Support Vector Machine

*E-mail: pima.vn@gmail.com

Mô tả chung

Support Vector Machine (SVM) là mô hình học có giám sát được phát minh bởi Vladimir Naumovich Vapnik và Alexey Yakovlevich Chervonenkis năm 1963. Mô hình này được dùng chủ yếu là trong bài toán **phân lớp** (classification) dữ liệu. Trong dự án này, các bạn sẽ tìm hiểu hiểu về SVM và các biến thể của nó áp dụng trong bài toán **phân lớp nhị phân** sử dụng một **siêu phẳng** phân tách.

Yêu cầu cơ bản

Lý thuyết:

- Phát biểu bài toán phân loại nhị phân có giám sát tổng quát theo ngôn ngữ toán học.
- Phát biểu hàm mục tiêu cần tối ưu và các điều kiện ràng buộc khi phân tách theo mô hình SVM.
- Để tối ưu cho hàm mục tiêu nói trên, trong phần lớn các trường hợp, người ta giải bài toán đối ngẫu thay vì bài toán gốc. Hãy trình bày ý tưởng này và giải thích lý do.
- Trình bày cụ thể các bước cần thực hiện trong việc giải một bài toán phân loại nhị phân có giám sát (tức bao gồm huấn luyện và truy vấn trên mô hình đã huấn luyện) sử dụng mô hình SVM.
- Trong nhiều trường hợp dữ liệu ta có không thể phân cách tuyến tính được và có nhiều thông tin nhiễu. Để giải quyết các trường hợp này người ta đã đề xuất Soft Margin SVM và Kernel SVM. Hãy tìm hiểu và trình bày các phương pháp này. Các phương pháp này thay đổ quy trình thực hiện như thế nào?

Thực hành:

- Tìm hiểu cách xây dựng và sử dụng mô hình SVM và các biến thể trong thư viện `scikit-learn`. Thử nghiệm trên các bộ dữ liệu tự sinh để quan sát trực quan kết quả, nhận xét và so sánh giữa các biến thể.
- Tìm một bộ dữ liệu thực tế, quan sát, phân tích ở mức cơ bản các đặc điểm của bộ dữ liệu đã chọn. Áp dụng mô hình SVM và các biến thể của SVM vào bộ dữ liệu này. So sánh, nhận xét về kết quả và đánh giá mô hình.

Câu hỏi nâng cao

Lý thuyết:

- Trong một số trường hợp, tối ưu SVM có thể biến đổi thành bài toán tối ưu không ràng buộc và có thể giải được bằng Gradient Descent. Hãy tìm hiểu và phân tích vấn đề này. Áp dụng kiến thức đã tìm hiểu vào dữ liệu thực tế và so sánh, đánh giá.

- Kết quả của bài toán SVM là một phân lớp cứng. Hãy đề xuất hoặc tìm hiểu và trình một số giải pháp để gán một giá trị điểm (có thể hiểu là độ tự tin hay xác suất) cho dự đoán của SVM.
- Có thể mở rộng SVM cho bài toán phân lớp dữ liệu có thể thuộc về một trong nhiều hơn 2 lớp hay không? Nếu có, tìm hiểu và trình bày ý tưởng và tính chất của phương pháp mở rộng đã tìm hiểu.

Thực hành:

- Cài đặt mô hình SVM và các biến thể theo các bước đã phân tích (chỉ sử dụng các thư viện hỗ trợ như numpy, cvxopt,..).

Kiến thức

- Toán: Ma trận, các phép toán trên ma trận, vi tích phân nhiều biến, bài toán tối ưu lồi, phương pháp nhân tử Lagrange,...
- Một số từ khóa: Supervised Learning, Binary Classification, Hyperplane, Convex Optimization, Linearly Seperable, Linear Classifier, Maximum Margin, Dual Problem,...

Tham Khảo

[1] Các bài giảng PiMA 2021.
[2] https://en.wikipedia.org/wiki/Support_vector_machine
[3] https://scikit-learn.org/stable/modules/svm.html