Máquinas de vetor de suporte

Carlos Diego Rodrigues

14 de dezembro de 2021

Universidade Federal do Ceará

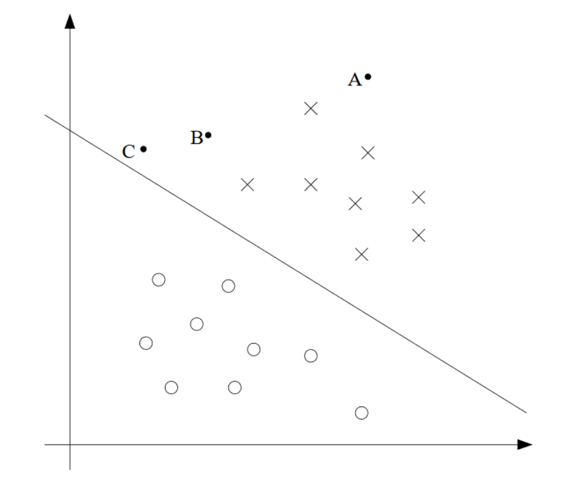
Máquinas de vetor de suporte - *SVM*

- Máquinas de vetor de suporte estão entre os métodos mais populares e mais eficientes para problemas de classificação.
- Consiste em encontrar uma borda (ou margem) de classificação que separa exemplos de classes distintas.
- Qual a melhor margem separadora? Vetores de suporte.
- A partir de uma borda linear é possível encontrar outras bordas de maior complexidade que satisfação algumas condições na definição da função de borda.
- *Kernell tricks*

Margens

- Quando fazemos uma regressão logística, temos como resultado um modelo θ de tal forma que ao aplicarmos a função sobre um exemplo x^i, o resultado $h_\theta(x^i) \geq 0,5 \leftrightarrow \theta^T x^i \geq 0$.
- Se $\theta^T x^i \gg 0$ então temos alta confiança que $y^i = 1$ enquanto se $\theta^T x^i \ll 0$ temos alta confiança que $y^i = 0$.
- Podemos dizer assim que o hiperplano $\theta^T x = 0$ é um hiperplano separador para as duas classes.

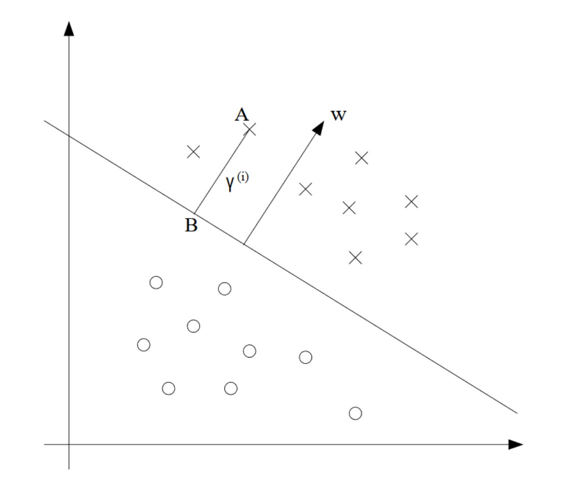
Margens



Margens geométricas

- Primeiro vamos considerar que $y^i \in \{-1, +1\}$.
- Dado um exemplo (x^i, y^i), vamos definir uma margem funcional de (w, b) com relação ao exemplo o valor $\gamma^i = y^i(w^T x^i + b)$
- Se $y^i = 1$ então queremos que o valor da margem seja positivo, enquanto se $y^i = 0$ queremos o valor da margem negativo.
- Restrição: $y^i(w^T x^i + b) \geq 0$

Margens geométricas



Normalização da margem

- Normalização da margem (w, b):

$$w^T \left(x^i - \gamma^i \frac{w}{||w||} \right) + b = 0$$

- Resolvendo para γ^i temos:

$$\gamma^i = y^i \left(\left(\frac{w}{||w||} \right) x^i \right) + \frac{b}{||w||}$$

- Menor margem é a medida de qualidade do modelo:

$$\gamma = \min_{i=1,\cdots,m} y^i$$

Problema de otimização

- Queremos maximizar a margem para aumentar a confiança do modelo:

$$\max_{\gamma, w, b} \gamma$$
$$s.a : y^i(w^T x^i + b) \geq \gamma \qquad i = 1, \cdots, m$$
$$||w|| = 1$$

- Podemos utilizar a relação entre γ e $||w||$ para reescrever o problema de uma outra forma:

$$\max_{\gamma, w, b} \frac{\hat{\gamma}}{||w||}$$
$$s.a : y^i(w^T x^i + b) \geq \hat{\gamma} \qquad i = 1, \cdots, m$$

Modelo final

- Fazendo $\hat{\gamma} = 1$ chegamos na minimização de $||w||$:

$$\min_{w,b} \frac{1}{2}||w||^2$$
$$y^i(w^T x^i + b) \geq 1 \qquad\qquad i = 1, \cdots, m$$

- Problema de otimização não linear (quadrático)
- Técnicas consolidadas para resolvê-lo, incluindo a decomposição lagrangeana.

Decomposição Lagrangeana

- Dado um problema de otimização qualquer:

$$
\begin{aligned}
&\min_{w} \; f(w) \\
&s.a: g_i(w) \leq 0 \qquad\qquad i = 1, \cdots, k \\
&\qquad h_i(w) = 0 \qquad\qquad i = 1, \cdots, l
\end{aligned}
$$

- Definimos o dual Lagrangeano como a função:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

onde α e β são chamados multiplicadores de Lagrange.

Decomposição Lagrangeana

- Considere, para um w fixo o valor:

$$\theta(w) = \max_{\alpha \geq 0, \beta} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

- Otimizando em w temos:

$$\min_{w} \theta(w) = \min_{w} \max_{\alpha \geq 0, \beta} \mathcal{L}(w, \alpha, \beta)$$

- Sob certas condições (funções g são convexas e funções h são afins), o problema acima é equivalente a:

$$\max_{\alpha \geq 0, \beta} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

Condições de Karush-Khan-Tucker (KKT)

- Para encontrar um trio $(w*, \alpha^*, \beta^*)$ ótimo é necessário e suficiente que satisfação:

$$\frac{d}{dw_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0 \qquad i = 1, \cdots, n \tag{1}$$

$$\frac{d}{d\beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \qquad i = 1, \cdots, l \tag{2}$$

$$\alpha_i^*(g_i(w^*)) = 0, \qquad i = 1, \cdots, k \tag{3}$$

$$g_i(w^*) \leq 0, \qquad i = 1, \cdots, k \tag{4}$$

$$\alpha_i^* \geq 0, \qquad i = 1, \cdots, k \tag{5}$$

Aplicando no modelo SVM

- Relembrando o modelo:

$$\min_{w,b} \frac{1}{2}||w||^2$$
$$y^i(w^T x^i + b) \geq 1 \qquad\qquad i = 1, \cdots, m$$

- Portanto:

$$f(w,b) = \frac{1}{2}||w||^2$$
$$g_i(w,b) = 1 - y^i(w^T x^i + b)$$

- A função Lagrangeana é então:

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \cdot \left(y^i(w^T x^i + b) - 1\right)$$

Observando as condições de KKT

- Primeiramente derivando em relação a w:

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^i x^i = 0$$

- Portanto:

$$w = \sum_{i=1}^{m} \alpha_i y^i x^i$$

- Agora derivando em relação a b:

$$\frac{d}{db}\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^i = 0$$

Utilizando as condições de KKT para voltar ao modelo

- Vamos inserir os resultados anteriores para eliminar os fatos w e b:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \cdot \Big(y^i (w^T x^i + b) - 1 \Big)$$

- Usando o fato $w = \sum_{i=1}^{m} \alpha_i y^i x^i$:

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j (x^i)^T x^j - b \sum_{i=1}^{m} \alpha_i y^i$$

- Agora sabendo que $\sum_{i=1}^{m} \alpha_i y^i = 0$:

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j \langle x^i, x^j \rangle$$

Modelo dual

- Chegamos a um modelo que depende apenas da variável lagrangeana α:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m}\sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j \langle x^i, x^j \rangle$$

$$s.a : \alpha_i \geq 0, \; i = 1, \cdots, m \qquad \sum \alpha_i y^i = 0$$

- A partir do valor ótimo de α para este modelo podemos reestabelecer os valores ótimos para w e b.