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
  - Thực hiện PCA [1]
  - Chuẩn hóa (Normalize) và làm sạch dữ liệu trong dataset
  - Tiến hành huấn luyện các mô hình

- Em đã thực hiện huấn luyện dataset này trên ba mô hình khác nhau là Logistic Regression, kNN và SVM.

2 Methods

Trong đồ án này, em thực hiện việc phân loại (Classification) bằng 3 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.
- kNN là một thuật toán học máy được sử dụng để phân loại dữ liệu. kNN hoạt động bằng cách tìm k mẫu gần nhất với mẫu cần phân loại và phân loại mẫu đó dựa trên lớp của các mẫu gần nhất.
- 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 tiền xử lý dữ liệu EDA và PCA.
- Sau đó chuẩn hóa và làm sạch lại tập dữ liệu rồi mới tiến hành huấn luyện cho mô hình.

3 Experiment

3.1 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.

3.2 Exploratory Data Analysis (EDA)

- Sau khi thực hiện việc tải dataset lên Colab, 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.

3.3 Principal Component Analysis (PCA)

- Có nhiều kỹ thuật giảm chiều được sử dụng cho học máy (PCA, t-SNE,...). Trong đồ án này, em chọn kỹ thuật PCA để giảm chiều dữ liệu.
- PCA (Principal Component Analysis) là một kỹ thuật giảm chiều phổ biến được sử dụng trong học máy. PCA hoạt động bằng cách chuyển đổi các thuộc tính của dữ liệu sang một hệ thống tọa độ mới, trong đó các thuộc tính mới có liên quan chặt chẽ với nhau.
- Bộ dữ liệu Breast Cancer Wisconsin này là bộ dữ liệu có không gian 30 chiều, em tiến hành giảm chiều của bộ dữ liệu này xuống còn 2 chiều, để xem liệu các biến có thể tách thành các cụm riêng biệt được không. Bộ dữ liệu 2 chiều này được chia làm 2 lớp, với màu tối là benign và màu sáng là malignant (*Figure 1*).

3.4 Evaluation

- Evaluation metric: Accuracy.
- Trong đồ án này, để đánh giá và so sánh hiệu suất của các mô hình, em sử dụng độ đo Accuracy.
- *Table 1* thể hiện kết quả bằng độ đo Accuracy của từng phương pháp trên dataset Breast Cancer Wisconsin.

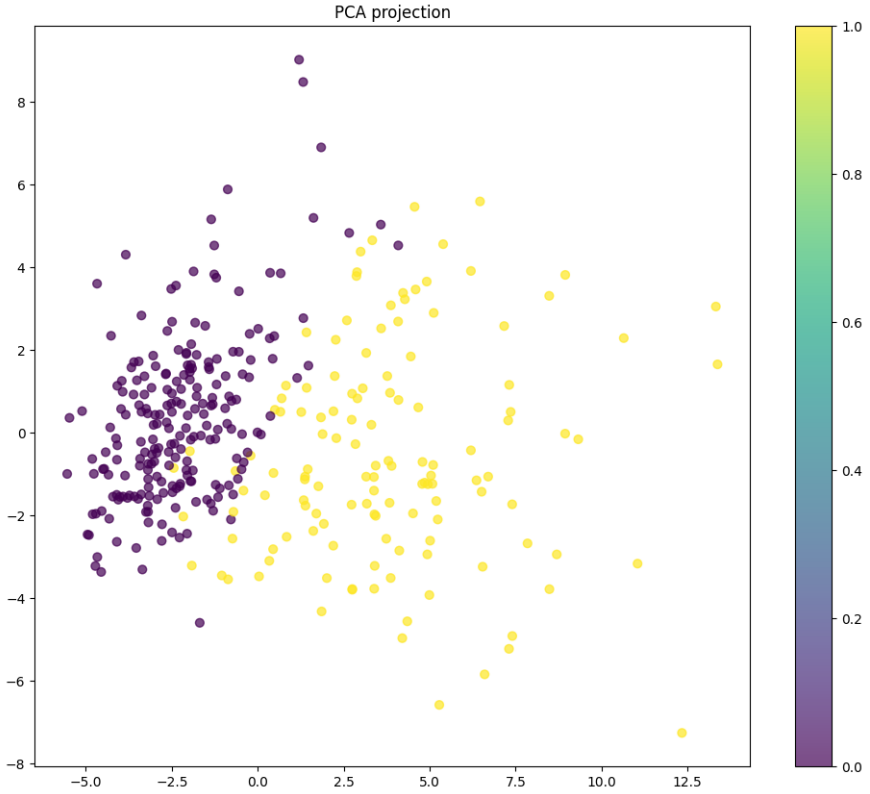

Figure 1: Chiếu dữ liệu từ không gian 30 chiều xuống không gian 2 chiều

Model	LR	kNN	SVM
Acc	98.09	97.14	99.04

Table 1: *Đánh giá các mô hình qua độ đo Accuracy*

4 Conclusion

- Trước khi huấn luyện một mô hình máy học, ta phải trải qua các bước tiền xử lý dữ liệu. Trong đồ án này, em đã thực hiện 2 bước tiền xử lý dữ liệu là EDA và PCA.
- Qua thực nghiệm, ta có thể dễ dàng nhận thấy, mô hình hoạt động tốt nhất trên dataset Breast Cancer Wisconsin là SVM.

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

References

[1] Saima Nazir Sara Ibrahim and Sergio A. Velastin. Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis. 10 2021.