Đại học quốc gia TP.HCM
Trường đại học công nghệ thông tin

Ngành Khoa học máy tính

Môn học: CS116 - Lập trình Python cho máy học

Đồ án cuối kì

Sinh viên:
Nguyễn Quốc Trường
21521604

Giảng viên:
LT: Nguyễn Duy Khánh
TH: Chế Quang Huy

Contents

1 Introduction

- Đồ án này em thực hiện bài toán phân loại (Classification) trên dataset Breast Cancer Wisconsin, đây là một dataset dùng để xây dựng mô hình dự đoán dựa trên đặc trưng của ảnh.
- Khi thực hiện đồ án này, em đã tiến hành theo các bước:

  - Tải dataset
  - Thực hiện EDA
  - Chuẩn hóa (Normalize) dữ liệu trong dataset
  - Tiến hành huấn luyện các mô hình

- Em đã thực hiện huấn luyện dataset này trên hai mô hình khác nhau là Logistic Regression và SVM[1].

2 Exploratory Data Analysis (EDA)

2.1 Dataset

Breast Cancer Wisconsin (Diagnostic) Data Set
- Số lượng mẫu: Dataset chứa dữ liệu từ 569 mẫu.
- Đặc trưng (Features): Có 30 đặc trưng được sử dụng để mô tả các đặc điểm hình học của nhân của tế bào.

  - 10 đặc trưng được tính từ các giá trị thống kê của các tế bào bị biến đổi (mean, standard và worst).
  - 3 đặc trưng là kích thước trung bình, kích thước SE và kích thước tối đa của nhân tế bào.
  - 16 đặc trưng là giá trị của biểu thức "smoothness," "compactness," "concavity," "concave points," "symmetry," và "fractal dimension."

- Biến phụ thuộc (Target): Biến phụ thuộc là nhãn của tế bào, được gán nhãn là "malignant" (ác tính) hoặc "benign" (lành tính).
- Nguồn gốc: Dataset được thu thập từ Trung tâm Y học Bệnh viện Tuyến vú Wisconsin.
- Mục tiêu: Mục tiêu chính của dataset này là xây dựng mô hình dự đoán liệu một tế bào có tính ác tính hay lành tính dựa trên các đặc trưng hình ảnh của nó.

2.2 EDA

- Sau khi thực hiện việc tải dataset, em thực hiện bước phân tích khám phá dữ liệu (EDA). EDA là một quá trình khám phá dữ liệu và hiểu rõ hơn về dữ liệu. EDA có thể giúp chúng ta hiểu được các thuộc tính nào quan trọng nhất đối với việc phân tích, và các mẫu trong dataset có phân bố như thế nào.
- Các bước khi em thực hiện EDA:

  - Kiểm tra tổng quan dữ liệu.
  - Kiểm tra kích thước dữ liệu (số lượng mẫu, số lượng thuộc tính).
  - Kiểm tra mô tả thống kê của các thuộc tính (giá trị trung bình, độ lệch chuẩn, min-max, phân vị)
  - Kiểm tra các giá trị thiếu, giá trị null

- Kết quả kiểm tra cho thấy:

  - Dataset bao gồm 569 mẫu, mỗi mẫu có 9 thuộc tính.
  - Các thuộc tính trong dataset đều có giá trị số (trừ thuộc tính diagnosis).
  - Không có giá trị thiếu trong dataset.

- Sau khi thực hiện kiểm tra, em tiến hành thay thế lại dữ liệu trong cột diagnosis thành giá trị số ('M': 1, 'B': 0) và vẽ heatmap cho bộ dataset.
- Từ heatmap cho thấy, có mối tương quan mạnh mẽ giữa một số đặc trưng như bán kính hạt nhân, số điểm lõm, chu vi và diện tích với tính chất ung thư.

3 Proposed Methods

- Trong đồ án này, em thực hiện việc phân loại (Classification) bằng 2 mô hình khác nhau.

  - Logistic Regression là một thuật toán học máy được sử dụng để phân loại dữ liệu. Logistic regression hoạt động bằng cách xây dựng một mô hình toán học để mô tả mối quan hệ giữa các thuộc tính đầu vào và lớp đầu ra.

- SVM là một thuật toán học máy được sử dụng để phân loại dữ liệu. SVM hoạt động bằng cách tìm một siêu phẳng (hyperplane) có thể tách các mẫu của hai lớp ra một cách tốt nhất.

- Trước khi bắt đầu huấn luyện các mô hình, em đã thực hiện lần lượt các bước phân tích dữ liệu (EDA) và chuẩn hóa dữ liệu.

3.1 Normalize Data

- Trong đồ án, em sử dụng cả hai phương pháp chuẩn hóa dữ liệu là Robust Scaler và Standard Scaler.
- Robust Scaler và Standard Scaler là hai phương pháp thường được sử dụng để chuẩn hóa dữ liệu trong quá trình tiền xử lý dữ liệu cho các bài toán học máy. Cả hai phương pháp đều giúp giảm khoảng cách giữa các giá trị thuộc tính khác nhau, nhưng chúng có những cách tiếp cận khác nhau để xử lý các giá trị bất thường (outliers).
- Standard Scaler:

- Trừ giá trị trung bình của mỗi thuộc tính khỏi các giá trị của thuộc tính đó.
- Chia các giá trị đã trừ trung bình bằng độ lệch chuẩn của thuộc tính đó.
- Kết quả thu được là các giá trị chuẩn hóa có trung bình là 0 và độ lệch chuẩn là 1.

- Robust Scaler:

- Sử dụng median (trung vị) và interquartile range (IQR) để thay thế cho mean (trung bình) và standard deviation (độ lệch chuẩn).
- Trừ median của mỗi thuộc tính khỏi các giá trị của thuộc tính đó.
- Chia các giá trị đã trừ median bằng IQR của thuộc tính đó.
- IQR là khoảng cách giữa quartile thứ 3 (75%) và quartile thứ 1 (25%) của các giá trị trong thuộc tính.

3.2 Logistic Regression

- Logistic Regression là một mô hình phân loại được sử dụng để dự đoán khả năng một điểm dữ liệu thuộc về một lớp nhất định. Mô hình này dựa trên hàm logistic, là một hàm phi tuyến biến đổi xác suất từ một giá trị từ 0 đến 1.
- Mô hình logistic regression sử dụng một hàm logistic để biến đổi kết quả của mô hình linear regression thành một giá trị từ 0 đến 1. Hàm logistic được định nghĩa như sau:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

- Mô hình linear regression được định nghĩa như sau:

$$y = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b \tag{2}$$

- Đầu ra của mô hình linear regression là một giá trị số thực. Để biến đổi giá trị này thành một giá trị từ 0 đến 1, chúng ta sử dụng hàm logistic như sau:

$$p = \sigma(y) \tag{3}$$

- Tham số của mô hình logistic regression được tìm bằng phương pháp tối ưu hóa tối thiểu hóa hàm mất mát. Hàm mất mát thường được sử dụng là hàm cross-entropy:

$$L = -\sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{4}$$

3.3 Support vector machine (SVM)

- SVM là một mô hình phân loại được sử dụng để phân loại các điểm dữ liệu vào các lớp khác nhau. Mô hình này dựa trên khái niệm support vector, là các điểm dữ liệu nằm trên ranh giới giữa các lớp.
- Mô hình SVM tìm cách tìm ra một siêu phẳng trong không gian mà có thể phân tách các điểm dữ liệu của hai lớp một cách tối ưu. Siêu phẳng này phải có khoảng cách lớn nhất từ các điểm dữ liệu của hai lớp. Các điểm dữ liệu nằm trên siêu phẳng này được gọi là các support vector.
- Tham số của mô hình SVM được tìm bằng phương pháp tối ưu hóa tối thiểu hóa hàm mất mát. Hàm mất mát thường được sử dụng là hàm hinge loss:

$$L = \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) \tag{5}$$

3.4 Other methods

- Ngoài 2 mô hình ở trên được em thực hiện thử nghiệm trên tập Private Test, em còn chạy thử một số mô hình khác trên tập Public Test.

- *k-Nearest Neighbors (kNN)* là một thuật toán máy học phân loại dựa trên khoảng cách giữa các điểm dữ liệu. Để dự đoán nhãn của một điểm dữ liệu mới, kNN sẽ tìm k điểm dữ liệu gần nhất với điểm dữ liệu đó và sử dụng nhãn của k điểm dữ liệu này để dự đoán nhãn của điểm dữ liệu mới.

- *Random Forest* là một thuật toán máy học phân loại dựa trên kết hợp của nhiều decision tree. Mỗi decision tree trong random forest sẽ được đào tạo trên một tập dữ liệu ngẫu nhiên của tập dữ liệu ban đầu. Để dự đoán nhãn của một điểm dữ liệu mới, random forest sẽ sử dụng kết quả dự đoán của tất cả các decision tree trong tập hợp và sử dụng phương pháp bỏ phiếu đa số để đưa ra dự đoán cuối cùng.

- *Decision Tree* là một thuật toán máy học phân loại dựa trên việc phân chia dữ liệu thành các nhánh nhỏ hơn và nhỏ hơn. Mỗi nhánh sẽ được phân chia dựa trên một thuộc tính của dữ liệu. Để dự đoán nhãn của một điểm dữ liệu mới, decision tree sẽ bắt đầu từ đỉnh của cây và đi xuống theo các nhánh cho đến khi đến một nhánh có nhãn.

4 Experiment

4.1 Metrics

- Trong đồ án này, để đánh giá và so sánh hiệu suất của các mô hình, em sử dụng 2 độ đo là Accuracy và F1-score.
- Công thức Accuracy:

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total}} \tag{6}$$

- Công thức F1-score:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

với precision và recall được tính như sau:

$$precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{8}$$

$$recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{9}$$

- Trong đó:

- true positives (TP) là số lượng mẫu được dự đoán đúng là dương.
- true negatives (TN) là số lượng mẫu được dự đoán đúng là âm.
- false positives (FP) là số lượng mẫu được dự đoán là dương nhưng thực tế là âm.
- false negatives (FN) là số lượng mẫu được dự đoán là âm nhưng thực tế là dương.
- total là tổng số mẫu trong tập dữ liệu.

4.2 Evaluation

- Trong đồ án này, em sử dụng hai mô hình chính là Logistic Regression và SVM, thực hiện trên dataset được chuẩn hóa bằng hai phương pháp là Robust Scaler và Standard Scaler.
- Đối với mô hình SVM, em có sử dụng SVM mặc định và SVM với C = 0.1.
- *Table 1* thể hiện kết quả của từng phương pháp trên dataset Breast Cancer Wisconsin.

Model	LR+RC	LR+SC	SVM+RC	SVM+SC	SVM(C=0.1)+SC
Acc	86.7073	85.6098	83.4146	86.7073	85.6098
F1	86.6831	85.5861	83.3514	86.6672	85.5861

Table 1: *Kết quả đánh giá các mô hình (%)*

- Ngoài các phương pháp trên, em còn sử dụng một số phương pháp khác để thử nghiệm như kNN, Random Forest, Decision Tree... Nhưng kết quả thu được chưa được cao trên tập Public Test, do đó em không thực hiện thử nghiệm trên tập Private Test.

5 Conclusion

- Trước khi huấn luyện một mô hình máy học, ta phải trải qua các bước phân tích và tiền xử lý dữ liệu. Trong đồ án này, em đã thực hiện EDA và chuẩn hóa dữ liệu.
- Qua thực nghiệm, ta có thể dễ dàng nhận thấy, trong đồ án và qua các xử lý của em, mô hình hoạt động tốt nhất trên dataset Breast Cancer Wisconsin là Logistic Regression với phương pháp chuẩn hóa Robust Scaler.
- Trên thực tế, em có tìm hiểu và biết được một vài phương pháp kết hợp với các cách xử lý dữ liệu khác cho kết quả cao hơn. Tuy nhiên, em vẫn chưa cài đặt được.
- Em sẽ tiếp tục tìm hiểu và thử nghiệm để cải tiến đồ án này.
- Dường dẫn dưới đây là một số công việc liên quan đến đồ án mà em thực hiện thêm ngoài những thứ được thầy chỉ dẫn.

Source: https://github.com/lIlDrAgOoN/CS116.O11.KHTN

References

[1] Nello Cristiamini and J Shawe-Taylor. An introduction to support vector machines. *Cambridge university press*, 2000.