UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Física
Bacharelado em Engenharia Física
Tópicos especiais de engenharia física

Deomar Santos da Silva Junior / 00260682

Algoritmo de *Support Vector Machines* para classificação de diferentes espécies da flor Iris

Introdução

O *dataset* escolhido é de informações acerca de três espécies diferentes de flores Iris [1]. É um *dataset* bem conhecido para aplicações inicias de algoritmos de classificação, pois pode ser separado de forma linear e, também, é possível observar visualmente as separações gerando assim um bom *dataset* para exemplificação e ensino de algoritmos de *machine learning* de classificação.

Support Vector Machines (SVM)

A técnica de SVM basicamente define bordas entre diferentes categorias do mesmo conjunto de dados. Essas bordas podem ser lineares ou não lineares, caso no qual é utilizada uma função kernel para defini-las.

Por exemplo, se há um conjunto de dados definidos por duas *features* (x1 e x2) - figura 1 [2], e, dentro desse conjunto de dados, há duas classificações, dados do grupo azul e dados do grupo verde, que evidentemente podem ser separados por uma reta, pode-se então utilizar a técnica de SVM para definir a melhor reta que separa os dois conjuntos, assim é possível distinguir qualquer dado que pertença à fronteira de separação.

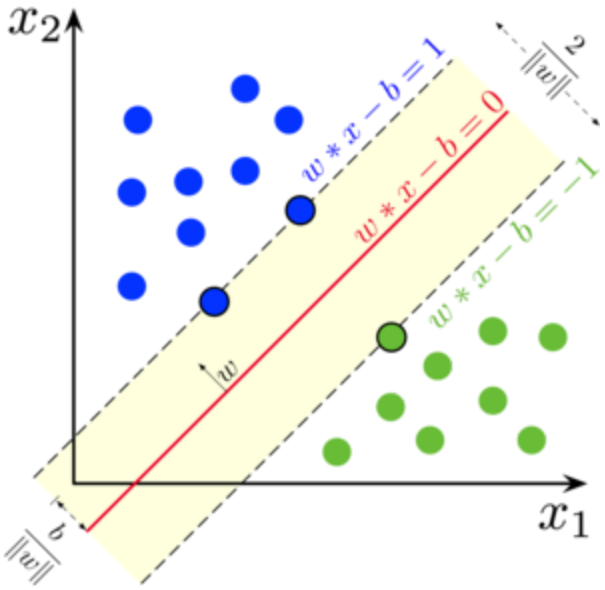

Figura 1. *Dataset* genérico com duas *features* e dois tipos de dados [2].

Dentro do contexto de análise de dados, pode-se "chutar" uma reta entre os pontos e, então, maximizar a distância dessa reta entre os pontos dos dois conjuntos de dados.

Matematicamente, a reta que separa os planos é escrita pela equação (para mais dimensões se define um hiperplano):

$$wx_i - b = 0$$

Onde ***w*** é um vetor perpendicular à reta de separação, **x** é um ponto do conjunto de dados que está exatamente em cima da reta e **b** uma constante.

Pode-se, então, definir outras duas retas que delimitam as regiões dos conjuntos azul e verde. De forma que utiliza-se a mesma equação acima com uma pequena modificação.

$$wx_i - b = 1$$

Para a reta que delimite o conjunto azul e:

$$wx_i - b = -1$$

Para a reta que delimita o conjunto verde. Novamente os pontos ***x*** estão exatamente em cima das retas (figura 1).

Logo pode-se inferir facilmente que para qualquer outro ponto do conjunto azul:

$$wx_i - b \geq 1$$

e

$$wx_i - b \leq -1$$

Para qualquer ponto do conjunto verde.

Introduzindo-se a variável y_i para identificar o ponto ***i*** no dois conjuntos de forma que y=1 para pontos no conjunto azul e y = -1 para pontos no conjunto verde.

Multiplicando essa variável pela equação das retas, tem-se, para os dois casos:

$$y_i(wx_i - b) - 1 \geq 0$$

E:

$$y_i(wx_i - b) - 1 = 0$$

Para os pontos x_i entre as bordas.

Como se quer a reta central que mais se distancia dos dois conjuntos de dados, introduz-se a distância euclidiana entre as retas de borda de forma a encontrar a máxima distância de separação. Então, dado um ponto no conjunto azul (x_{azul}) em cima da borda e um ponto no conjunto verde (x_{verde}). pode-se definir a distância entre as retas como a distância (dist) entre os pontos, na direção de ***w*** como:

$$dist = (x_{azul} - x_{verde})w/|w|$$

Substituindo-se os valores na inequação geral acima, tem-se que:

$$x_{azul} = 1 - b \text{ e } x_{verde} = 1 + b$$

Substituindo na distância:

$$dist = 2w/|w|$$

Ou seja, a distância entre os planos é 2/|w|.

Logo, a maximização da distância entre os planos envolve a restrição acima. Para a maximização com restrições se utiliza o multiplicador de Lagrange (L) que, neste caso, tem a seguinte forma:

$$L = (1/2)||w||^2 + \sum_{i}^{N} \alpha_i[y_i w.x_i + b) - 1]$$

Onde α são os multiplicadores e N o número de pontos do *dataset*.

Diferenciando-se L pelos seus dois parâmetros ***w*** e ***b***, e após algumas operações algébricas, chega-se a, para a primeira diferenciação:

$$w = \sum_{i}^{N} \alpha_i y_i x_i$$

E, para a segunda diferenciação:

$$\sum_{i}^{N} \alpha_i y_i = 0$$

Substituindo na primeira equação do multiplicador, obtém-se:

$$L = \sum_{i}^{N} \alpha_i + (1/2) \sum_{i}^{N} \sum_{j}^{N} \alpha_i \alpha_j y_i y_j x_i x_j$$

Que pode ser entendida como a função de custo que depende apenas dos dados ***x*** e dos multiplicadores α.

Logo, utiliza-se a função de custo para obter os multiplicadores α e, por consequência, encontrar o vetor ***w*** perpendicular ao plano.

Para encontrar o conjunto de alfas iterativamente se utiliza o método do gradiente descendente de forma que a iteração de α_i no tempo ***i*** para α_{i+1} é dada por:

$$\alpha_{i+1} = \alpha_i - \eta dL/d\alpha_i$$

Onde η é a taxa de ajuste do gradiente. α deve respeitar a condição:

$$\sum_{i}^{N} \alpha_i y_i = 0$$

Ou seja, é necessário normalizar cada novo α iterado.

Após encontrar os multiplicadores, calcula-se o vetor ***w*** que separa os conjuntos através de:

$$w = \sum_{i}^{N} \alpha_i y_i x_i x_i$$

Implementação do algoritmo para classificação do *dataset* Iris

Com o objetivo de explorar o algoritmo se estudou o *dataset* de três espécies de flores Iris [1]. Informações acerca das dimensões das flores são avaliadas como as dimensões das sépalas.

Para simplificação e aplicação do modelo, o algoritmo foi aplicada apenas para duas espécies de flores a partir da classificação de duas features (figura 2).

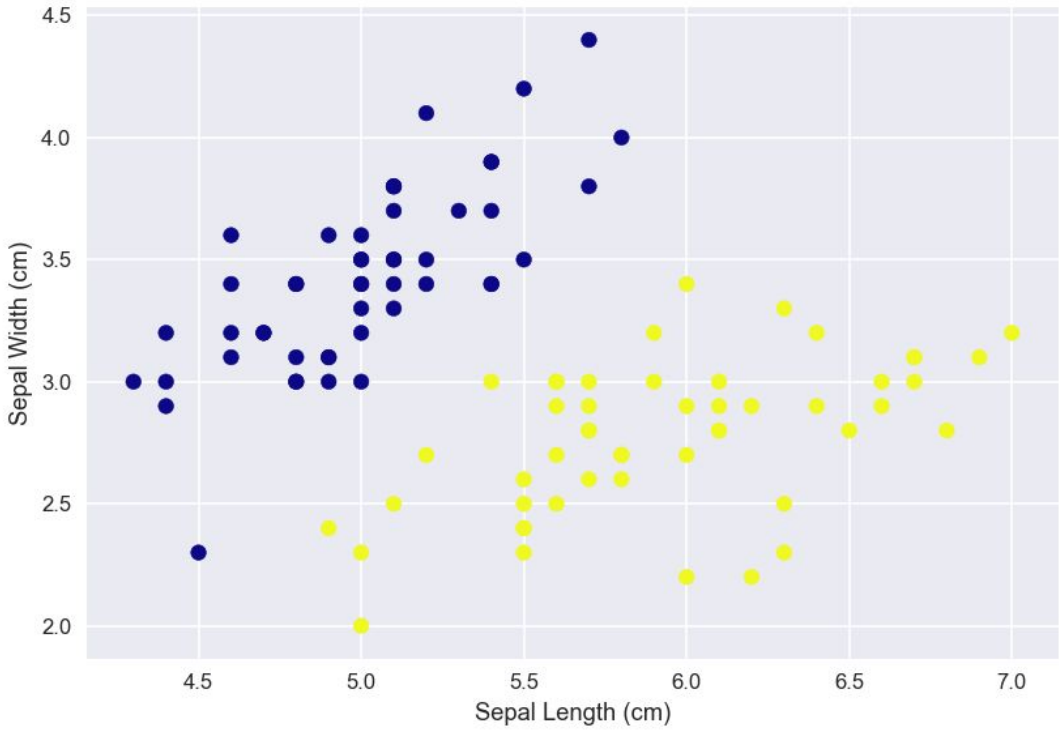

Figura 2. Duas classes de dados do *dataset* Iris.

A cor azul representa a espécie 1 enquanto a cor amarela representa a espécie 2. Observa-se que o *dataset* é bem separado visualmente.

Com a finalidade de mostrar visualmente a separação dos dados encontrada pelo algoritmo, pode-se as seguintes operações algébricas de forma a dividir os dados por uma constante em vez de uma reta com equação definida.

Para isso, sabe-se que os pontos acima da primeira reta respeitam a seguinte condição:

$$wx_i - b \geq 1$$

E que ***w*** é dado por:

$$w = \sum_{i}^{N} \alpha_i y_i x_i x_i$$

Então tem-se:

$$\sum_{i}^{N} \alpha_i y_i x_i x_i \; . \; x_i \geq 1 + b$$

Como b é uma constante, reescreve-se para uma nova constante b' de forma que:

$$\sum_{i}^{N} \alpha_i y_i x_i x_i \; . \; x_i \geq b'$$

Ou seja, ao plotar o lado esquerdo da equação para todos os pontos x_i, uma das espécies deve estar acima da constante b' enquanto os outros pontos abaixo.

Aplicando as equações do modelo descrito acima, iterando o alfa inicial a uma taxa de 0.0001 por 200 iterações, plota-se o lado esquerdo da equação para o conjunto de dados analisado (figura 3):

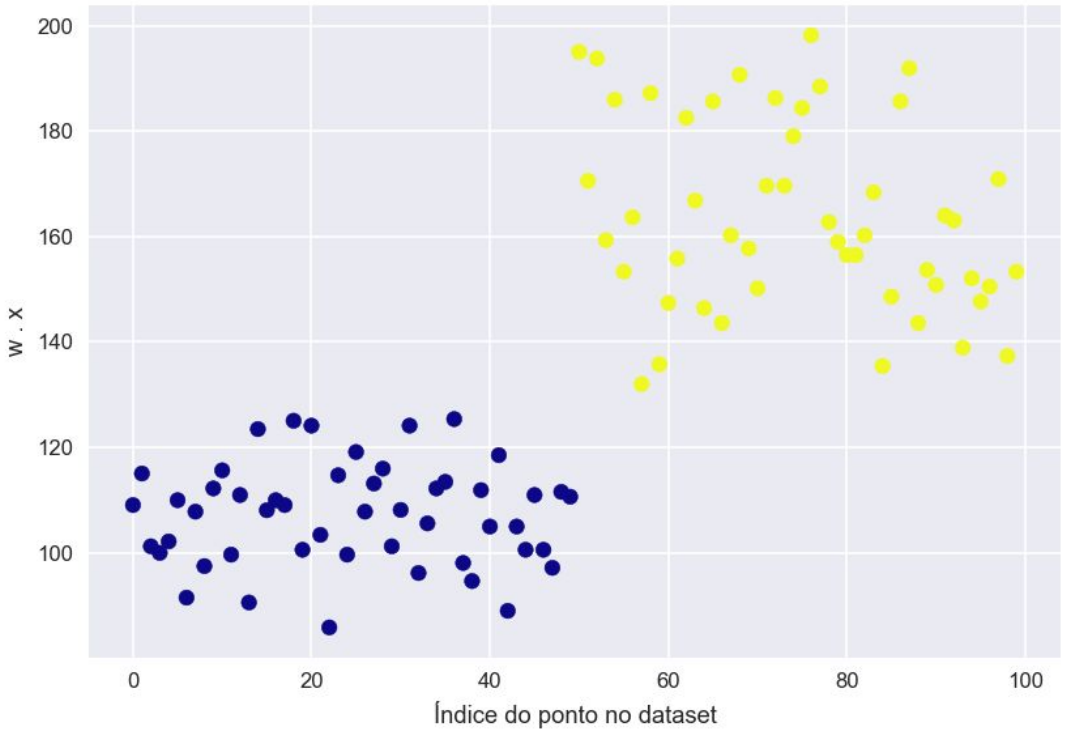

Figura 3. Plot do produto escalar do vetor ‘w’ pelos pontos do dataset.

Como resultado, observa-se que os dados foram separados em dois grupos por uma constante que, grosseiramente, pode ser aproximada por 125.

Observação: O eixo x apenas representa o índice do ponto dentro do dataset, então nenhuma interpretação deve ser feita nesse eixo.

Para classificar novos pontos, pode-se resolver o lado esquerdo da equação acima e classificar os novos pontos dentro das duas categorias de espécie se estiverem acima ou abaixo da constante b’.

Conclusão

O algoritmo separou o conjunto de dados em duas classificações diferentes em relação a uma constante permitindo que uma classificação futura de novos dados seja feita. Após as operações algébricas realizadas, pode-se mostrar visualmente o resultado do algoritmo.

Referências:

[1] https://www.kaggle.com/uciml/iris <Disponível em 20/07/2020>

[2] https://en.wikipedia.org/wiki/Support_vector_machine <Disponível em 20/07/2020>