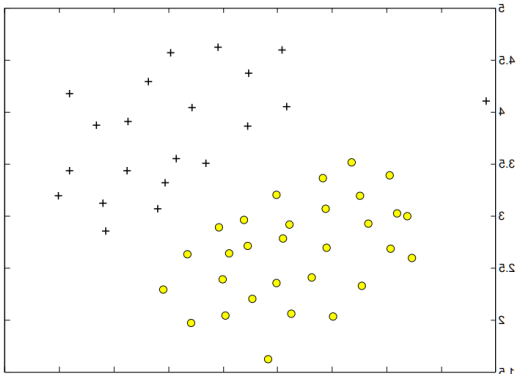
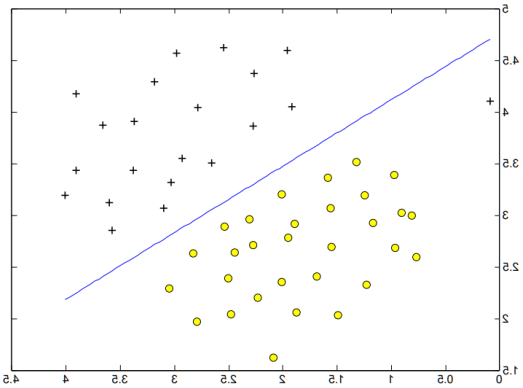
Bài tập: Các phương pháp học có giám sát và ứng dụng

Bài 1. Cho dữ liệu là một tập điểm 2D có thể phân tách bằng một đường thẳng. Dữ liệu lưu ở file data1.txt trong đó 2 cột đầu là tọa độ X, Y, cột cuối cùng là nhãn của điểm đó

a. Vẽ minh họa dữ liệu như hình



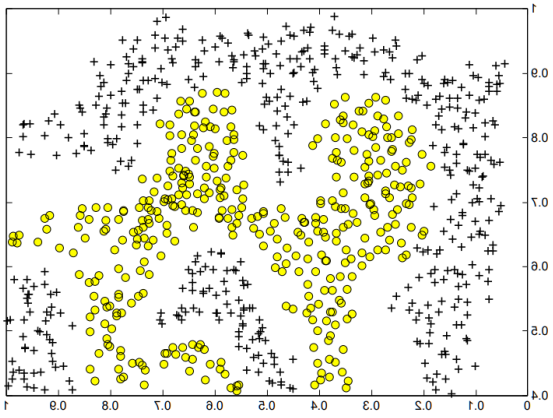
b. Sử dụng giải thuật Linear SVM để tìm ra đường thẳng đó. Vẽ minh họa như hình



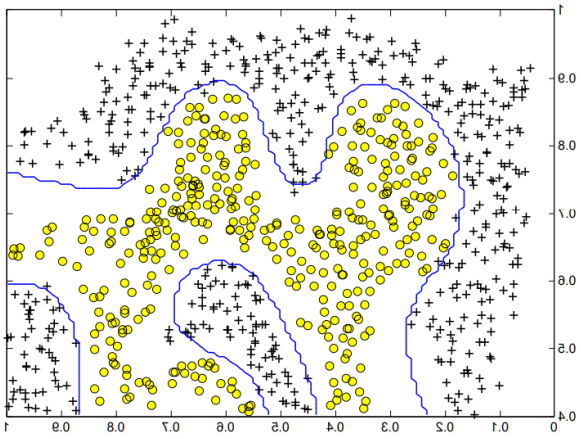
c. Thay đổi tham số C, vẽ minh họa kết quả giống câu b, đánh giá, giải thích kết quả

Bài 2. Tương tự bài 1, sử dụng dữ liệu từ file data1.txt và data2.txt để xây dựng bộ phân lớp SVM sử dụng Gaussian Kernel, tối ưu các tham số C và σ để sinh ra đường biên tốt nhất.

a. Đọc, vẽ minh họa từng dataset, ví dụ:



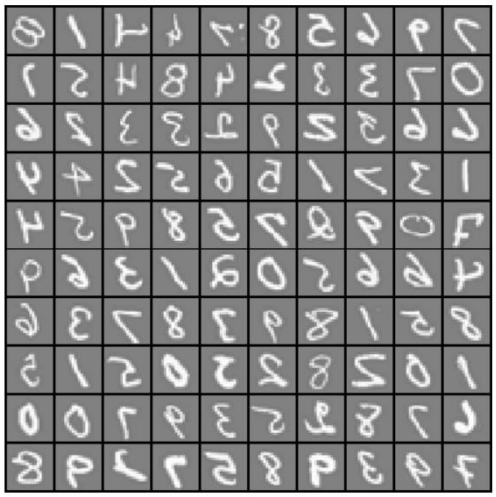
b. Xây dựng bộ phân lớp SVM với nhân Gaussian, vẽ minh họa kết quả, ví dụ như hình



c. Tối ưu các tham số, vẽ minh họa kết quả

Bài 3. Cho CSDL bao gồm 5000 mẫu ký tự viết tay ở file hand_digit_X.npy và hand_digit_y.npy. Mỗi mẫu là ảnh xám có kích thước 20x20. Mỗi pixel được thể hiện bằng giá trị thực thể hiện mức xám ở vị trí đó. Ảnh được "duỗi" ra thành một vector 400 chiều. Như vậy ma trận X sẽ có kích thước 5000x400

a. Đọc và vẽ minh họa dữ liệu, kết quả giống như hình:



b. Tách ngẫu nhiên dữ liệu ra 70% để huấn luyện và 30% để test
c. Huấn luyện sử dụng các giải thuật học máy khác nhau, đánh giá, so sánh kết quả của các bộ phân lớp khác nhau
d. Với bộ phân lớp đạt kết quả tốt nhất, vẽ ma trận nhầm lẫn (confusion matrix)

Bài 4. Tự triển khai lại thuật toán K-NN:

a. Triển khai lại thuật toán K-NN chỉ sử dụng bộ thư viện chuẩn của Python (không được dùng numpy)

b. So sánh, đánh giá độ chính xác + tốc độ của giải thuật tự triển khai so với giải thuật của thư viện Scikit-learn
c. So sánh tốc độ của giải thuật khi tối ưu sử dụng các thuật toán sắp xếp (sorting)/ có nhớ

Bài 5. Xây dựng lại giải thuật multi-class SVM

a. Xây dựng các bộ phân lớp SVM nhị phân cho từng cặp lớp trong bộ CSDL Iris
b. Xây dựng bộ phân lớp SVM nhiều lớp sử dụng chiến lược one-vs-one
c. Xây dựng bộ phân lớp SVM nhiều lớp sử dụng chiến lược one-vs-all
d. Với mỗi mẫu đưa vào, tìm cách trả về xác suất bộ phân lớp dự đoán cho từng lớp thay vì một lớp cố định
e. Tìm cách phát hiện trong trường hợp ảnh đưa vào không thuộc lớp nào cả

Bài 6*. Xây dựng hệ thống tra cứu sử dụng K-NN: Giả sử bạn có CSDL khuôn mặt của 1 triệu người ở Hà Nội, mỗi người có 5 ảnh, mỗi ảnh được biểu diễn bởi một vector đặc trưng 1024 chiều.

a. Xây dựng CSDL giả lập cho bài toán trên, CSDL bao gồm 1 triệu class, 5x1 triệu = 5 triệu mẫu -> ma trận huấn luyện là 5 triệu x 1024
   Xây dựng CSDL sao cho các vector đặc trưng của cùng một người thì hầu hết gần nhau và các vector đặc trưng của hai người bất kỳ thì hầu hết xa nhau
b. Tìm cách vẽ minh họa CSDL
c. Huấn luyện giải thuật KNN trên CSDL đã sinh ra
d. Với một khuôn mặt mới đưa vào, tra cứu ảnh đó thuộc người nào hoặc không xuất hiện trong CSDL
e. Đánh giá tốc độ nhận dạng/ tra cứu hiện tại, tìm cách cải thiện tốc độ tra cứu

Bài 7*. Xây dựng hệ thống phân loại thư rác sử dụng CSDL tại địa chỉ:
https://spamassassin.apache.org/old/publiccorpus/

a. Download dữ liệu, hiển thị một vài mẫu
b. Tiền xử lý dữ liệu, bao gồm các công đoạn:

   - Chuyển tất cả ký tự thành chữ viết thường

   - Xóa bỏ các thẻ HTML như <html> </body>,…

   - Chuẩn hóa địa chỉ email: thay tất cả địa chỉ email bằng ký tự {email}

   - Chuẩn hóa số: thay tất cả địa chỉ số xuất hiện bằng ký tự {number}

   - Chuẩn hóa đơn bị: thay tất cả ký tự $ bằng ký tự {dollar}

   - Chuẩn hóa kiểu chữ (stemming): Ví dụ: “discount”, “discounts”, “discounted” và “discounting” được chuẩn hóa về “discount”, tìm kiếm và sử dụng các thư viện stemmer sẵn có để sử dụng, ví dụ NLTK

   - Xóa tất cả các ký tự không phải là chữ: ví dụ như tab, xuống dòng, cách

Kết quả đạt được giống như hình sau:

```
anyon know how much it cost to host a web portal well it depend on how
mani visitor your expect thi can be anywher from less than number buck
a month to a coupl of dollarnumb you should checkout httpaddr or perhap
amazon ecnumb if your run someth big to unsubscrib yourself from thi
mail list send an email to emailaddr
```

c. Xây dựng bộ từ vựng: từ tất cả các từ trong tất cả các email, xây dựng bộ từ vựng sao cho mỗi từ trong đó xuất hiện ít nhất 100 lần, với dữ liệu ở trên, bộ từ vựng bao gồm 1899 từ. Trong các hệ thống thực tế, bộ từ vựng bao gồm từ 10000 đến 50000 từ.
d. Với mỗi email, xây dựng một vector đặc trưng có 1899 chiều, trong đó mỗi giá trị là 0/1 thể hiện từ đó có ở trong email hay không
e. Xây dựng hệ thống phân loại thư rác dựa trên vector đặc trưng / nhãn đã đưa vào, hiển thị các thư rác tiêu biểu nhất
f. Thay vì sử dụng giá trị 0/1 cho vector đặc trưng, sử dụng giá trị thực từ 0-1 thể hiện tần suất xuất hiện của các từ trong email