Machine Learning

Data Science Academy

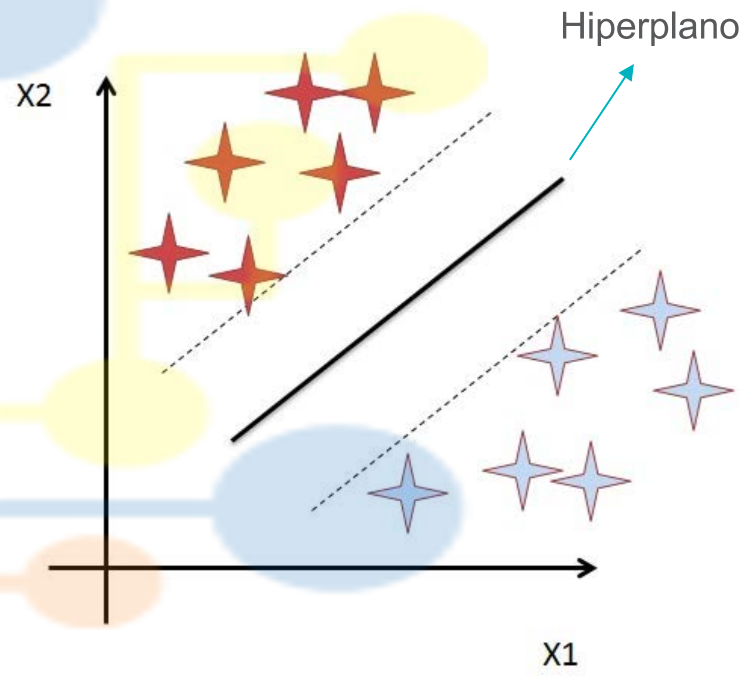
Seja muito bem-vindo(a)!

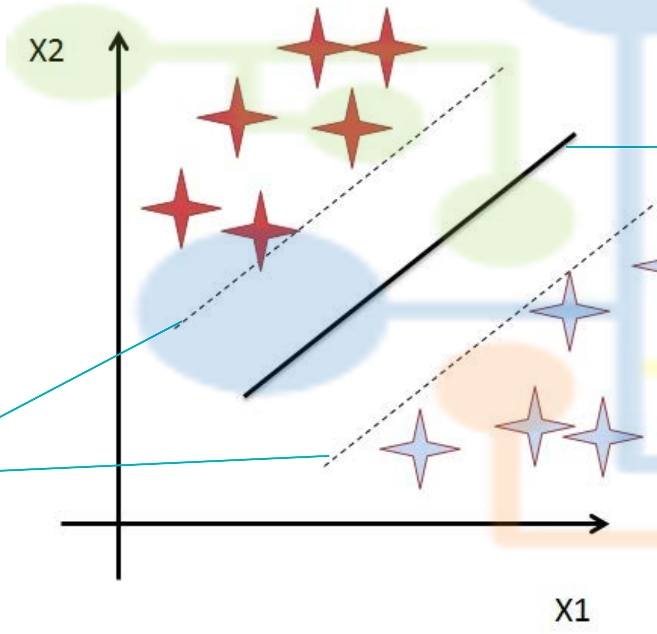
Data Science Academy

O Que São SVMs?

Regressão Linear

SVM


Hiperplano
X2
X1



Hiperplano de separação das classes

Margens

Algumas características das SVM's:

- Em caso de outliers o modelo SVM busca a melhor forma possível de classificação e, se necessário, desconsidera o outlier;
- É um classificador criado para fornecer separação linear;
- Funciona muito bem em domínios complicados, em que existe uma clara margem de separação;
- Não funciona bem em conjuntos de dados muito grandes, pois o tempo de treinamento é muito custoso;
- Não funciona bem em conjuntos de dados com grande quantidade de ruídos;
- Se as classes estiverem muito sobrepostas deve-se utilizar apenas evidências independentes.

Data Science Academy

Machine Learning

Data Science Academy

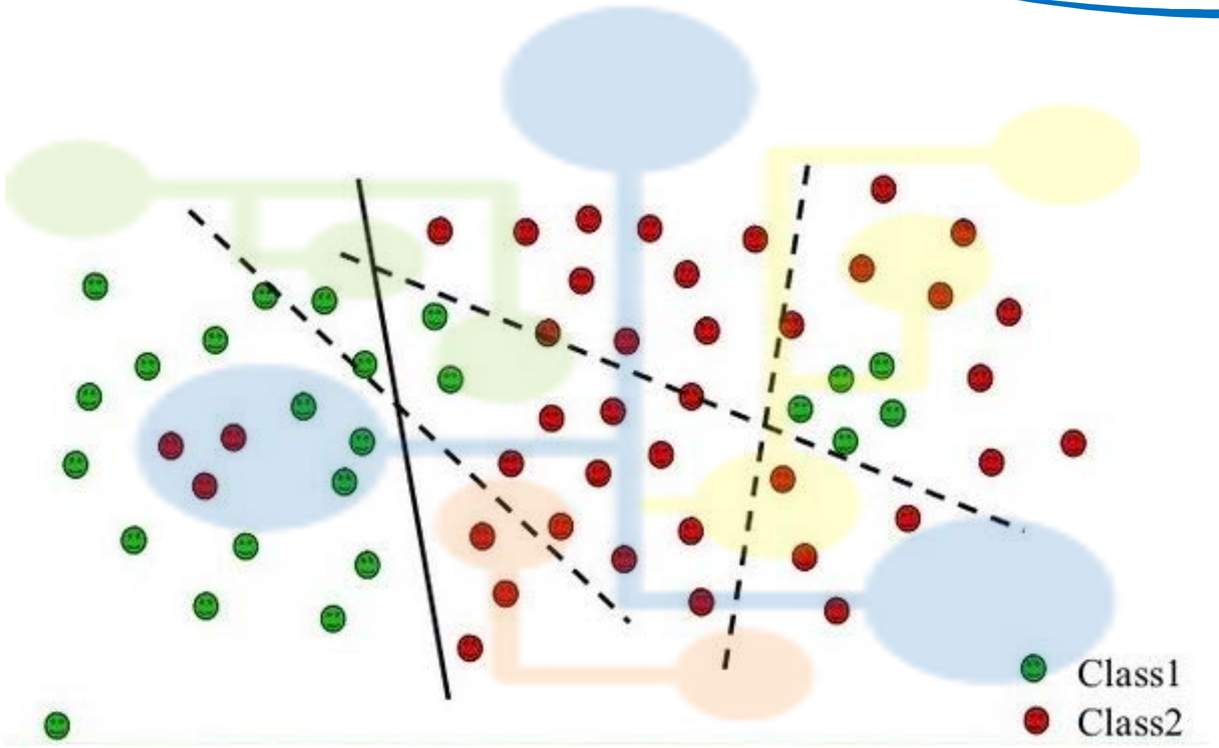
Funcionamento das SVMs

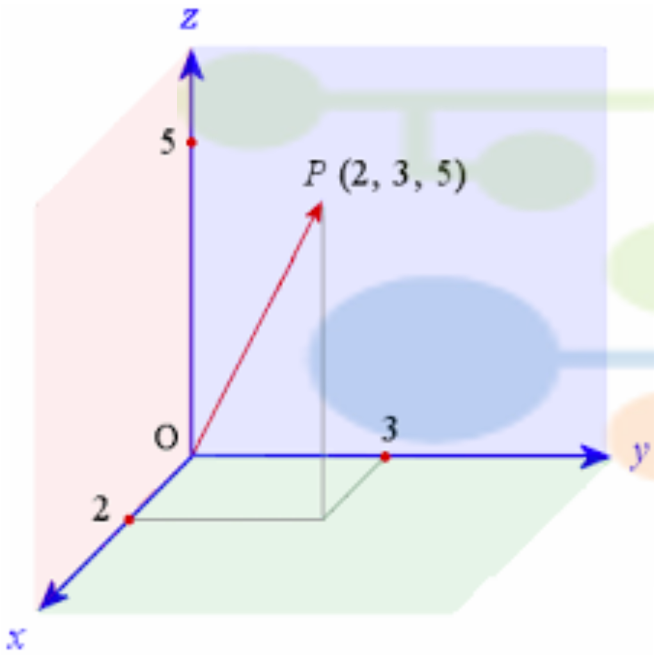
Support Vector Machines (SVM's) são modelos de aprendizagem supervisionada, que possuem algoritmos de aprendizagem que analisam dados e reconhecem padrões, utilizados para classificação e análise de regressão.

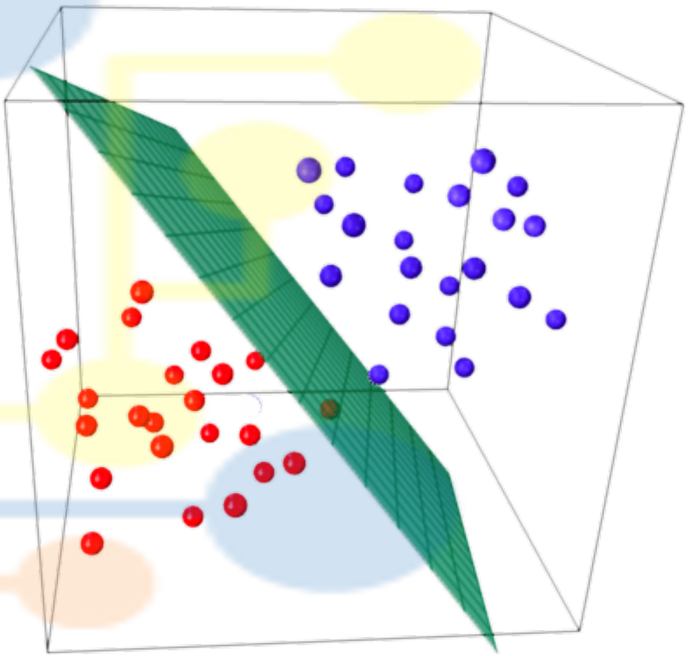
Funcionamento do Modelo SVM para dados linearmente separáveis.

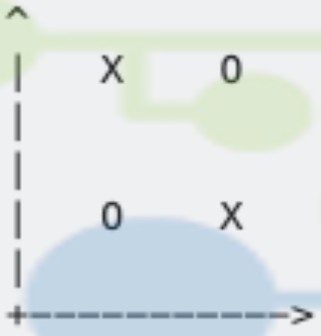
Dados Não Linearmente Separáveis

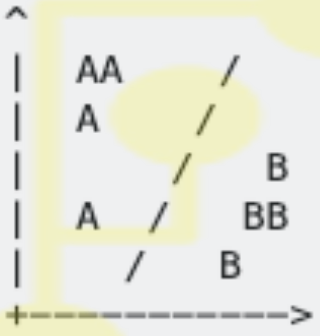
Dados Linearmente Separáveis


Class1
Class2


z
5
P (2, 3, 5)
O
3
y
2
x





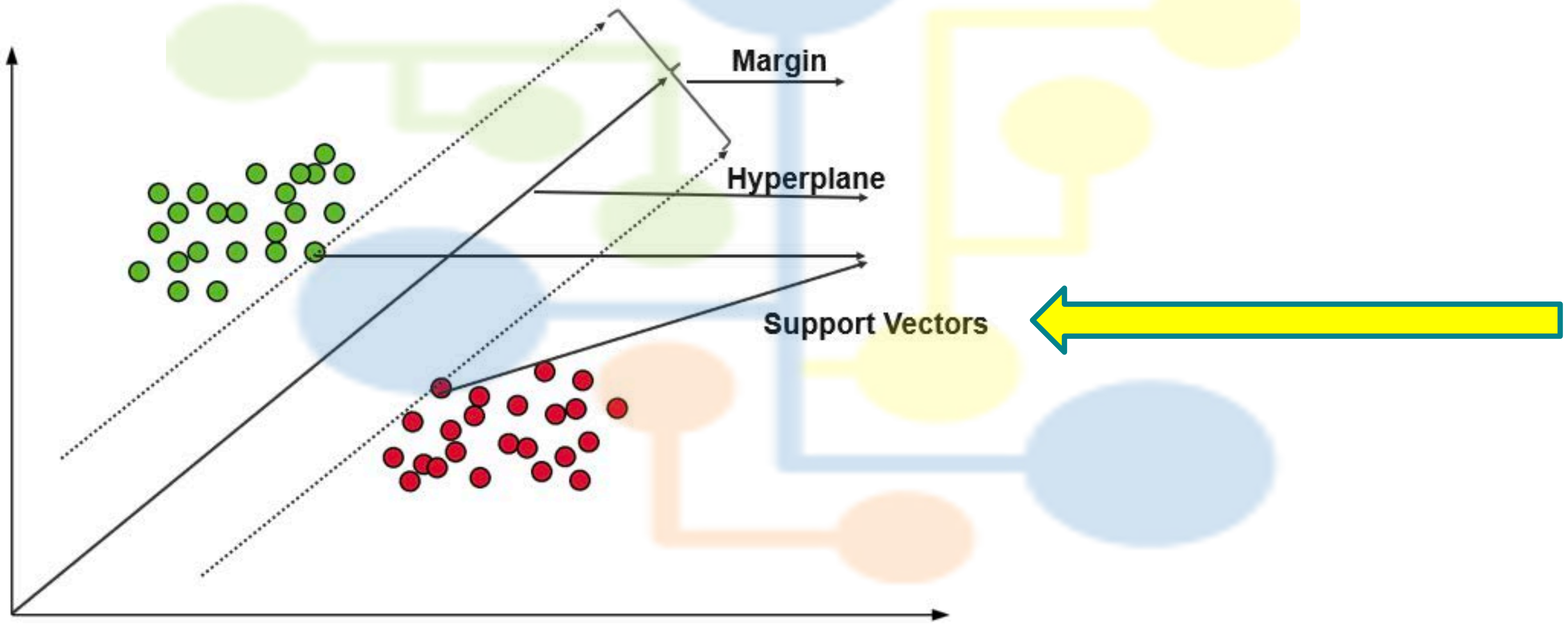


Dados Não Linearmente Separáveis

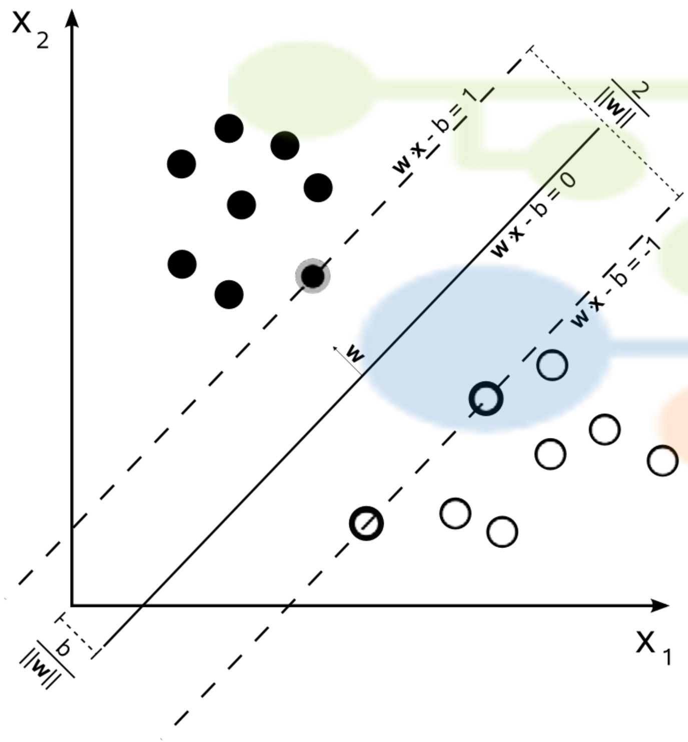
Dados Linearmente Separáveis

Precisamos de Função de Kernel Para a Separação

Funcionamento do Modelo SVM para dados linearmente separáveis.


Margin
Hyperplane
Support Vectors

Uma Dose de Matemática



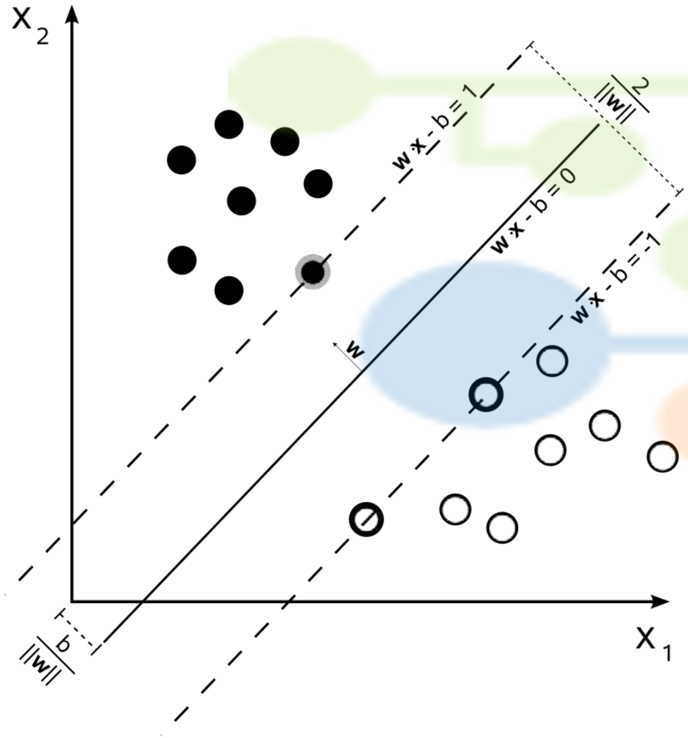
Encontrar o valor de y:

$$y^{(i)} = \begin{cases} -1 & \text{if } \mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b \le -1 \\ 1 & \text{if } \mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b \ge 1 \end{cases}$$

Distância Mínima Entre os VS:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{\|\mathbf{w}\|}{2} + C\sum_{i=1}^{N} \xi^{(i)}, \\ s.t. \quad & y^{(i)}(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b) \ge 1 - \xi^{(i)}, \qquad \forall i \in \{1, \dots, N\} \\ & \xi^{(i)} \ge 0, \qquad \forall i \in \{1, \dots, N\} \end{aligned}$$

Uma Dose de Matemática



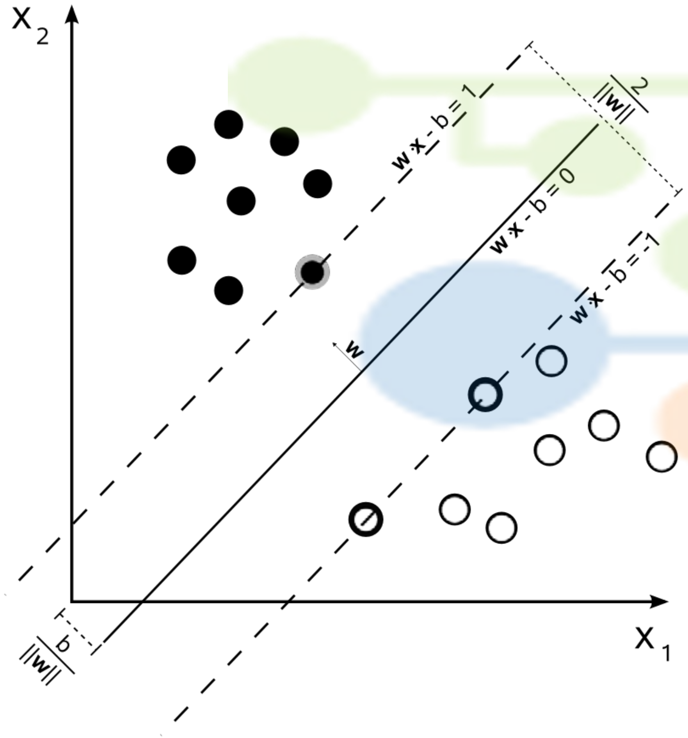
Distância Mínima Entre os VS:

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{\|\mathbf{w}\|}{2} + C\sum_{i=1}^{N} \xi^{(i)}, \\
s.t. \quad & y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)}, \qquad \forall i \in \{1,\ldots,N\} \\
& \xi^{(i)} \geq 0, \qquad \forall i \in \{1,\ldots,N\}
\end{aligned}
$$

Maximizar a Distância Mínima (Otimização):

$$
\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^{N} \alpha^{(i)} - \sum_{i=1}^{N}\sum_{j=1}^{N} \left(y^{(i)}\alpha^{(i)}\phi\left(\mathbf{x}^{(i)}\right)^{T}\phi\left(\mathbf{x}^{(j)}\right) y^{(j)}\alpha^{(j)} \right) \\
s.t. \quad & 0 \leq \alpha^{(i)} \leq C,
\end{aligned}
$$

Uma Dose de Matemática



Maximizar a Distância Mínima (Otimização):

$$\max_{\alpha} \quad \sum_{i=1}^{N} \alpha^{(i)} - \sum_{i=1}^{N}\sum_{j=1}^{N} \left(y^{(i)} \alpha^{(i)} \underbrace{\phi\left(\mathbf{x}^{(i)}\right)^{T} \phi\left(\mathbf{x}^{(j)}\right)} y^{(j)} \alpha^{(j)} \right)$$

$$s.t. \quad 0 \leq \alpha^{(i)} \leq C,$$

Inner Product
O Kernel Trick é este mapeamento.

Coeficiente aprendido durante o treinamento, um para i e outro para j.

Uma Dose de Matemática

O produto escalar entre dois vetores (dot product) mostra como os vetores são "semelhantes". Se os vetores representam pontos no seu conjunto de dados, o produto escalar informa se eles são semelhantes ou não.

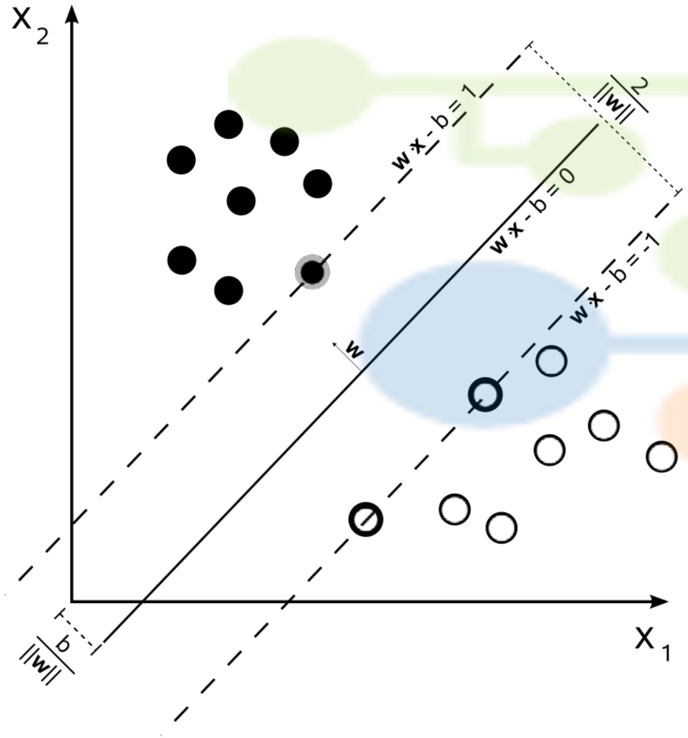
Mas, em alguns (muitos) casos, o produto escalar não é a melhor métrica de similaridade.

Por exemplo:

Talvez os pontos com produto escalar baixo sejam semelhantes por outras razões. Você pode ter itens de dados que não estão bem representados como pontos ou pode não haver separação linear.

Então, em vez de usar o produto escalar, você usa um "kernel", que é apenas uma função que recebe dois pontos e fornece uma medida de sua similaridade. O SVM aplica esse conceito que é denominado Truque do Kernel (Kernel Trick).

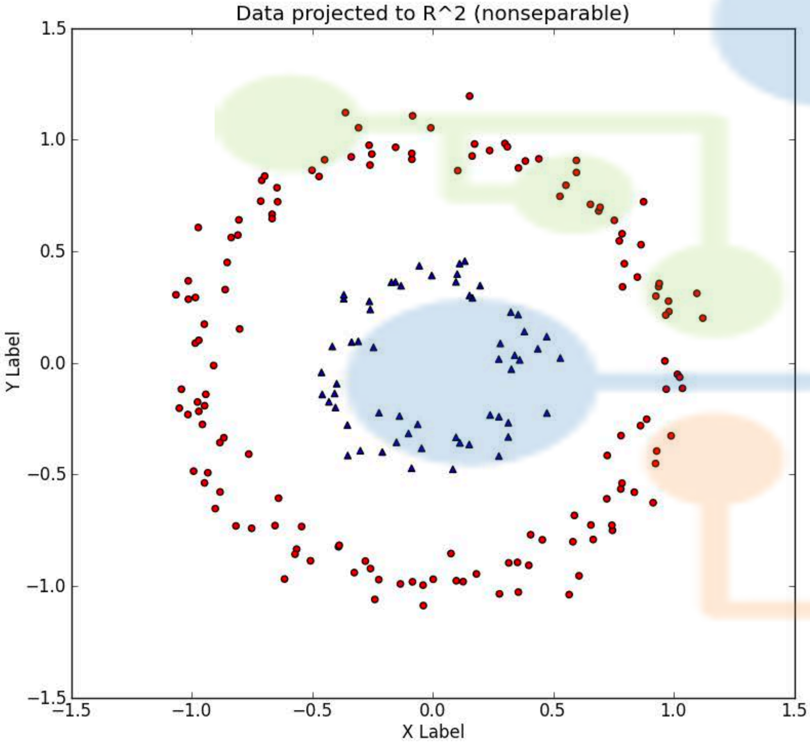
Uma Dose de Matemática

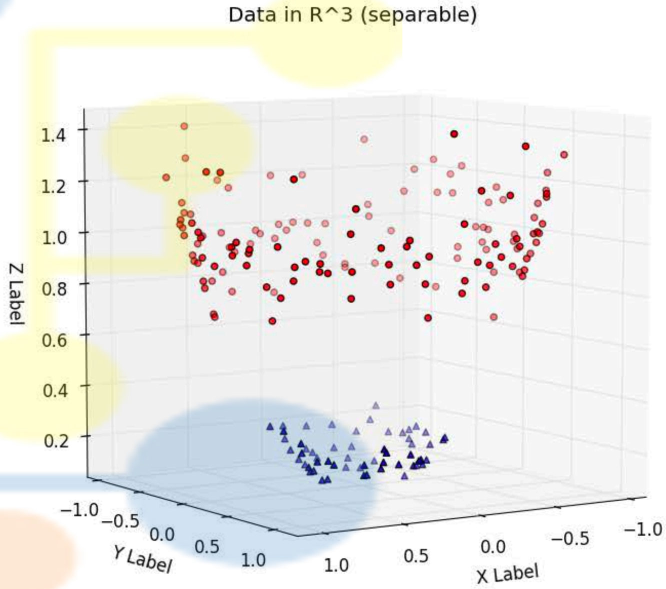


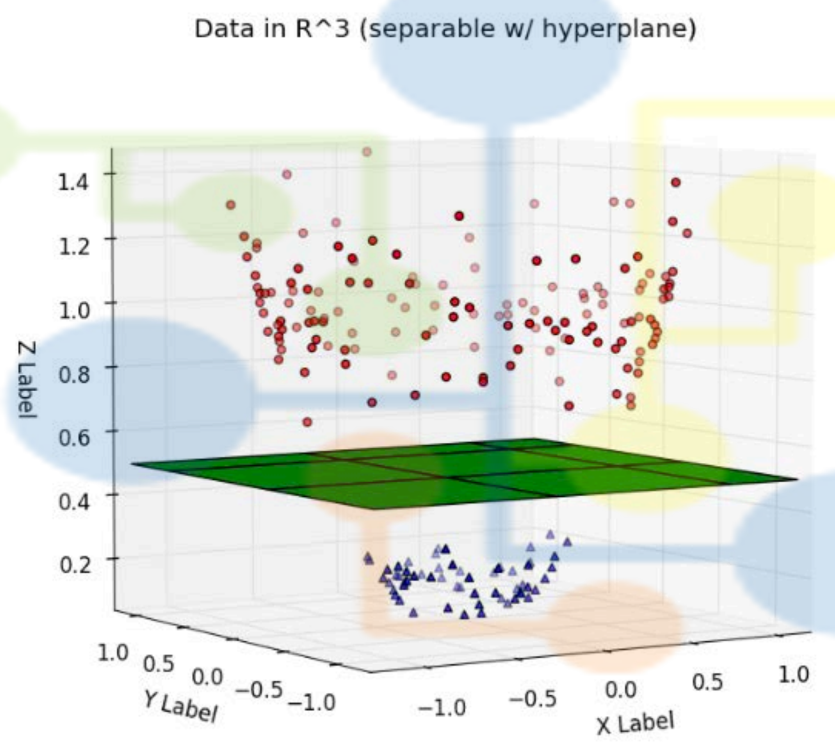
Previsões do Modelo:

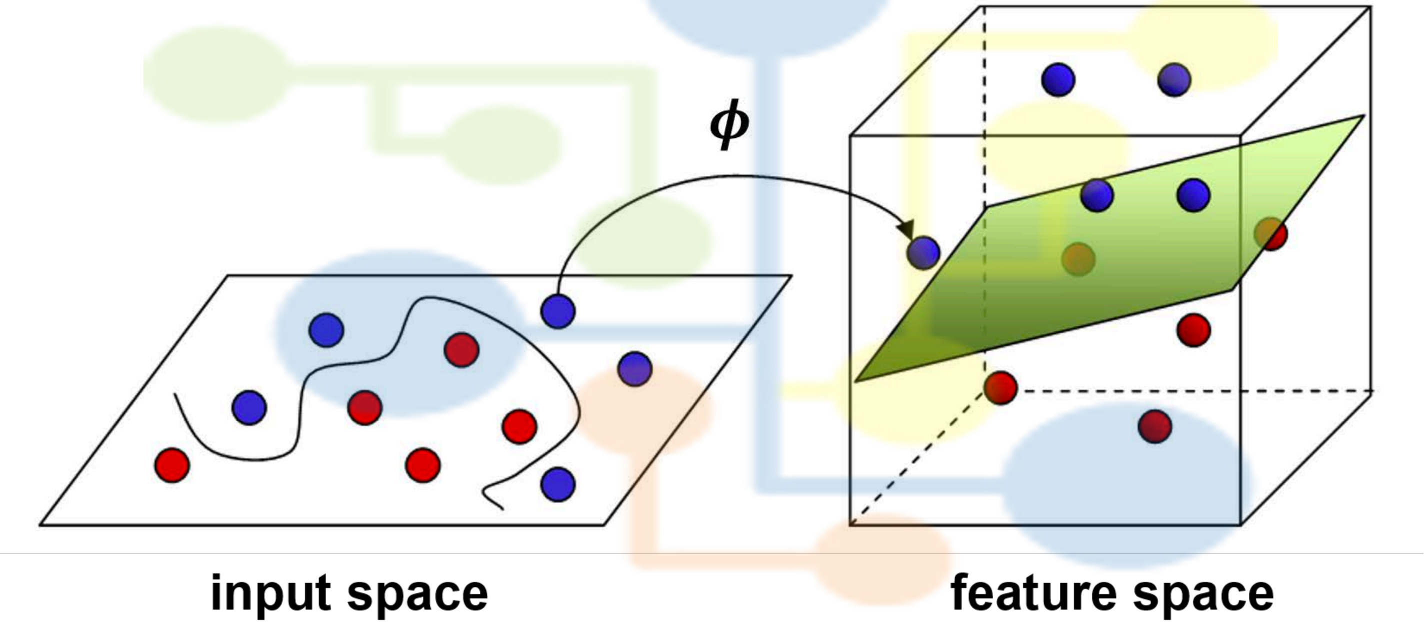
$$
\begin{aligned}
y^{\text{test}} &= \text{sign}\left(\mathbf{w}^{\mathbf{T}}\phi\left(\mathbf{x}^{\text{test}}\right) + b\right) \\
&= \text{sign}\left(\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \phi\left(x^{(i)}\right)^{T} \phi\left(\mathbf{x}^{\text{test}}\right) + b\right)
\end{aligned}
$$

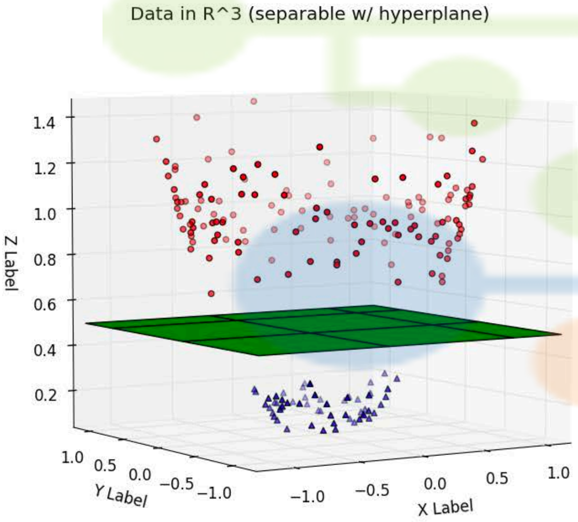
Funcionamento do Modelo SVM para dados NÃO linearmente separáveis.

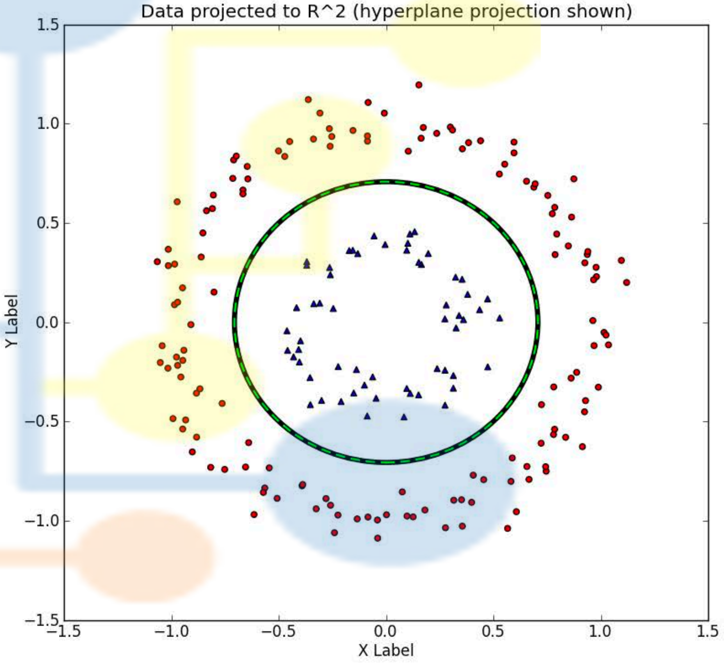

Data projected to R^2 (nonseparable)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Y Label
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X Label


Data in R^3 (separable)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Z Label
−1.0
−0.5
0.0
0.5
1.0
Y Label
1.0
0.5
0.0
−0.5
−1.0
X Label


Data in R^3 (separable w/ hyperplane)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Z Label
1.0
0.5
0.0
−0.5
−1.0
Y Label
−1.0
−0.5
0.0
0.5
1.0
X Label


ϕ
input space
feature space


Data in R^3 (separable w/ hyperplane)
Z Label
Y Label
X Label


Data projected to R^2 (hyperplane projection shown)
Y Label
X Label

Machine Learning

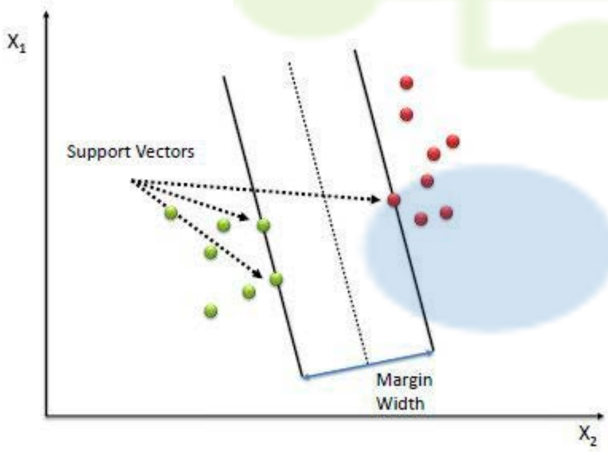
Data Science Academy

SVM's com Margens Rígidas

x

SVM's com Margens Flexíveis

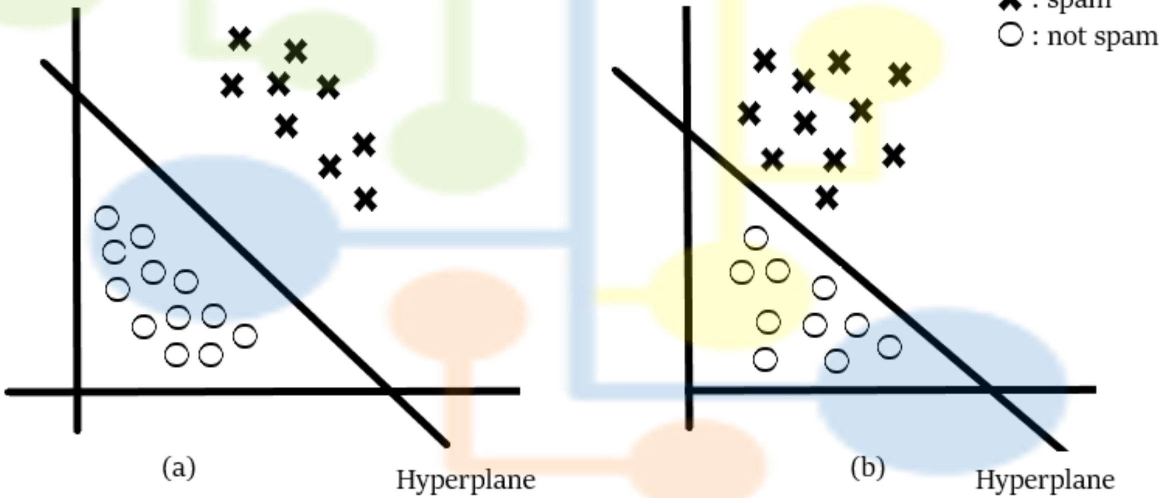
As máquinas de vetores de suporte (chamadas SVMs) são um algoritmo de aprendizado supervisionado que pode ser usado para problemas de classificação e regressão como classificação de vetores de suporte (SVC) e regressão de vetores de suporte (SVR).



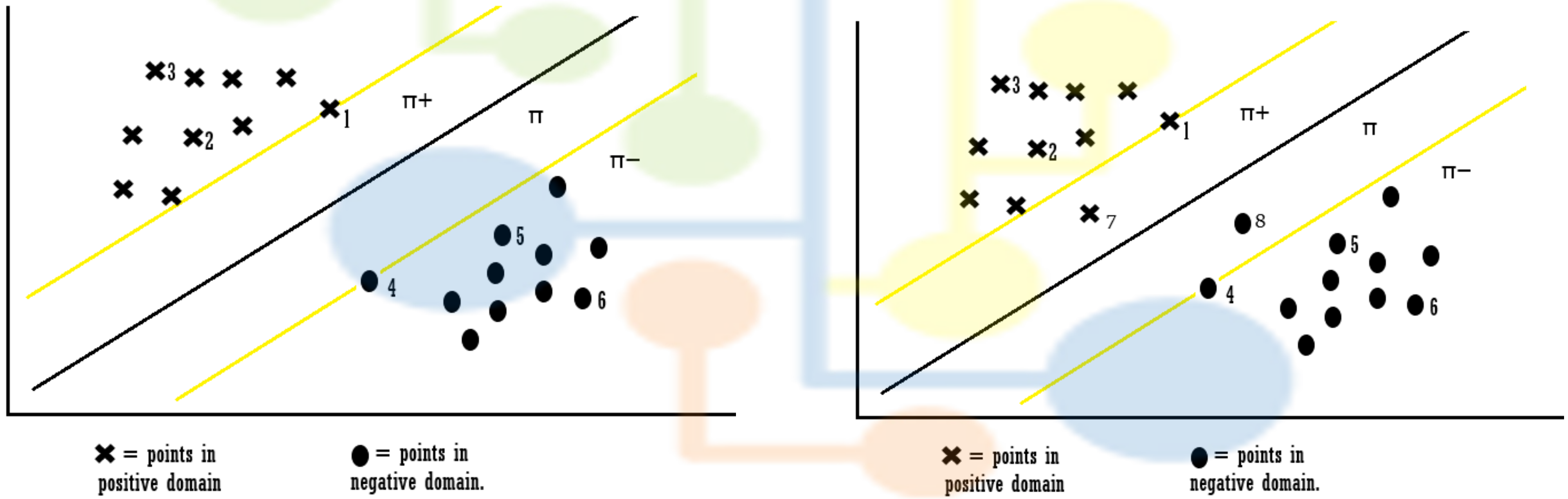
Os pontos mais próximos ao hiperplano são chamados de pontos do vetor de suporte e a distância dos vetores do hiperplano é chamada de margem.

A intuição básica a ser desenvolvida aqui é que quanto mais pontos SV adicionais, do hiperplano, maior a probabilidade de classificar corretamente os pontos em suas respectivas regiões ou classes. Os pontos SV são muito críticos na determinação do hiperplano porque se a posição dos vetores muda, a posição do hiperplano é alterada.

Tecnicamente, esse hiperplano também pode ser chamado de hiperplano de maximização de margem.


: spam
: not spam
(a)
Hyperplane
(b)
Hyperplane

Margens Rígidas



Margens Rígidas

Se os pontos são linearmente separáveis, apenas o nosso hiperplano é capaz de distinguir entre eles e se algum erro for introduzido (outliers por exemplo), não será possível separá-los.

Esse tipo de SVM é chamado SVM de Margem Rígida (já que temos restrições muito rígidas para classificar corretamente cada ponto de dados).

Margens Flexíveis

Basicamente, consideramos que os dados são linearmente separáveis e isso pode não ser o caso no cenário da vida real.

Precisamos de uma atualização para que nossa função possa pular alguns valores discrepantes e poder classificar pontos quase linearmente separáveis. Por esse motivo, apresentamos uma nova variável Slack (ξ) chamada Xi.

Distância Mínima Entre os VS:

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{\|\mathbf{w}\|}{2} + C\sum_{i=1}^{N} \xi^{(i)}, & \\
s.t. \quad & y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)}, & \forall i \in \{1, \ldots, N\} \\
& \xi^{(i)} \geq 0, & \forall i \in \{1, \ldots, N\}
\end{aligned}
$$

Margens Flexíveis

Se ξi = 0, os pontos podem ser considerados corretamente classificados.
Senão, se ξi > 0, pontos são classificados incorretamente.

Portanto, se ξi > 0 significa que Xi (variáveis) está na dimensão incorreta, podemos pensar em ξi como um termo de erro associado a Xi (variável).

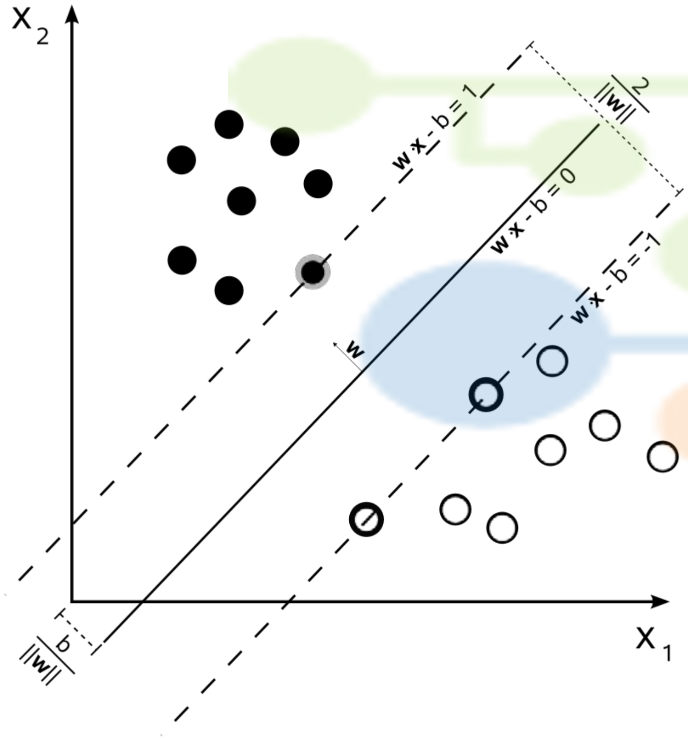
$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

Resumindo

A margem rígida é aquela que separa claramente os pontos positivos e negativos.

A margem flexível também é chamada SVM linear “barulhenta”, pois inclui alguns pontos classificados incorretamente.

Parâmetro de Regularização C



Distância Mínima Entre os VS:

$$\min_{\mathbf{w},b} \quad \frac{\|\mathbf{w}\|}{2} + C\sum_{i=1}^{N}\xi^{(i)},$$

$$s.t. \quad y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)}, \qquad \forall i \in \{1,\dots,N\}$$

$$\xi^{(i)} \geq 0, \qquad \forall i \in \{1,\dots,N\}$$

Maximizar a Distância Mínima (Otimização):

$$\max_{\alpha} \quad \sum_{i=1}^{N}\alpha^{(i)} - \sum_{i=1}^{N}\sum_{j=1}^{N}\left(y^{(i)}\alpha^{(i)}\phi\left(\mathbf{x}^{(i)}\right)^{T}\phi\left(\mathbf{x}^{(j)}\right)y^{(j)}\alpha^{(j)}\right)$$

$$s.t. \quad 0 \leq \alpha^{(i)} \leq C,$$

Parâmetro de Regularização C

O parâmetro de regularização C no Modelo SVM é responsável pelo treinamento do modelo com hiperplano de margem flexível ou rígida.

Quanto **maior** o valor de C **menor** a margem do hiperplano selecionada para o treinamento de um modelo.

Quanto **menor** o valor de C **maior** a margem do hiperplano escolhida para o treinamento de um modelo.

Para obter resultados de classificação mais precisos (menos amostras classificadas incorretamente), é necessário selecionar C com grande valor.


Continue Trilhando uma Excelente Jornada de Aprendizagem!
Muito Obrigado!