là 95.61%.
- Nhận xét, nhóm chúng em thấy với tham số k lớn thì độ chính xác ở tập train giảm còn ở tập kiểm thử khá là bất thường (tăng/giảm). Nhóm chúng em nhận thấy rằng đối với tập train, thì có thể thấy nếu tăng k lên thì việc so khớp sự giống nhau giữa các dữ liệu càng khó hơn, cũng giống như việc học càng khó thì sẽ nhận biết và tiếp thu được nhiều thứ.

4. Link demo: https://www.youtube.com/watch?v=bcUGlbKe5j0