Algoritmo SMO e *Kernels*

Carlos Diego Rodrigues

14 de dezembro de 2021

Universidade Federal do Ceará

Algoritmo SMO e *Kernels*

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

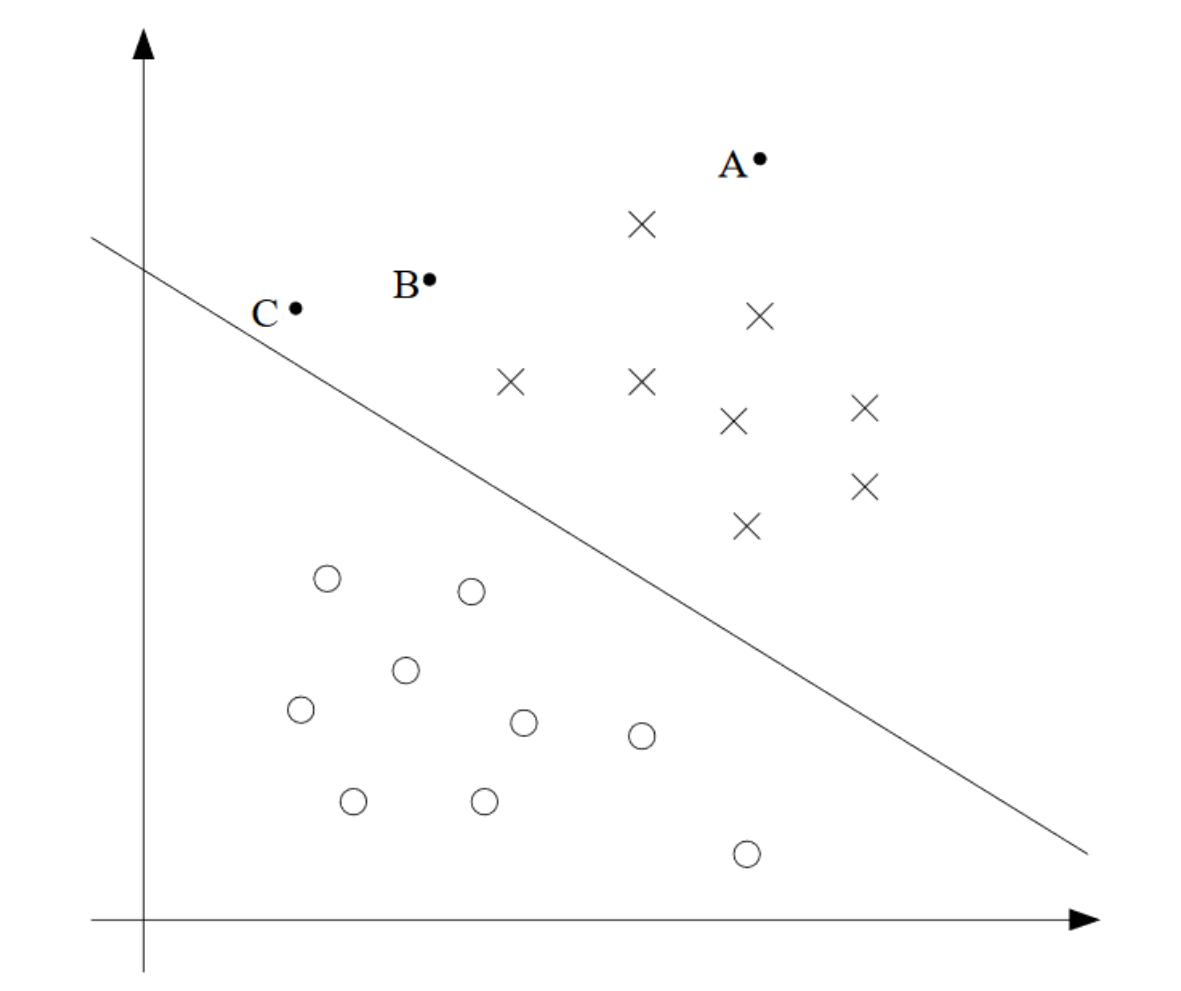
$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m}\sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j \langle x^i, x^j \rangle$$

$$s.a : \sum \alpha_i y^i = 0$$

$$\alpha_i \geq 0, \ i = 1, \cdots, m$$

- Vamos elaborar uma estratégia de otimização iterativa observando propriedades particulares do modelo e suas condições de otimalidade.
- O que ocorre com os casos que não são linearmente separáveis?
- *Kernels*

Relembrando o que queremos fazer



Voltando ao modelo

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

$$s.a : \sum \alpha_i y^i = 0$$

$$\alpha_i \geq 0, \ \ i = 1, \cdots, m$$

- Partindo de uma solução inicial (por exemplo $\alpha_i = 0$, $i = 1, \cdots, m$), observe que não é possível modificar apenas um único coeficiente sem violar a restrição

$$\sum \alpha_i y^i = 0$$

- Portanto, se quisermos mudar iterativamente de solução devemos considerar pelo menos dois exemplos (x^a, y^a) e (x^b, y^b) de tal forma que $y^a + y^b = 0$.

Desenvolvendo em função do par a, b

- Considerando exatamente dois exemplos podemos manipular a restrição, isolando os termos de α_a e α_b:

$$\alpha_a y^a + \alpha_b y^b = -\sum_{i \notin \{a,b\}} \alpha_i y^i$$

- Fixando todos os valores que de α que não sejam aqueles de a e b temos:

$$\alpha_a y^a + \alpha_b y^b = \beta$$

- Por esta igualdade, podemos ainda escrever tudo em função de α_b:

$$\alpha_b = y^b (\beta - \alpha_a y^a)$$

Modelo simplificado em uma única variável

- Assim o valor de

$$W(\alpha) = W(\alpha_1, \cdots, \alpha_a, \cdots, \alpha_b, \cdots, \alpha_m)$$

passa a

$$W(\alpha) = W(\alpha_1, \cdots, \alpha_a, \cdots, y^b\left(\beta - \alpha_a y^a\right), \cdots, \alpha_m)$$

.

- Considerando fixos os demais α_i para $i \notin \{a, b\}$ temos um modelo que depende de uma única variável $\alpha_a \geq 0$ com uma simples função quadrática $A\alpha_a^2 + B\alpha_a + C$ a ser otimizada.
- Naturalmente o valor ótimo desta função ocorre onde

$$\alpha_a = \frac{-B}{2A}$$

quando $\frac{-B}{2A} \geq 0$.

Algoritmo SMO

Algorithm 1 SMO

Entrada: X, Y: conjunto de atributos e classes

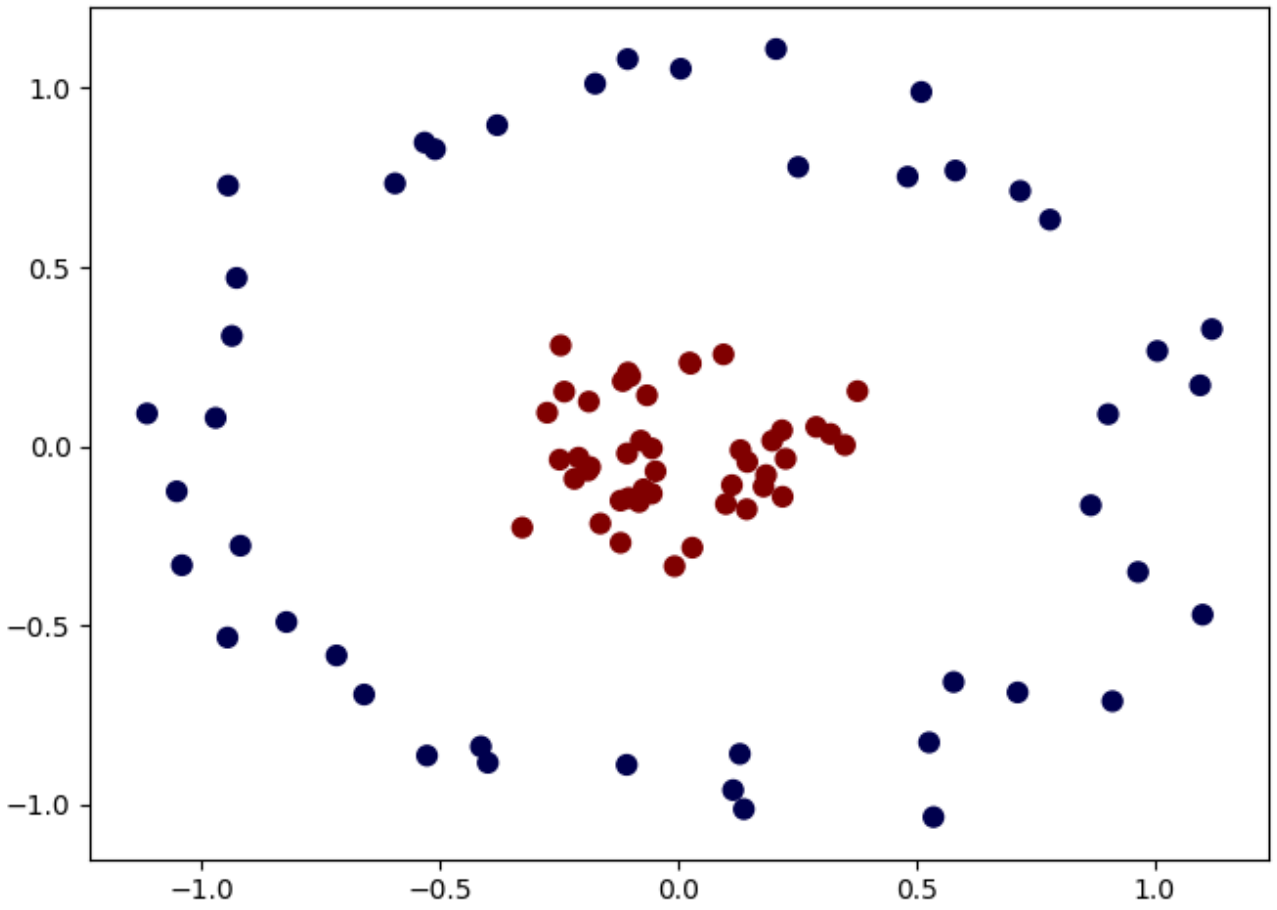
Saída: α: coeficientes lagrangeanos

1: Encontre um valor inicial para α
2: **repeat**
3: Escolha dos exemplos (x^a, y^a) e (x^b, y^b).
4: Faça $\alpha_b = y^b (\beta - \alpha_a y^a)$.
5: Encontre e otimize a equação $A\alpha_a^2 + B\alpha_a + C$.
6: Atualize os valores de α_a e α_b.
7: **until** convergência
8: Retorna α

Casos não linearmente separáveis

- Podemos ter casos onde uma separação linear seria inviável para a classificação.
- No caso do SVM isto indicaria um modelo inviável que não resultaria em qualquer solução.
- Classicamente existem duas propostas para lidar com estes casos:
  1. uma relaxação do modelo original, aplicando penalidades aos casos erroneamente classificados;
  2. uma transformação dos pontos originais para um espaço de dimensão ampliada, onde neste espaço eles sejam linearmente separáveis.
- É possível combinar ambos os métodos!

Exemplo não-separável



Penalidades

- Retomando o modelo original:

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

- Vamos adicionar fatores que permitam, para cada exemplo, que a sua restrição seja violada, mas estes fatores vão ter um custo na função objetivo.

$$\min_{w,b} \frac{1}{2}||w||^2 + C \cdot \sum_{i=1}^{m} \epsilon_i$$
$$y^i(w^T x^i + b) \geq 1 - \epsilon_i \qquad i = 1, \cdots, m$$
$$\epsilon_i \geq 0 \qquad i = 1, \cdots, m$$

Aplicando a decomposição Lagrangeana

- O modelo dualizado é:

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

$$s.a : \sum \alpha_i y^i = 0$$

$$0 \leq \alpha_i \leq C, \ i = 1, \cdots, m$$

- Toda a ideia do algoritmo SMO permanece.
- Temos que averiguar a restrição $\alpha_i \leq C$ quando vamos resolver para uma única variável.

Kernels

- O espaço vetorial dos atributos originais das instâncias pode não oferecer a possibilidade de uma separação linear.
- Podemos construir um mapeamento do espaço dos atributos para um espaço expandido, chamado espaço de características. Isto é feito aplicando-se uma função ϕ a cada instância.
- Como vimos anteriormente, nossos modelos e algoritmos podem ser escritos dependendo quase que exclusivamente do produto interno dos atributos $\langle x^i, x^j \rangle$.
- Definimos a função **Kernel** como o produto interno das características:

$$K(x^i, x^j) = \langle \phi(x^i), \phi(x^j) \rangle$$

Efeitos do *Kernel*

- Portanto a qualquer momento em nosso desenvolvimento se trocarmos $\langle x^i, x^j \rangle$ por $K(x^i, x^j)$ estaremos aplicando o SVM sobre um outro espaço, mais adequado à separação linear.
- Isto também é compatível com a ideia das penalidades, agora no espaço de características.
- O chamado truque dos *Kernels* acontece quando a função $K(x^i, x^j)$ é fácil de calcular, porém as transformações $\phi(x^i)$ e $\phi(x^j)$ são complexas (e desnecessárias!)
- Podemos portanto trabalhar apenas com o produto vetorial $K(x^i, x^j)$ como uma medida de similaridade entre os objetos x^i e x^j sem necessariamente expandi-los.

Condições para uma função *Kernel*

- Como saber se existe um mapeamento de características ϕ para uma determinada função candidata $K(x^i, x^j)$ a um Kernel?
- **Teorema de Mercer**: Dada uma função $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, dizemos que K é um *Kernel* se, e somente se, para quaisquer pontos $x^1, \cdots, x^m$, com $m < \infty$ a matriz de kernel

$$\begin{bmatrix} K(x^1, x^1) & K(x^1, x^2) & \cdots & K(x^1, x^m) \\ K(x^2, x^1) & K(x^2, x^2) & \cdots & K(x^2, x^m) \\ \vdots & \vdots & \ddots & \vdots \\ K(x^m, x^1) & K(x^m, x^2) & \cdots & K(x^m, x^m) \end{bmatrix}$$

é semidefinida positiva ($x^T K x \geq 0, \ \forall x \in \mathbb{R}^m$).

Exemplos de *Kernels*

- *Kernel* polinomial:

$$K(x^i, x^j) = \left(\langle x^i, x^j \rangle + c\right)^d$$

(espaço de características com $\binom{n+d}{d}$ dimensões)

- *Kernel* gaussiano:

$$K(x^i, x^j) = exp\left(\frac{\langle x^i - x^j, x^i - x^j \rangle}{2\sigma^2}\right)$$

(espaço de características infinito!)