Inteligência Artificial

Raoni F. S. Teixeira

Aula 11 - SVM

1 Introdução

Support Vector Machine (SVM) é um classificador que separa os dados encontrando o hiperplano com a maior margem possível.

Nesta aula, consideramos um problema de classificação binária com rótulos $\{+1, -1\}$. O classificador é definido pela equação:

$$h(\mathbf{x}) = \mathbf{sign}(\mathbf{w}^T\mathbf{x} + b),$$

onde $\mathbf{w}$ é o vetor de pesos, $\mathbf{x}$ é o vetor de entrada, b é o termo bias, e $\mathbf{sign}$ é uma função que devolve o sinal $+$ ou $-$ dos rótulos.

2 SVM

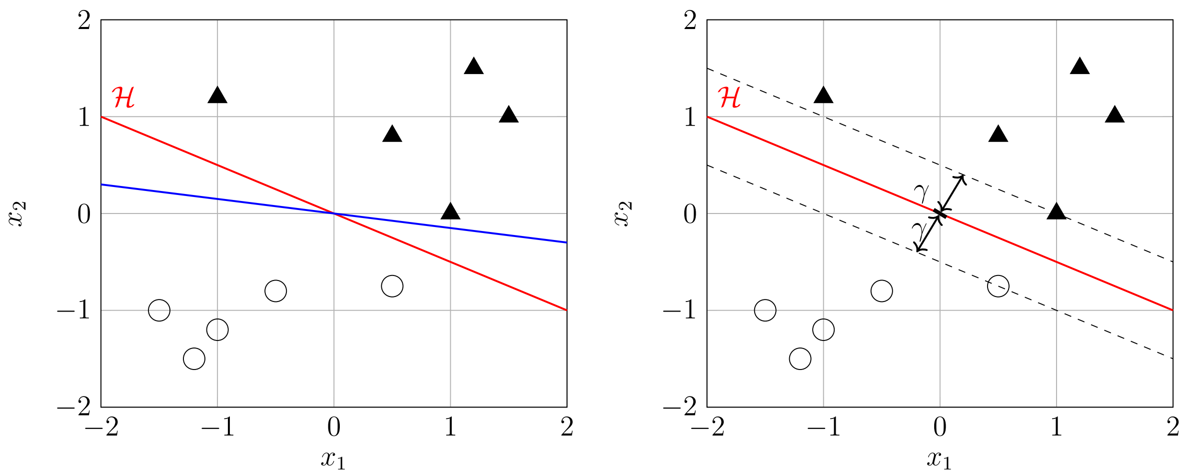
Existem infinitos hiperplanos capazes de separar um conjunto de dados linearmente separável. A pergunta central é: *Qual é o melhor hiperplano?*

O SVM responde a essa questão ao escolher o hiperplano que maximiza a distância para os pontos mais próximos de ambas as classes, ou seja, o **hiperplano com margem máxima**.

A Figura 1 ilustra o hiperplano de margem máxima em um conjunto de dados bidimensional. Na Figura 1a, comparamos dois hiperplanos para o mesmo conjunto de dados: o hiperplano de margem máxima (vermelho) e um hiperplano alternativo (azul). A Figura 1b destaca os pontos mais próximos ao limite de decisão, conhecidos como **vetores de suporte**, e a margem γ, que representa a distância entre o hiperplano e esses pontos.

2.1 Margem

A margem γ é a distância entre o hiperplano e os pontos mais próximos das duas classes.



(a) (b)

Figura 1: Hiperplano de margem máxima em um conjunto de dados bidimensional: (a) compara dois hiperplanos e (b) destaca os vetores de suporte e a margem.

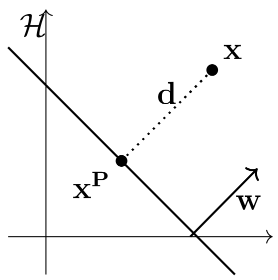


Figura 2: Distância entre um ponto $\mathbf{x}$ e o hiperplano $\mathcal{H}$.

Como calcular a distância de um ponto ao hiperplano?

Considere um hiperplano definido por $\mathbf{w}$ e b, descrito como $\mathcal{H} = \{\mathbf{x} \mid \mathbf{w}^T\mathbf{x}+b = 0\}$. Considere também o vetor $\mathbf{d}$ conectando o ponto $\mathbf{x}$ ao hiperplano $\mathcal{H}$. A distância de um ponto $\mathbf{x}$ até $\mathcal{H}$ é dada por:

$$\|\mathbf{d}\| = \frac{|\mathbf{w}^\mathbf{T}\mathbf{x} + b|}{\|\mathbf{w}\|_2}.$$

A Figura 2 mostra a distância entre um ponto $\mathbf{x}$ e o hiperplano $\mathcal{H}$, ilustrando os conceitos descritos. A projeção $\mathbf{x}^P$ de $\mathbf{x}$ no hiperplano é definida por:

$$\mathbf{x}^P = \mathbf{x} - \mathbf{d}.$$

Como $\mathbf{d}$ é paralelo a $\mathbf{w}$, podemos escrever:

$$\mathbf{d} = \alpha \mathbf{w}.$$

Sabendo que $\mathbf{x}^P \in \mathcal{H}$, temos:

$$\mathbf{w}^T \mathbf{x}^P + b = 0.$$

Substituindo $\mathbf{x}^P$ em termos de $\mathbf{d}$:

$$\mathbf{w}^T(\mathbf{x} - \mathbf{d}) + b = \mathbf{w}^T(\mathbf{x} - \alpha \mathbf{w}) + b = 0.$$

Logo,

$$\alpha = \frac{\mathbf{w}^T \mathbf{x} + b}{\mathbf{w}^T \mathbf{w}}.$$

O comprimento de $\mathbf{d}$ é então:

$$\|\mathbf{d}\|_2 = \underbrace{|\alpha| \cdot \|\mathbf{w}\|_2}_{\text{propriedade da norma}} = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\underbrace{\mathbf{w}^T \mathbf{w}}_{\text{sempre } \geq 0}} \|\mathbf{w}\|_2 = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\mathbf{w}^T \mathbf{w}} \sqrt{\mathbf{w}^T \mathbf{w}} = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|_2}.$$

Finalmente, a margem de um hiperplano $\gamma(\mathbf{w}, b)$ é a menor distância para qualquer ponto do conjunto de dados:

$$\gamma(\mathbf{w}, b) = \min_{\mathbf{x} \in D} \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|_2}.$$

Como a margem e o hiperplano são invariantes à escala, temos:

$$\gamma(\beta \mathbf{w}, \beta b) = \gamma(\mathbf{w}, b), \quad \forall \beta \neq 0.$$

Se o hiperplano maximiza γ, ele estará centralizado entre as duas classes, garantindo a maior separação possível entre elas.

3 Classificador de Margem Máxima

Podemos formular nossa busca pela margem máxima como um problema de otimização restrito. O objetivo é maximizar a margem sob as restrições de que todos os pontos de dados devem estar no lado correto do hiperplano:

$$\underbrace{\max_{\mathbf{w}, b} \gamma(\mathbf{w}, b)}_{\text{maximiza margem}} \quad \text{sujeito a} \quad \underbrace{\forall i \; y_i \big(\mathbf{w}^T \mathbf{x}_i + b\big) \geq 0}_{\text{garante a separação}}$$

Substituindo a definição de γ, obtemos:

$$\underbrace{\max_{\mathbf{w},b} \underbrace{\frac{1}{\|\mathbf{w}\|_2} \min_{\mathbf{x}\in D} \left|\mathbf{w}^T\mathbf{x}+b\right|}_{\gamma(\mathbf{w},b)}}_{\text{maximiza margem}} \quad \text{sujeito a} \quad \underbrace{\forall i\; y_i\big(\mathbf{w}^T\mathbf{x}_i+b\big) \geq 0}_{\text{garante separação}}$$

Como o hiperplano é invariante em escala, podemos fixar a escala de $\mathbf{w}$ e b como quisermos. Vamos ser estratégicos e escolher de modo que:

$$\min_{\mathbf{x}\in D} \left|\mathbf{w}^T\mathbf{x}+b\right| = 1.$$

Podemos adicionar esse redimensionamento como uma restrição de igualdade. Assim, nosso objetivo torna-se:

$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} \cdot 1 = \min_{\mathbf{w},b} \|\mathbf{w}\|_2 = \min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w}.$$

Aqui, utilizamos o fato de que $f(z) = \sqrt{z}$ é uma função monotonicamente crescente para $z \geq 0$ e $\|\mathbf{w}\|_2 \geq 0$. Logo, o $\mathbf{w}$ que maximiza $\|\mathbf{w}\|$ também maximiza $\mathbf{w}^T\mathbf{w}$.

O novo problema de otimização se torna:

$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w} \quad \text{sujeito a} \quad \forall i\; y_i\big(\mathbf{w}^T\mathbf{x}_i+b\big) \geq 1.$$

Esta formulação resulta em um problema de otimização quadrática, onde a função objetivo é quadrática e as restrições são lineares. Podemos resolvê-lo de forma eficiente usando algum solver QCQP – Quadratically constrained Quadratic Program. A solução é o hiperplano mais simples (com menor $\mathbf{w}^T\mathbf{w}$) que garanta uma margem de no mínimo uma unidade para todos os pontos.

Os vetores de suporte, que estão na margem, determinam a posição e orientação do hiperplano. Apenas eles influenciam o resultado final. Se um vetor de suporte for movido, o hiperplano será ajustado. Por outro lado, mover pontos que não são vetores de suporte não afetará o hiperplano, desde que esses movimentos não alterem as margens.

4 SVM com Restrições Suaves

Em dimensões baixas, frequentemente não existe um hiperplano que separe as classes. Para contornar isso, introduzimos variáveis de folga (ξ_i), que permitem algumas violações nas margens:

$$\min_{\mathbf{w},b,\xi} \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i, \quad \text{sujeito a} \quad y_i\big(\mathbf{w}^T\mathbf{x}_i+b\big) \geq 1-\xi_i,\ \xi_i \geq 0, \forall i.$$

Essas variáveis penalizam as violações na função objetivo. O parâmetro C controla o rigor do modelo:

- Valores altos de C forçam o modelo a minimizar violações, resultando em margens rígidas.
- Valores baixos permitem maior flexibilidade, priorizando uma solução mais simples.

4.1 Formulação Irrestrita

Podemos expressar ξ_i como:

$$\xi_i = \max(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b), 0).$$

Substituindo na função objetivo, obtemos a versão irrestrita:

$$\min_{\mathbf{w},b} \underbrace{\mathbf{w}^T\mathbf{w}}_{\text{regularização L2}} + C\sum_{i=1}^{n} \underbrace{\max\left[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b), 0\right]}_{\text{hinge-loss}}.$$

Essa é função **hinge loss** com $\|\mathbf{w}\|^2$ atuando como regularizador. Essa formulação permite otimizar os parâmetros por métodos como descida de gradiente, similar à regressão logística, mas com uma perda distinta.

5 Problemas Não Linearmente Separáveis

Para dados não linearmente separáveis, o SVM utiliza *kernels* para mapear os dados para espaços de dimensão mais alta (uma estratégia semelhante ao mapeamento da aula 7), onde a separação linear se torna possível. Funções de kernel comuns incluem:

- **Linear:** $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$,
- **Polinomial:** $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + c)^d$,
- **RBF (Radial Basis Function):** $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$.

Esses kernels permitem que o SVM lide com dados complexos e não linearmente separáveis, preservando o poder de generalização.

6 Exercícios

1. Considere o seguinte conjunto de dados bidimensional linearmente separável:
   - Classe $+1$: $(2, 3)$ e $(4, 5)$.
   - Classe -1: $(1, 1)$ e $(3, 2)$.

Responda às questões abaixo:

a) Escreva a equação genérica do hiperplano que separa as duas classes.

b) Calcule a margem do hiperplano obtido.

c) Identifique os vetores de suporte no conjunto de dados.

2. Um SVM com restrições suaves foi usado para classificar e-mails como "spam"ou "não spam". O parâmetro C afeta diretamente o desempenho do modelo. Responda:

a) Explique como C equilibra o compromisso entre margem rígida e margem flexível.

b) Determine como ajustar C em cada caso a seguir:

- Quando a precisão é mais importante que a generalização.
- Quando a generalização é mais importante que a precisão.