TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẠI HỌC QUỐC GIA TP.HCM

TÊN ĐỒ ÁN

Môn học: Python cho Khoa học Dữ liệu

Giảng viên hướng dẫn: ThS. Hà Minh Tuấn

Nhóm sinh viên thực hiện:

Nguyễn Thị Ngọc Anh	23280037
Trương Thị Quỳnh Giang	23280052
Trần Trung Kiên	23280066

TP.HCM, 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

2. Kết quả đạt được

3. Ý thức làm việc của sinh viên:

TP. Hồ Chí Minh, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

ThS. Hà Minh Tuấn

LỜI CẢM ƠN

Mục lục

Bảng ký hiệu và chữ viết tắt

Danh sách bảng

Danh sách hình

Tóm tắt

Giới thiệu

Chương 1 CƠ SỞ LÝ THUYẾT

1.1 Các phương pháp vectorize dữ liệu

1.1.1 One Hot

1.1.2 TF IDF

1.1.3 PhoBert

1.1.4 Một số phương pháp khác

1.2 Các mô hình học máy

1.2.1 Logistic Regression

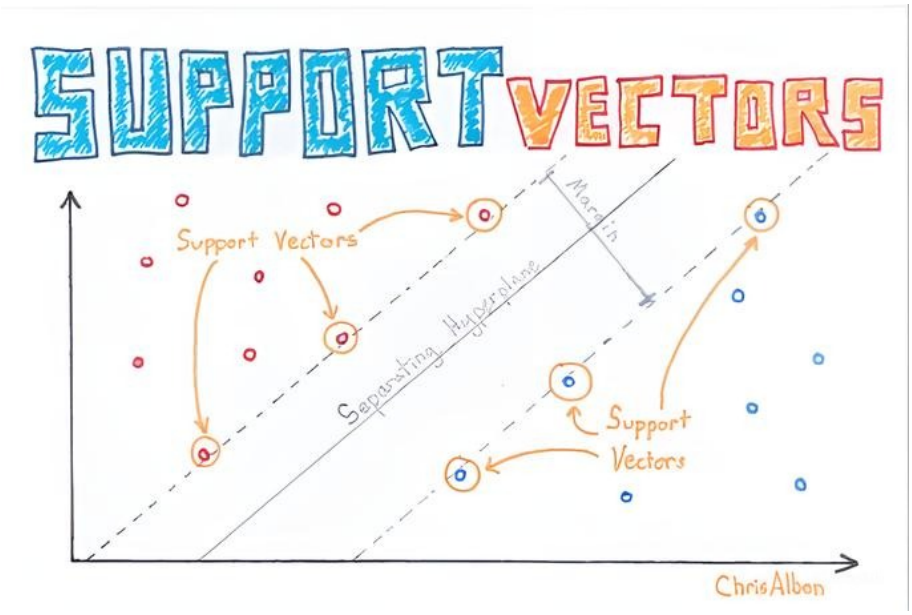
1.2.2 Support Vector Machine

Support Vector Machine (SVM) là mô hình học máy giám sát, mục tiêu của thuật toán là tìm ra một siêu phẳng (hyperplane) sao cho có thể phân tách tối ưu các điểm dữ liệu thuộc các lớp khác nhau. "Tối ưu" ở đây nghĩa là tìm ra siêu phẳng tạo ra khoảng cách lớn nhất (margin) giữa các lớp dữ liệu. Các điểm dữ liệu có khoảng cách nhỏ nhất đến siêu mặt phẳng (các điểm gần nhất) được gọi là các vector hỗ trợ (support vectors) Margin là độ rộng tối đa của dải không chứa bất kỳ điểm dữ liệu nào và song song với siêu phẳng.

Việc tính toán phân tách dữ liệu phụ thuộc vào hàm kernel. Có nhiều hàm kernel khác nhau: Tuyến tính (linear kernel), đa thức (polynomial kernel), gaussian, RBF (radial basis function), sigmoid kernel. Các hàm này xác định độ mượt và hiệu quả trong việc phân tách lớp, và việc tùy biến các siêu tham số của chúng có thể dẫn đến hiện tượng quá khớp (overfitting) hoặc thiếu khớp (underfitting). Trong khuôn khổ báo cáo này, nhóm không trình bày chi tiết các biểu thức toán học của từng loại kernel, bạn có thể tham khảo công thức đầy đủ ở tài liệu sau: [1] SVM không chỉ hỗ trợ phân loại nhị phân và tách các điểm dữ liệu thành hai lớp, SVM còn mở rộng để hỗ trợ các bài toán phân loại đa lớp thông qua cơ chế chia nhỏ bài toán đa lớp thành nhiều mô hình nhị phân.

a) Phương pháp OvO (One-to-One)

   Phương pháp OvO xây dựng bộ phân loại nhị phân cho mỗi cặp lớp trong tập dữ liệu. Với tập dữ liệu gồm K lớp, số lượng mô hình được tạo ra là *công



Hình 1.1: Hình 1. Support Vector Machine

thức toán học* [K(K-1)]/2. Mỗi mô hình được huấn luyện để phân biệt hai lớp cụ thể. Khi dự đoán, mỗi mô hình đưa ra một phiếu bầu cho một trong hai lớp, và lớp nhận được nhiều phiếu nhất sẽ được chọn làm kết quả cuối cùng. Phương pháp OvO thường hiệu quả khi số lượng lớp không quá lớn và mỗi mô hình nhị phân tương đối nhẹ.

b) Phương pháp OvR (One-to-Rest)

Phương pháp OvR (OvA) chia dữ liệu đa lớp thành nhiều bài toán phân loại nhị phân, sau đó mỗi bộ phân loại nhị phân được huấn luyện trên mỗi bài toán phân loại nhị phân và đưa ra dự đoán bằng cách sử dụng mô hình có độ tin cậy cao nhất. Phương pháp này yêu cầu tạo ra một mô hình cho mỗi lớp.

1.2.3 Naive Bayes

Chương 2 THU THẬP VÀ KHÁM PHÁ DỮ LIỆU

2.1 Thu thập dữ liệu

2.2 Khám phá bộ dữ liệu

Chương 3 XÂY DỰNG MÔ HÌNH PHÂN TÍCH

3.1 Module Loader

3.2 Module Preprocessor

3.3 Module feature

3.4 Tối ưu tham số

3.4.1 Logistic Regression

3.4.2 Support Vector Machine

3.4.3 Naive Bayes

3.5 Huấn luyện mô hình

3.5.1 Logistic Regression

3.5.2 Support Vector Machine

3.5.3 Naive Bayes

Chương 4 KẾT QUẢ THỰC NGHIỆM

4.1 Đánh giá mô hình

4.2 So sánh mô hình

4.3 Ứng dụng thực tế

Tài liệu tham khảo

[1] V. Cortes, C. & Vapnik, "Support-vector networks," *Machine Learning*, 1995.