

1. Lý thuyết thuật toán SVM

1.1. Giới thiệu SVM (Support Vector Machine)

SVM (Máy vector hỗ trợ) là thuật toán phân loại có giám sát (Supervised Learning). Mục tiêu của SVM là tìm ra siêu phẳng (hyperplane) tốt nhất để phân tách các lớp dữ liệu sao cho khoảng cách biên (margin) giữa hai lớp là lớn nhất có thể.

- Dữ liệu càng gần biên \rightarrow càng ảnh hưởng mạnh đến việc xác định siêu phẳng.
- Những điểm nằm gần hoặc trên biên gọi là Support Vectors.

1.2. Hàm mục tiêu của SVM

Với tập dữ liệu huấn luyện $((x_i, y_i)), (y_i \in \{-1, 1\})$:

Ta cần tìm véc-tơ trọng số w và hệ số b sao cho:

$$[\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i]$$

với điều kiện:

$$[y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0]$$

(C) là hệ số điều chỉnh độ phạt sai số.

(ξ_i) là sai số cho từng điểm dữ liệu (cho phép một số điểm nằm sai biên).

2. Các tham số quan trọng trong SVM

2.1. Tham số C (Penalty parameter)

- Là hệ số điều chỉnh giữa độ chính xác phân loại và độ lớn của khoảng cách biên.
- Khi C lớn \rightarrow mô hình phạt mạnh các điểm sai \rightarrow biên hẹp hơn, dễ overfitting.
- Khi C nhỏ \rightarrow mô hình chấp nhận một số lỗi nhỏ, biên rộng hơn, dễ underfitting.

Tóm tắt:

C	Ý nghĩa	Đặc điểm
Nhỏ	Ưu tiên biên rộng	Có thể bỏ qua một số điểm sai
Lớn	Ưu tiên phân loại chính xác	Dễ quá khớp dữ liệu

2.2. Tham số Gamma (γ)

- Áp dụng cho các kernel phi tuyến (như RBF, poly).
- Điều khiển độ ảnh hưởng của một điểm dữ liệu lên ranh giới quyết định.
 - + Gamma lớn \rightarrow mỗi điểm ảnh hưởng mạnh, ranh giới phức tạp, dễ overfit.
 - + Gamma nhỏ \rightarrow ảnh hưởng yếu, ranh giới mượt hơn, dễ underfit.

Tóm tắt:

Gamma	Đặc điểm	Ảnh hưởng
Nhỏ	Ranh giới mượt, đơn giản	Underfitting
Lớn	Ranh giới phức tạp	Overfitting

2.3. Tham số Kernel

Kernel là hàm nhân giúp chuyển dữ liệu phi tuyến sang không gian đặc trưng có tuyến tính để phân tách được.

Một số kernel phổ biến:

Kernel	Công thức	Ứng dụng
linear	$(K(x, x') = x \cdot x')$	Dữ liệu tuyến tính
poly	$(K(x, x') = (x \cdot x' + 1)^d)$	Dữ liệu có mối quan hệ đa thức
rbf	$(K(x, x') = e^{-\gamma})$	