SUPPORT VECTOR MACHINE

Profa Mariana Kleina

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine – SVM (ou Máquina de Vetores de Suporte, em português) é um conjunto de métodos que analisam os dados e reconhecem padrões, usado para classificação (*Support Vector Classification* – SVC) e análise de regressão (*Support Vector Regression* –SVR).

Vamos estudar inicialmente o modelo de classificação, o qual serve de ponto de partida para a construção de modelos mais sofisticados como o da regressão.

SUPPORT VECTOR CLASSIFICATION (SVC)

É EMPREGADO PARA CLASSIFICAÇÃO DE PONTOS EM MÚLTIPLAS CLASSES, ENTRETANTO USAREMOS CLASSIFICAÇÃO EM APENAS DUAS CLASSES.

O SVC PODE TRABALHAR COM:

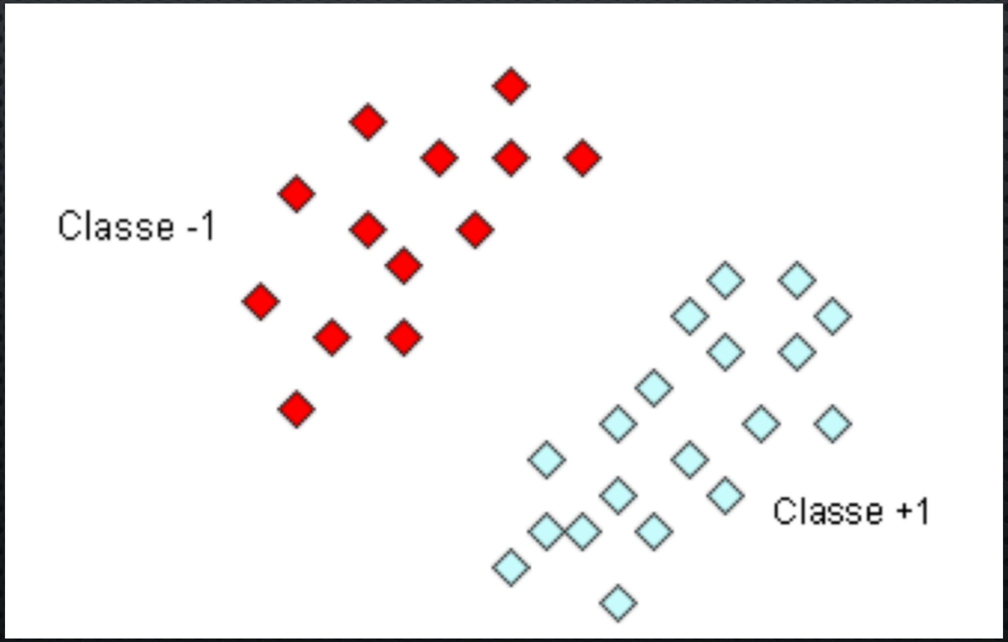
- <u>MARGENS RÍGIDAS</u> QUE CLASSIFICA APENAS DADOS LINEARMENTE SEPARÁVEIS;
- <u>MARGENS FLEXÍVEIS</u> QUE CLASSIFICA DADOS NÃO LINEARMENTE SEPARÁVEIS.

SVC COM MARGENS RÍGIDAS

SVC COM MARGENS RÍGIDAS

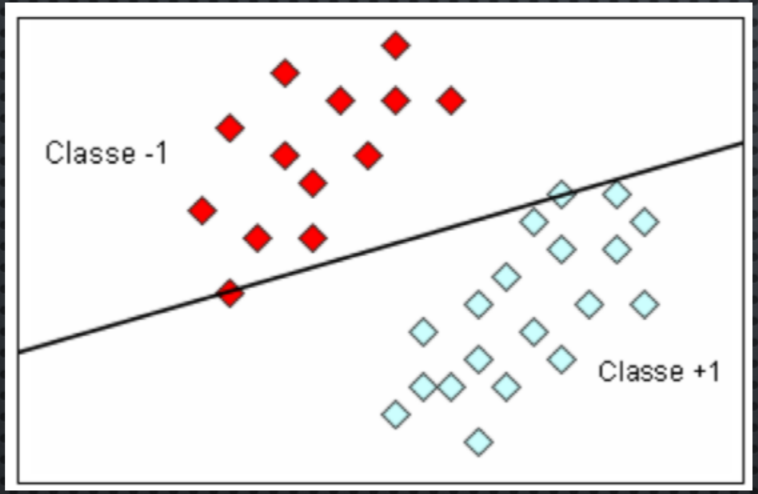
A PARTIR DE UM CONJUNTO DE VETORES DE TREINAMENTO PERTENCENTES A DUAS CLASSES LINEARMENTE SEPARÁVEIS, A TAREFA DO SVC É DETERMINAR UM HIPERPLANO SEPARADOR QUE CLASSIFIQUE CORRETAMENTE OS DADOS DE TESTE.

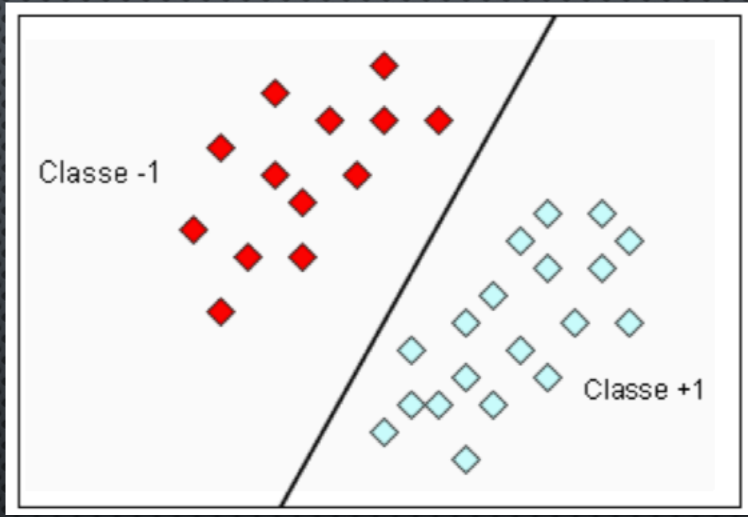
CONSIDERE O CONJUNTO DE DADOS LINEARMENTE SEPARÁVEL:

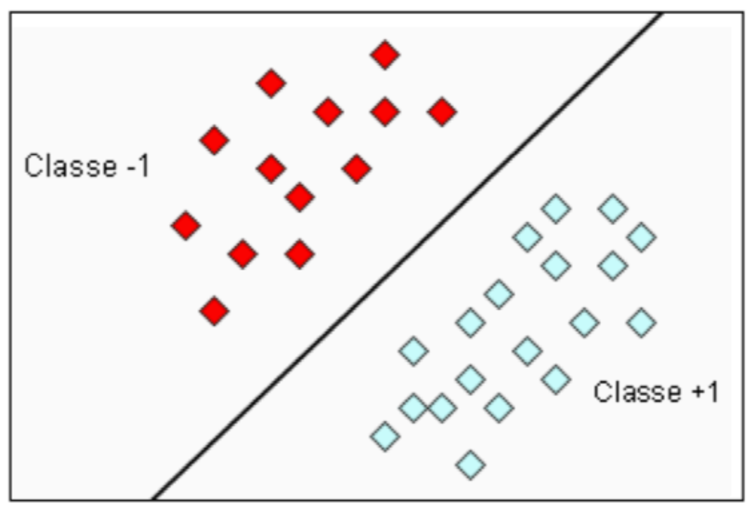


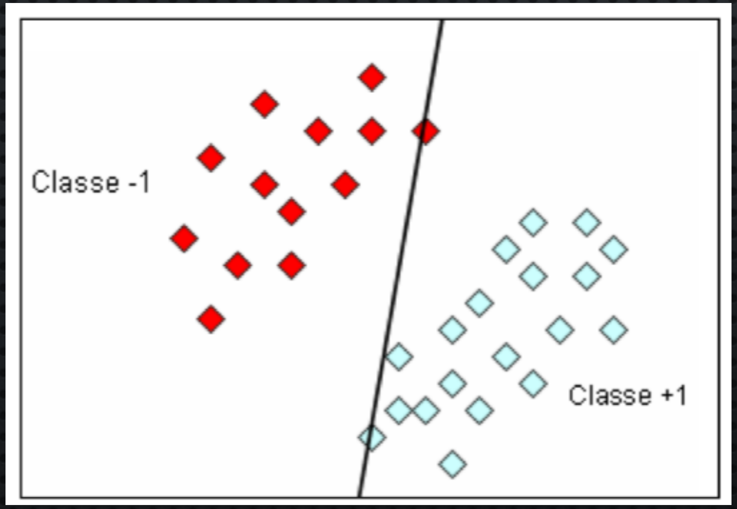
SVC COM MARGENS RÍGIDAS

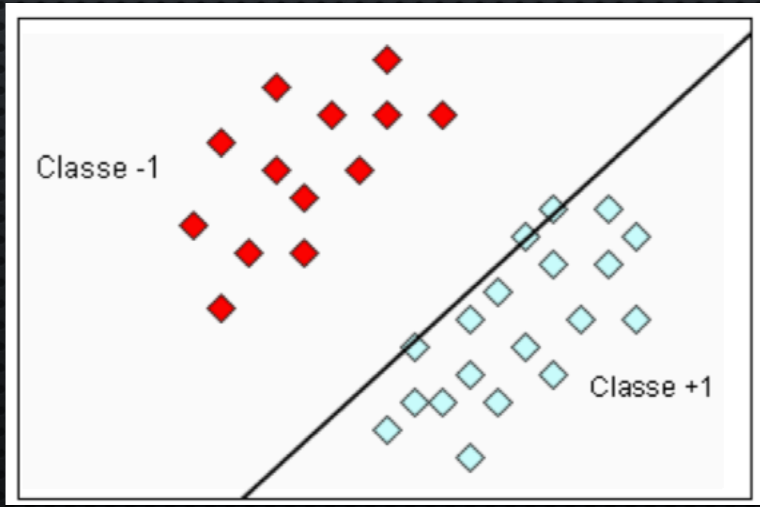
EXISTEM VÁRIAS FORMAS DE SEPARAR A AMOSTRA DE DADOS, SEM ERROS. NO ENTANTO HÁ SOMENTE UMA NA QUAL SE MAXIMIZA A MARGEM DE CLASSIFICAÇÃO.

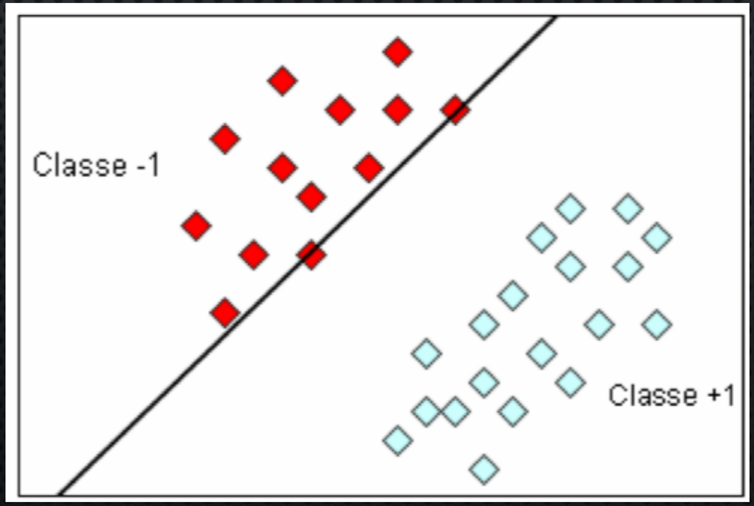






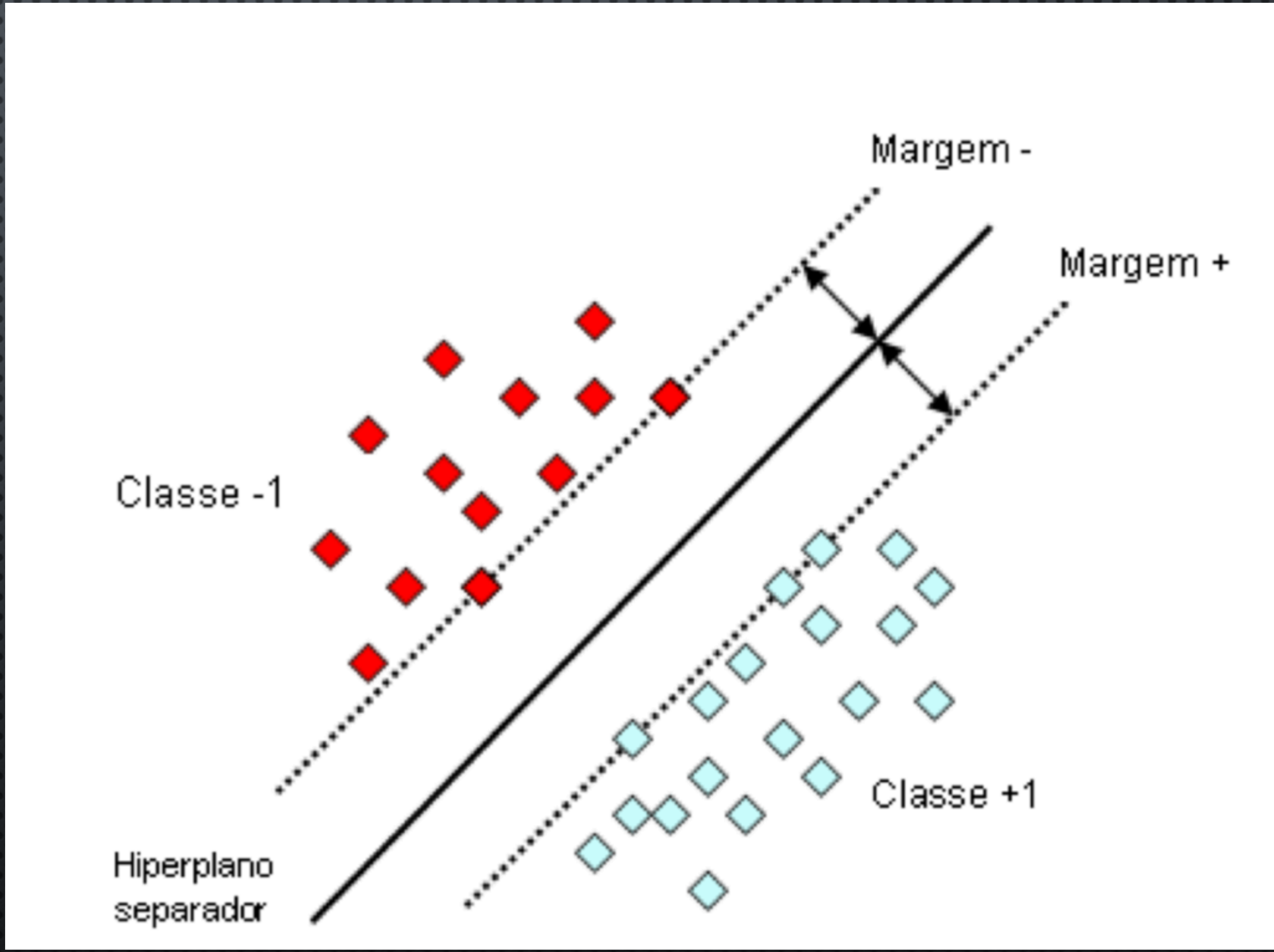






SVC COM MARGENS RÍGIDAS

DEFINE-SE MARGEM DE CLASSIFICAÇÃO COMO A DISTÂNCIA ENTRE O HIPERPLANO SEPARADOR E O VETOR MAIS PRÓXIMO DE CADA CLASSE.



SVC COM MARGENS RÍGIDAS

SEJA T UM CONJUNTO DE TREINAMENTO DADO POR:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \subseteq (X \times Y)^l$$

ONDE l É O NÚMERO DE PONTOS DE TREINAMENTO, $x_i \subseteq \mathbb{R}^n$ SÃO AS ENTRADAS E $y_i \in \{-1, +1\}$ SÃO AS SAÍDAS BINÁRIAS (DUAS CLASSES ROTULADAS EM $\{-1, +1\}$), $i = 1, \dots, l$.

SVC COM MARGENS RÍGIDAS

A CLASSIFICAÇÃO BINÁRIA É REALIZADA POR MEIO DA FUNÇÃO REAL $f: X \subseteq \mathbb{R}^n \to \mathbb{R}$, ATRIBUINDO-SE O VETOR DE ENTRADA $x = (x_1, x_2, \dots, x_n)$ À CLASSE POSITIVA SE $f(x) \geq 0$, E CASO CONTRÁRIO À CLASSE NEGATIVA. TEM-SE QUE A FUNÇÃO DE DECISÃO É DEFINIDA POR:

$$f(x) = \langle w, x \rangle + b$$

ONDE $w \in \mathbb{R}^n$: VETOR DE PESOS

$b \in \mathbb{R}$: *BIAS*

SVC com margens rígidas

A interpretação geométrica do problema consiste na divisão do espaço de entrada X em outros dois subespaços, realizada pelo hiperplano separador $\langle w, x\rangle + b = 0$, cuja dimensão é $n-1$.

Considerando que a margem negativa é definida por $\langle w, x\rangle + b = -1$ e a positiva é definida por $\langle w, x\rangle + b = +1$, um vetor de treinamento está corretamente classificado quando satisfaz as equações:

$$\begin{cases} \langle w, x_i\rangle + b \leq -1, \text{se } y_i = -1 \\ \langle w, x_i\rangle + b \geq +1, \text{se } y_i = +1 \end{cases}$$

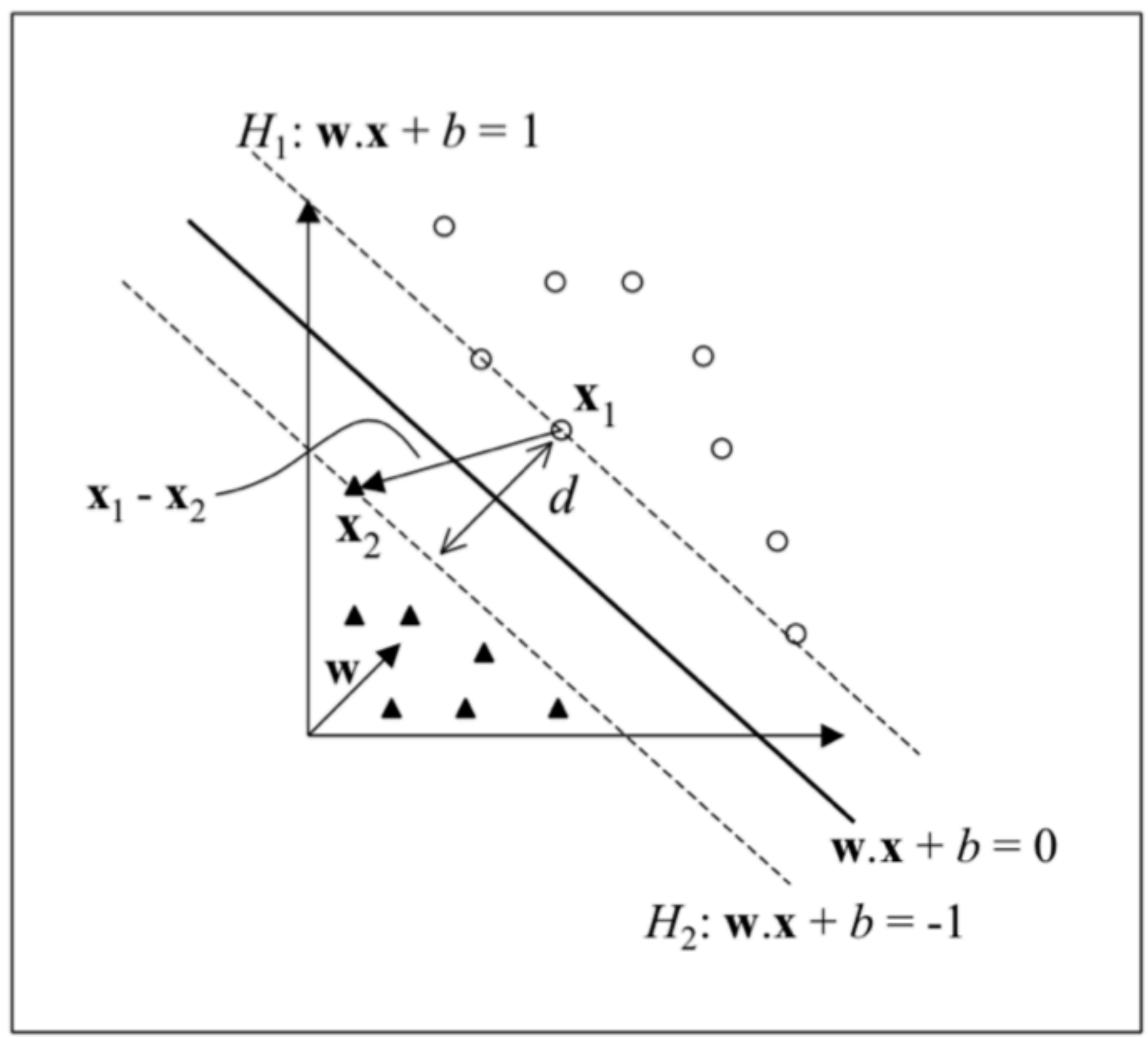
Reescrevendo as duas restrições em uma única expressão:

$$y_i(\langle w, x_i\rangle + b) \geq 1, \qquad \forall i = 1, \dots, l$$

SVC COM MARGENS RÍGIDAS

PODE-SE PROVAR QUE A DISTÂNCIA ENTRE DUAS MARGENS É DADA POR $d = \frac{1}{||w||}$.

QUEREMOS QUE ESSA DISTÂNCIA SEJA A MÁXIMA POSSÍVEL, OU SEJA, $maximizar\ d$, QUE É EQUIVALENTE A $minimizar\ ||w||$, QUE PARA SIMPLIFICAÇÕES POSTERIORES É EQUIVALENTE A $minimizar \frac{1}{2}||w||^2$.



SVC COM MARGENS RÍGIDAS

ASSIM, O PROBLEMA PRIMAL A SER RESOLVIDO É:

$$\begin{aligned} &minimizar \quad \frac{1}{2}||w||^2 \\ &sujeito\ a: \quad y_i(\langle w, x_i \rangle + b) \geq 1 \quad \forall i = 1, \dots, l \end{aligned}$$

ONDE $w \in \mathbb{R}^n$ E $b \in \mathbb{R}$ SÃO AS INCÓGNITAS DO PROBLEMA.

SVC COM MARGENS RÍGIDAS

O PROBLEMA PRIMAL PODE SER DE DIFÍCIL RESOLUÇÃO EM VIRTUDE DAS RESTRIÇÕES DE DESIGUALDADE.

POR ISSO, BUSCA-SE RESOLVER O PROBLEMA DUAL, QUE É FORMULADO A PARTIR DA FUNÇÃO LAGRANGEANA:

$$L(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{l} \alpha_i [y_i(\langle w, x_i\rangle + b) - 1]$$

ONDE $\alpha_i \geq 0$ SÃO OS MULTIPLICADORES DE LAGRANGE.

SVC COM MARGENS RÍGIDAS

A SOLUÇÃO DO PROBLEMA DE OTIMIZAÇÃO EM QUESTÃO É DETERMINADA MINIMIZANDO-SE A FUNÇÃO LAGRANGEANA EM RELAÇÃO ÀS VARIÁVEIS PRIMAIS E MAXIMIZANDO-SE EM RELAÇÃO AOS MULTIPLICADORES DE LAGRANGE, O QUE EQUIVALE A DETERMINAR O PONTO DE SELA DA FUNÇÃO.

PARA MINIMIZAR A FUNÇÃO LAGRANGEANA EM RELAÇÃO ÀS VARIÁVEIS PRIMAIS, CALCULAM-SE AS DERIVADAS PRIMEIRAS DESSA FUNÇÃO EM RELAÇÃO À w E b, E EM SEGUIDA IGUALAM-NAS À ZERO. ASSIM, TEM-SE:

Derivada	**Igualando a zero**
$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{l} y_i \alpha_i x_i$	$w = \sum_{i=1}^{l} y_i \alpha_i x_i$
$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i$	$\sum_{i=1}^{l} y_i \alpha_i = 0$

SVC COM MARGENS RÍGIDAS

SUBSTITUINDO $w = \sum_{i=1}^{l} y_i \alpha_i x_i$ E $\sum_{i=1}^{l} y_i \alpha_i = 0$ NA FUNÇÃO LAGRANGEANA, TEM-SE:

$$L(w, b, \alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

ASSIM, CONSIDERANDO UM CONJUNTO DE TREINAMENTO LINEARMENTE SEPARÁVEL $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$, RESOLVER O PROBLEMA DUAL SIGNIFICA ENCONTRAR OS MULTIPLICADORES DE LAGRANGE α_i, QUE :

$$maximize \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$sujeito\ a\text{:} \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$\alpha_i \geq 0, \qquad \forall i = 1, \dots, l$$

SVC COM MARGENS RÍGIDAS

ENCONTRADOS OS MULTIPLICADORES DE LAGRANGE ÓTIMOS α^*, CALCULA-SE O VETOR DE PESOS ÓTIMO w^*, UTILIZANDO $w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i$.

APESAR DO VALOR DE b NÃO APARECER NO PROBLEMA DUAL, O *BIAS* PODE SER ENCONTRADO POR MEIO DE:

$$b^* = -\frac{1}{2}\langle w^*, (x_r + x_s)\rangle$$

ONDE x_r E x_s SÃO QUAISQUER VETORES SUPORTE DE CADA CLASSE, SATISFAZENDO $\alpha_r, \alpha_s > 0$ E $y_r = -1, y_s = 1$.

SVC COM MARGENS RÍGIDAS

AS CONDIÇÕES DE COMPLEMENTARIDADE DE KKT (KARUSH-KUHN-TUCKER) FORNECEM INFORMAÇÕES MUITO ÚTEIS QUANTO À ESTRUTURA DO PROBLEMA EM QUESTÃO:

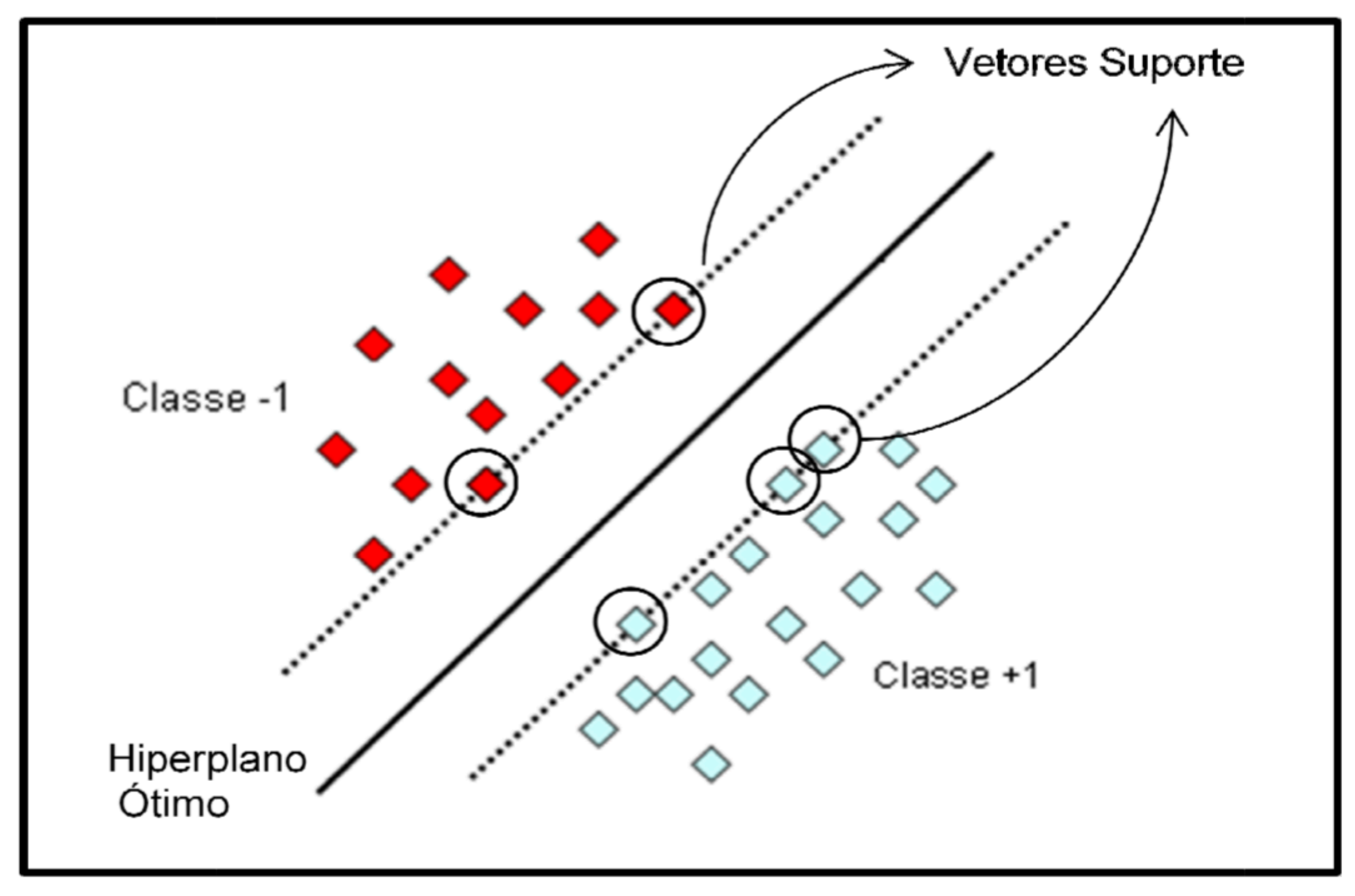
AS SOLUÇÕES ÓTIMAS α^*, w^* E b^* DEVEM SATISFAZER

$$\alpha_i^*[y_i(\langle w^*, x_i\rangle + b^*) - 1] = 0, \qquad \forall i = 1, \ldots, l$$

ISTO É, SOMENTE OS DADOS DE ENTRADA COM MARGEM IGUAL A 1, OS QUAIS ATENDEM À EXPRESSÃO $y_i(\langle w, x_i\rangle + b)$ =1 TEM SEU RESPECTIVO MULTIPLICADOR DE LAGRANGE $\alpha_i \neq 0$. TODOS OS DEMAIS $\alpha_i = 0$.

COMO SOMENTE ESTES PONTOS $(\alpha_i \neq 0)$ ESTÃO ENVOLVIDOS NO CÁLCULO DE w^*, ELES SÃO CHAMADOS DE **VETORES SUPORTE**. ESTES VETORES CONDENSAM TODAS AS INFORMAÇÕES, CONTIDAS NO CONJUNTO DE TREINAMENTO, NECESSÁRIAS PARA CLASSIFICAR OS DADOS DE TESTE.

SVC COM MARGENS RÍGIDAS



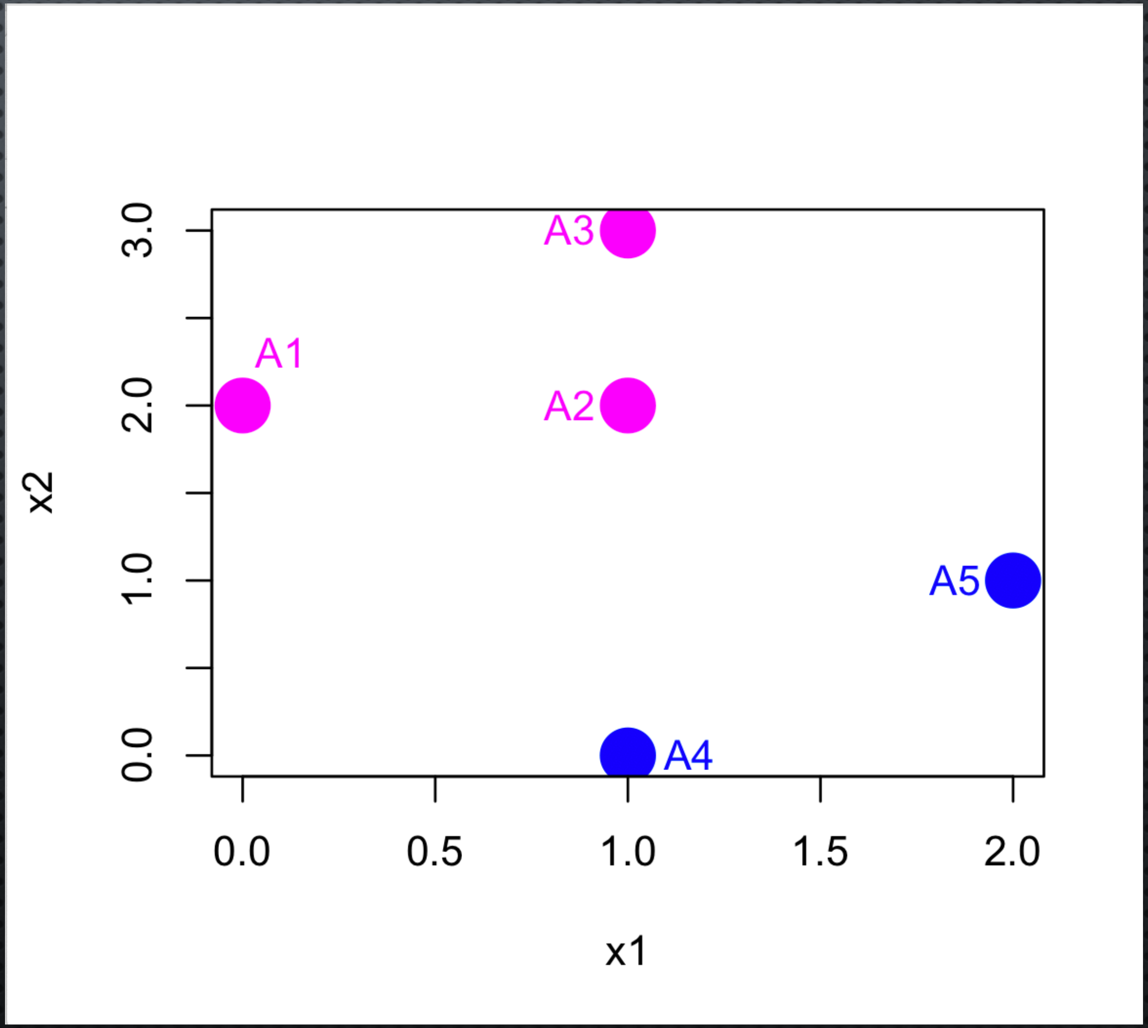
O HIPERPLANO ÓTIMO EXPRESSO NA REPRESENTAÇÃO DUAL É:

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i^* \langle x_i, x \rangle + b^*$$

SVC COM MARGENS RÍGIDAS

EXEMPLO: SEJA O CONJUNTO DE 5 PONTOS LINEARMENTE SEPARÁVEL:

PONTO	CLASSE
A1=(0,2)	-1
A2=(1,2)	-1
A3=(1,3)	-1
A4=(1,0)	+1
A5=(2,1)	+1

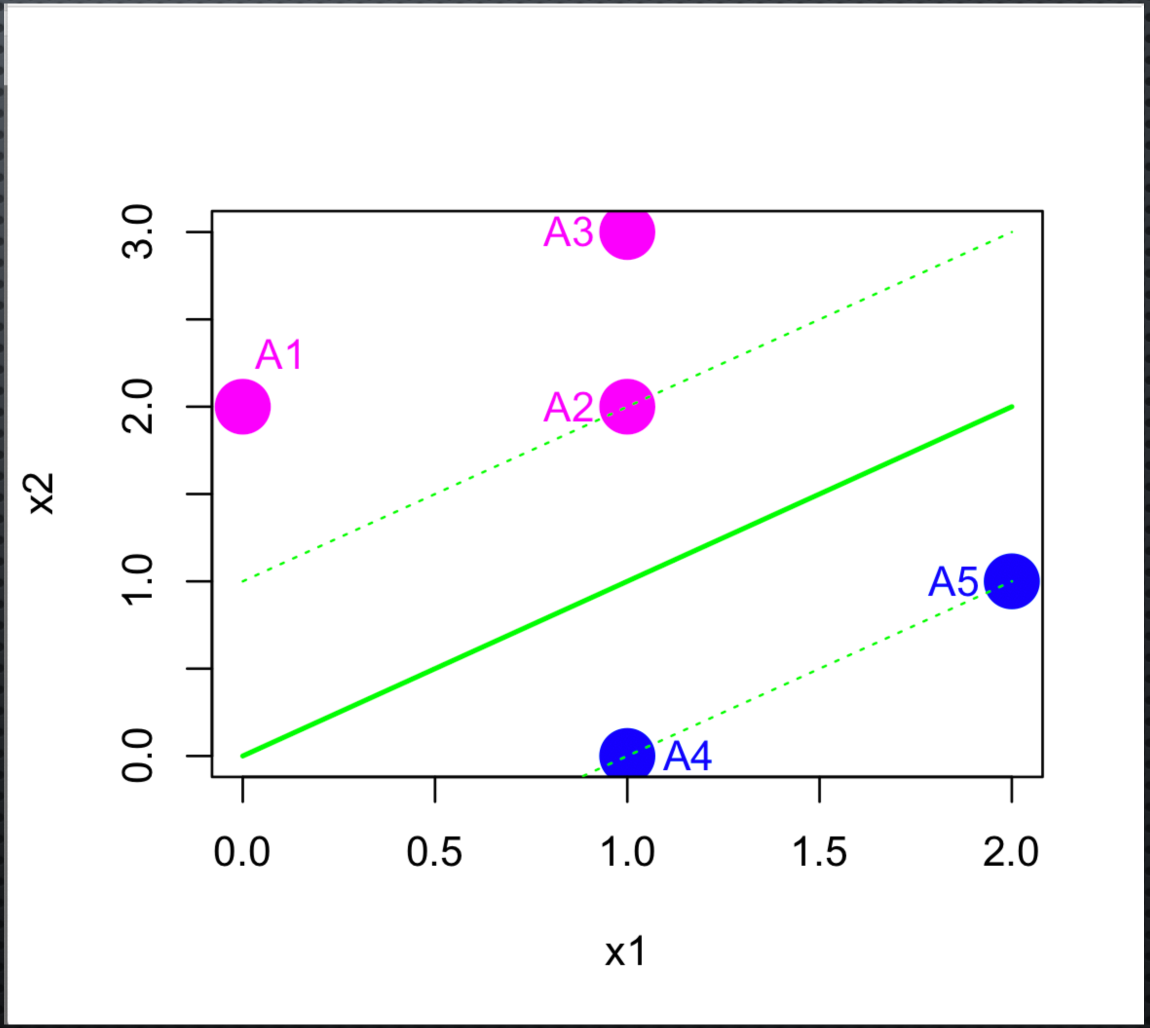


ENCONTRE A MELHOR FORMA DE SEPARAR AS DUAS CLASSES

SVC COM MARGENS RÍGIDAS

EXEMPLO: SEJA O CONJUNTO DE 5 PONTOS LINEARMENTE SEPARÁVEL:

PONTO	CLASSE
A1=(0,2)	-1
A2=(1,2)	-1
A3=(1,3)	-1
A4=(1,0)	+1
A5=(2,1)	+1



ENCONTRE A MELHOR FORMA DE SEPARAR AS DUAS CLASSES

SVC COM MARGENS FLEXÍVEIS

SVC COM MARGENS FLEXÍVEIS

USADO PARA DADOS NÃO LINEARMENTE SEPARÁVEIS

(VIOLAM $y_i(\langle w, x_i \rangle + b) \geq 1, \ \ \forall i = 1, \ldots, l$).

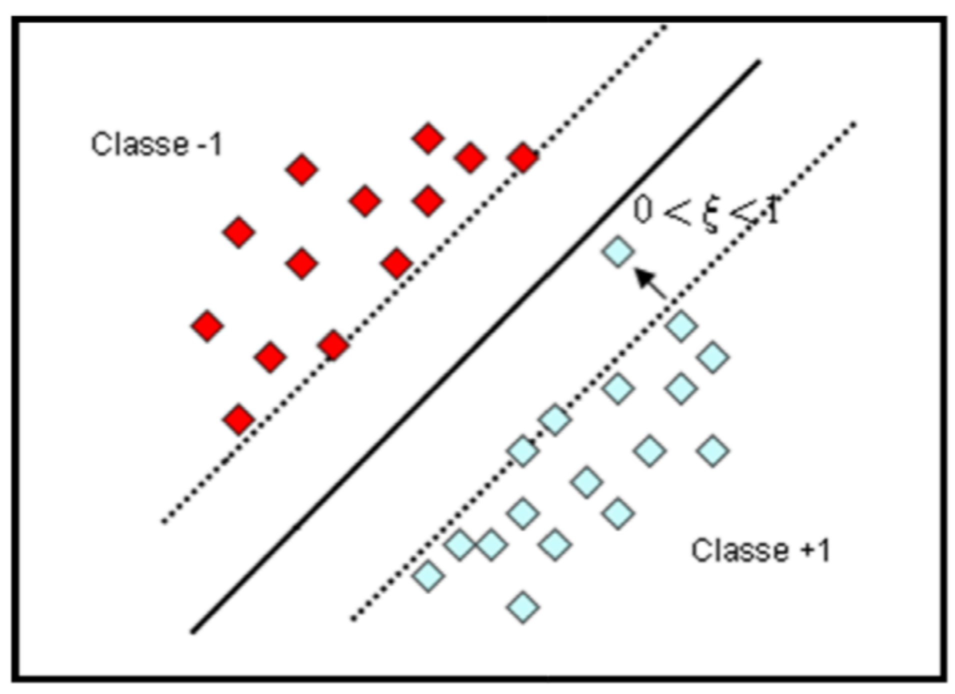
EMBORA NESTES CASOS NÃO SEJA POSSÍVEL CONSTRUIR UM HIPERPLANO SEPARADOR SEM ERROS DE CLASSIFICAÇÃO, É POSSÍVEL ENCONTRAR AQUELE QUE MINIMIZE A PROBABILIDADE DE ERRO JUNTO ÀS AMOSTRAS DE TREINAMENTO.

INTRODUZ-SE VARIÁVEIS DE FOLGA $\xi_i \geq 0$ ASSOCIADAS À CADA VETOR DE TREINAMENTO. UM VETOR ESTÁ CORRETAMENTE SEPARADO SE $\xi_i = 0$.

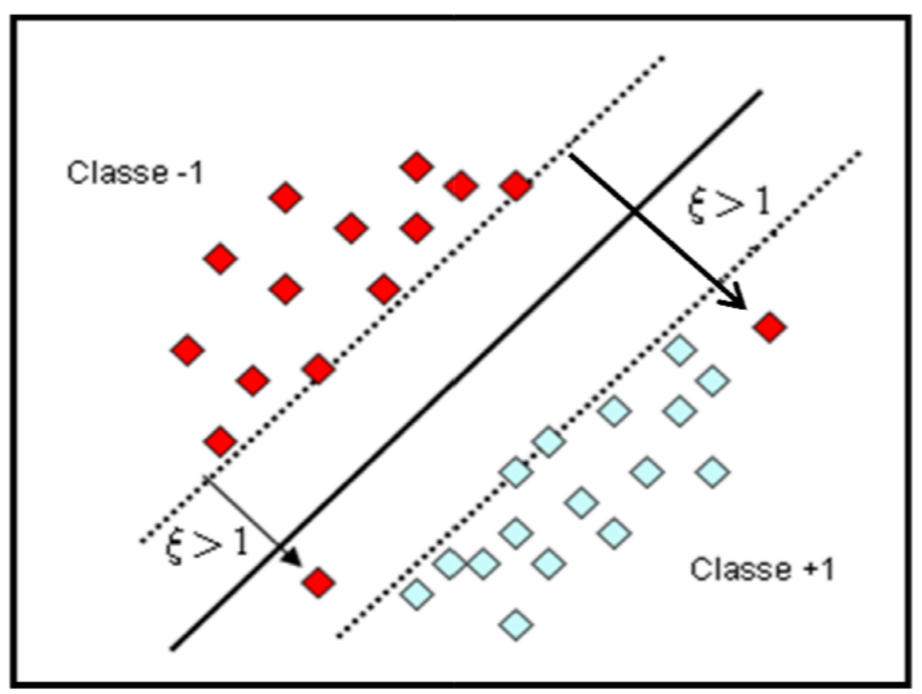
SVC COM MARGENS FLEXÍVEIS

TIPOS DE VIOLAÇÃO:

x_i ESTÁ DENTRO DA REGIÃO DO LADO CORRETO



x_i ESTÁ DENTRO DA REGIÃO DO LADO INCORRETO



SVC COM MARGENS FLEXÍVEIS

AO INSERIR VARIÁVEIS DE FOLGA, DIZ-SE QUE UM VETOR DE TREINAMENTO ESTÁ CORRETAMENTE CLASSIFICADO SE A RELAÇÃO É SATISFEITA:

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \qquad \forall i = 1, \dots, l$$

O PROBLEMA PRIMAL A SER RESOLVIDO É:

$$\begin{aligned} &minimizar && \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i \\ &sujeito\ a: && y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \forall i = 1, \dots, l \\ & && \xi_i \geq 0, \qquad \forall i = 1, \dots, l \end{aligned}$$

ONDE $w \in \mathbb{R}^n$ E $b \in \mathbb{R}$ SÃO AS INCÓGNITAS DO PROBLEMA E C É A CONSTANTE DE REGULARIZAÇÃO, POIS PONDERA OS TERMOS DA FUNÇÃO DE MINIMIZAÇÃO.

SVC COM MARGENS FLEXÍVEIS

O PROBLEMA DUAL É:

$$maximizar \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$sujeito\ a: \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad \forall i = 1, \ldots, l$$

SVC COM MARGENS FLEXÍVEIS

As condições de KKT relacionadas ao problema dual são definidas por:

$$\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] = 0, \qquad \forall i = 1, \dots, l$$

$$\xi_i(\alpha_i - C) = 0, \qquad \forall i = 1, \dots, l$$

Os pontos de entrada para os quais $\alpha_i > 0$ recebem o nome de vetores suporte.

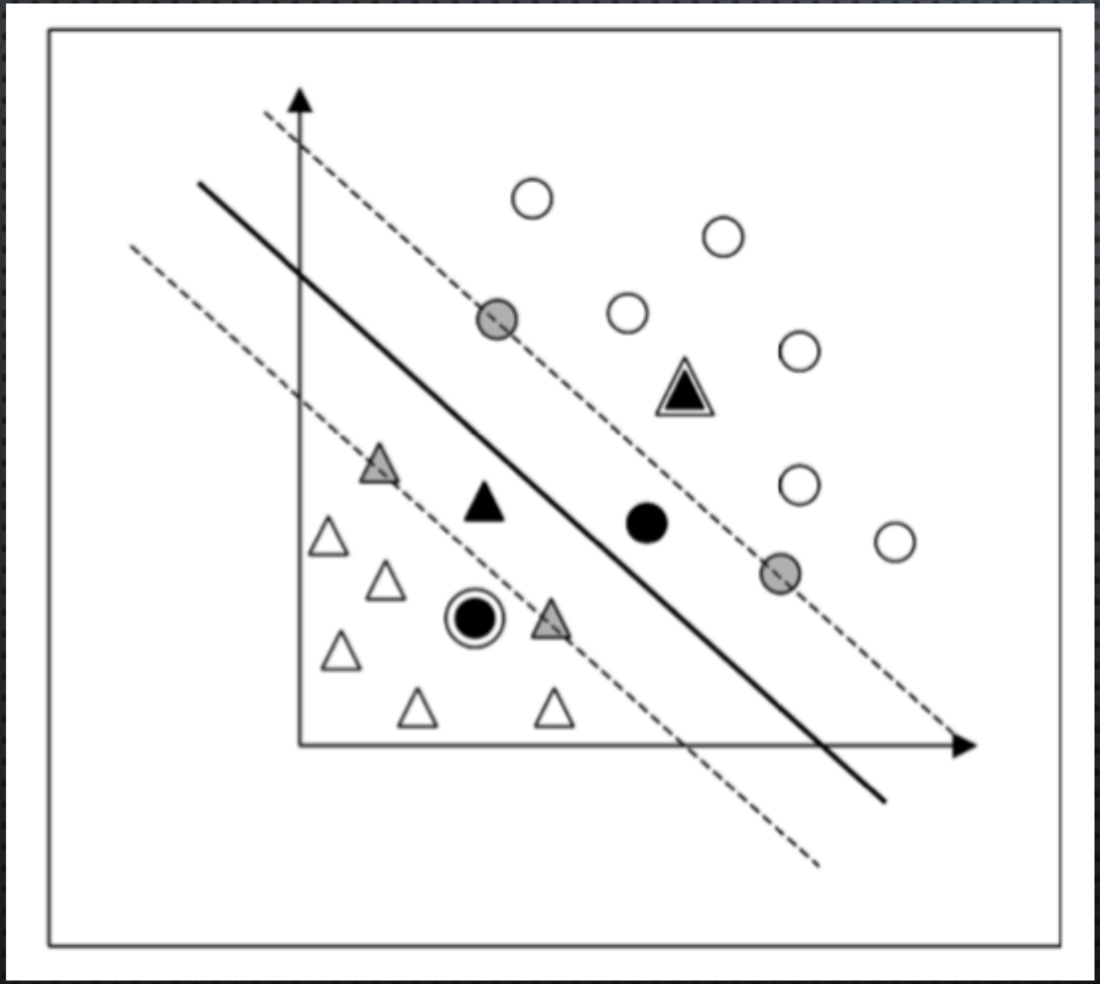
Se $0 < \alpha_i < C$ (chamados de vetores suportes não limitados), então por $\xi_i(\alpha_i - C) = 0$ tem se que $\xi_i = 0$ e estes estão sobre a margem de sua classe.

SVC COM MARGENS FLEXÍVEIS

SE $\alpha_i = C$ (CHAMADOS DE VETORES SUPORTE LIMITADOS), ENTÃO É POSSÍVEL QUE $\xi_i > 1$ (ERROS DE CLASSIFICAÇÃO) OU $0 < \xi_i \leq 1$ (CORRETAMENTE CLASSIFICADOS, PORÉM ENTRE AS MARGENS).

COM EXCEÇÃO DOS VETORES SUPORTE, SEJAM ELES LIMITADOS OU NÃO, TEM-SE QUE OS DEMAIS PONTOS (COM $\alpha_i = 0$) NÃO INFLUENCIAM NA CONSTRUÇÃO DO HIPERPLANO SEPARADOR ÓTIMO. AS VARIÁVEIS w^*, b^* E A SUPERFÍCIE DE DECISÃO SÃO DETERMINADOS COMO NO CASO DE MARGENS RÍGIDAS.

SVC COM MARGENS FLEXÍVEIS

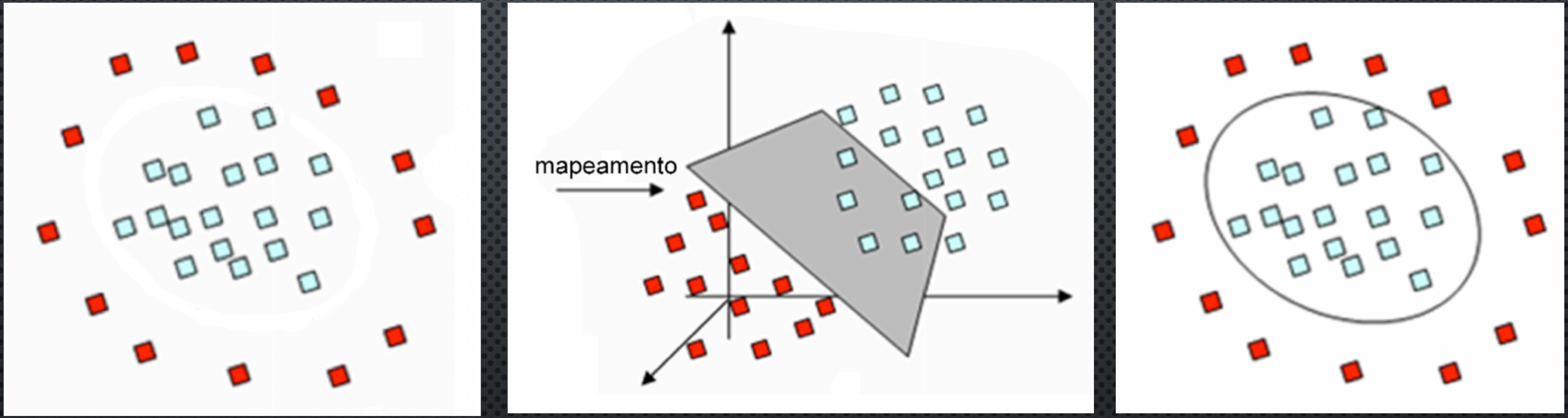


- VS NÃO LIMITADOS: PONTOS EM CINZA
- VS LIMITADOS: PONTOS EM PRETO E PONTOS EM PRETO COM BORDA (ERROS DE CLASSIFICAÇÃO)

SVC NÃO LINEAR

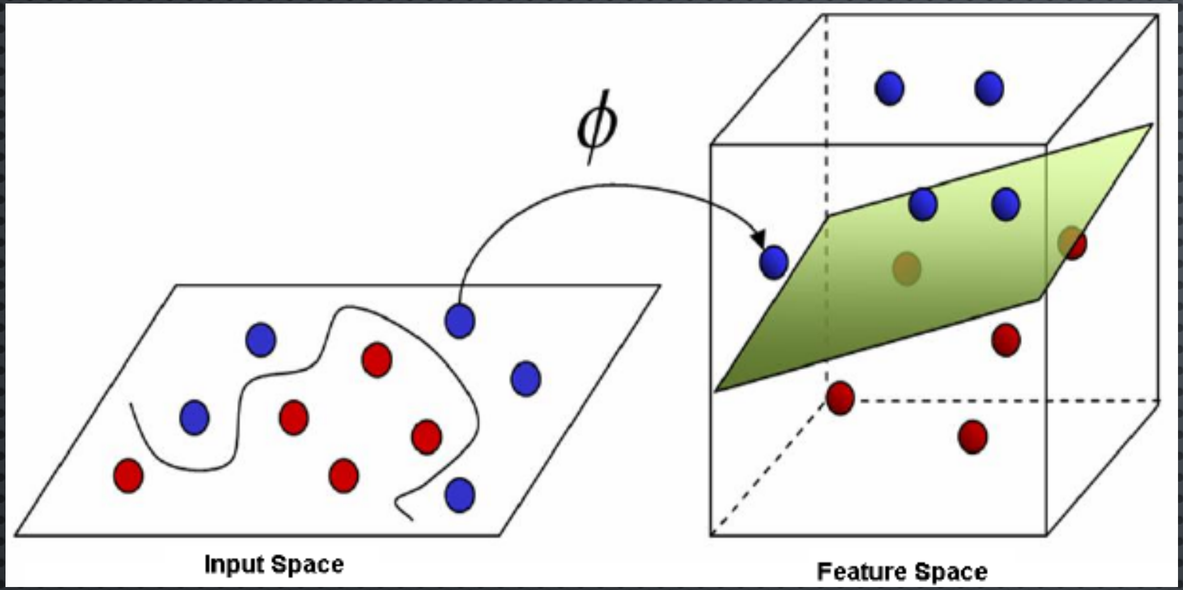
SVC NÃO LINEAR

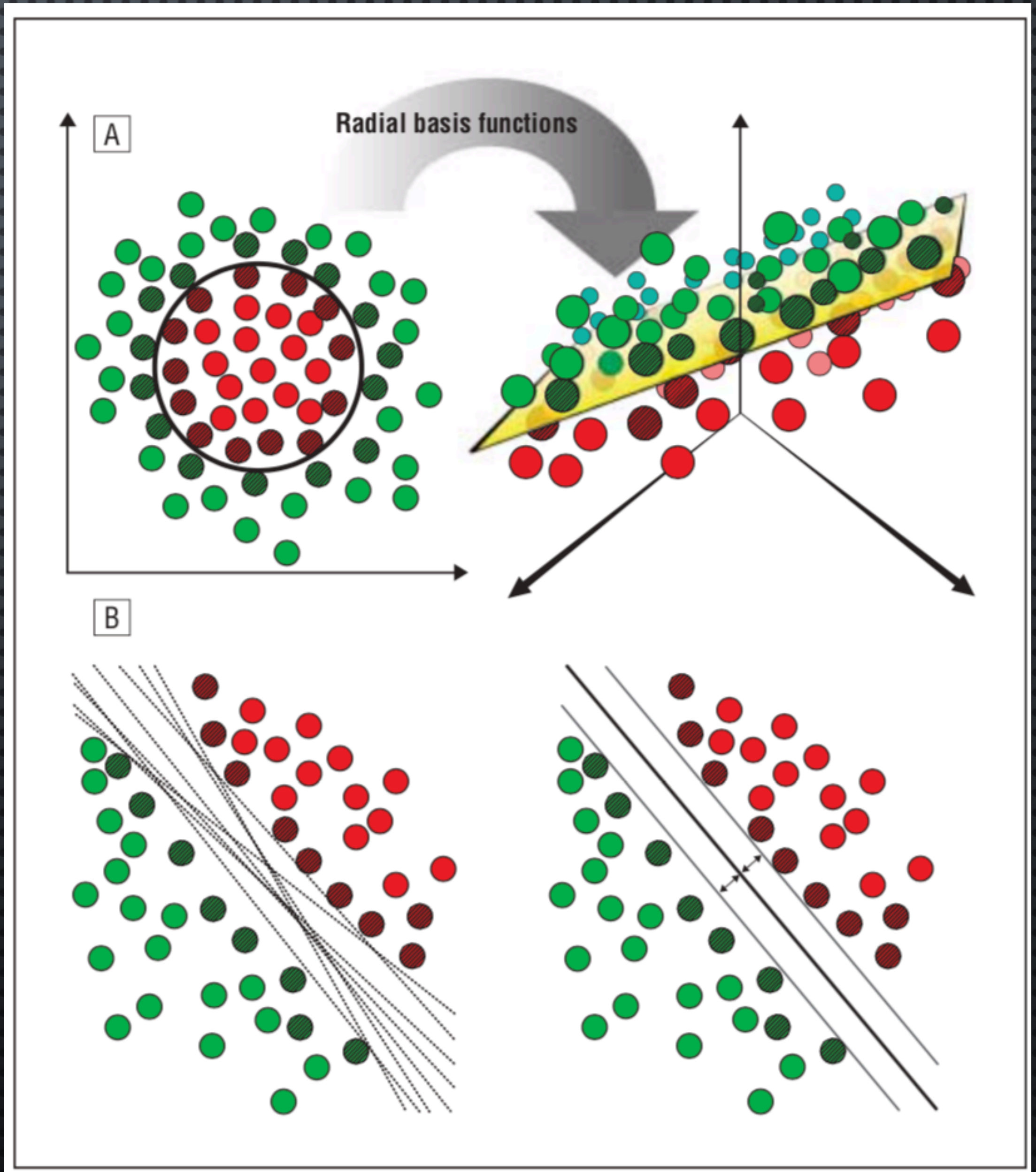
MAS E SE OS DADOS NÃO FOREM LINEARMENTE SEPARÁVEIS?

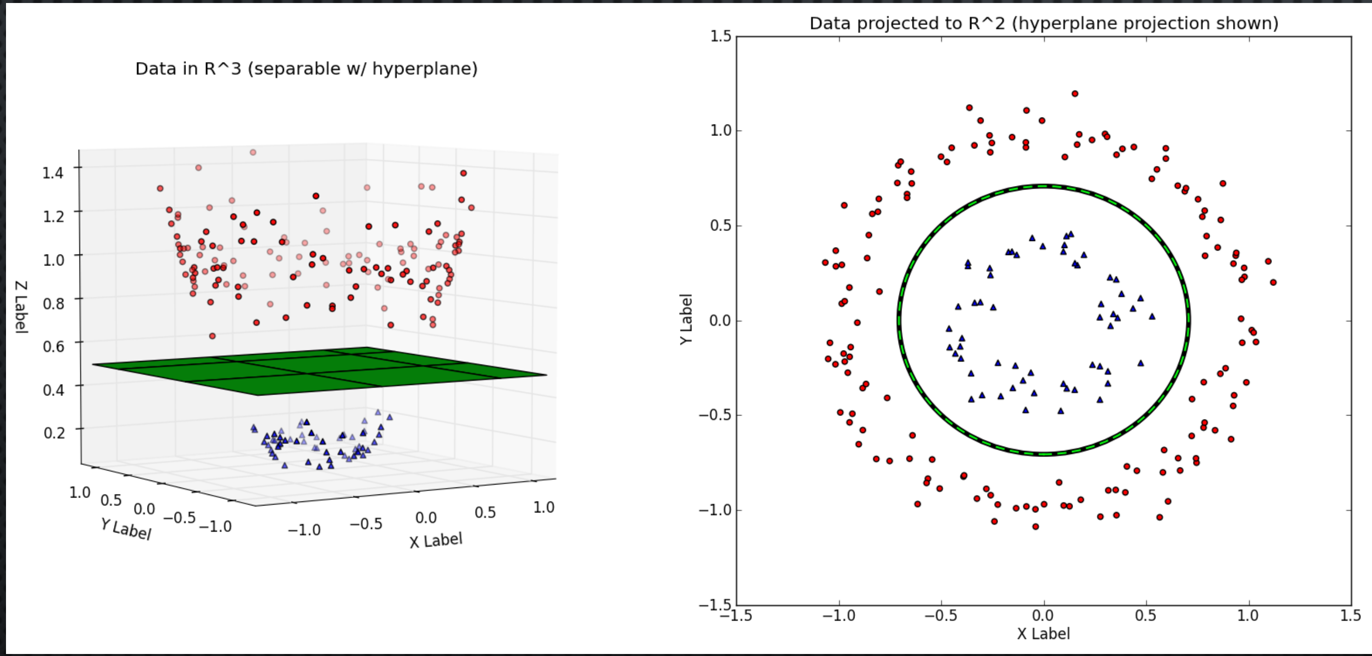


AUMENTA-SE A DIMENSÃO DO ESPAÇO ONDE A SEPARAÇÃO LINEAR TORNA-SE POSSÍVEL.

SVC NÃO LINEAR







SVC NÃO LINEAR

PARA ISSO, REALIZA-SE UM MAPEAMENTO NÃO LINEAR DOS DADOS DE ENTRADA X, PARA UM ESPAÇO DE ALTA DIMENSIONALIDADE F, CHAMADO DE ESPAÇO DAS CARACTERÍSTICAS. UMA TÉCNICA COMUM PARA REALIZAR ESTE PROCEDIMENTO É MUDAR A REPRESENTAÇÃO DOS DADOS POR MEIO DE:

$$x = (x_1, x_2, \dots, x_n) \longmapsto \phi(x) = (\phi_1(x), \dots, \phi_N(x))$$

ONDE $F = \{\phi\,(x) \mid x \in X\}$.

SVC NÃO LINEAR

NA PRÁTICA, BASTA SUBSTITUIR NA FUNÇÃO OBJETIVO DO PROBLEMA AS ENTRADAS x POR $\phi(x)$, ONDE $\phi\colon \mathbb{R}^n \rightarrow \mathbb{R}^N$, COM $N \gg n$.

$$\textit{maximizar} \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\textit{sujeito a:} \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad \forall i = 1, \dots, l$$

SVC NÃO LINEAR

NA PRÁTICA, BASTA SUBSTITUIR NA FUNÇÃO OBJETIVO DO PROBLEMA AS ENTRADAS x POR $\phi(x)$, ONDE $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^N$, COM $N \gg n$.

$$maximizar \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$sujeito\ a: \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad \forall i = 1, \dots, l$$

SVC NÃO LINEAR

Contudo, realizar o produto interno $\langle \phi(x_i), \phi(x_j) \rangle$ diretamente no espaço das características pode se tornar computacionalmente inviável, devido à sua alta dimensionalidade.

A fim de se evitar o custo computacional, o mapeamento pode ser realizado implicitamente pelas chamadas **<u>funções kernel</u>**, as quais dependem apenas das variáveis de entrada.

Definição: Um kernel é uma função K, tal que para todo $x, z \in X$ tem-se

$$K(x,z) = \langle \phi(x), \phi(z) \rangle$$

onde ϕ é um mapeamento de X para (um produto interno) o espaço característico F.

SVC NÃO LINEAR

Tipos de funções Kernel:

- Kernel Linear: $K(x_i, x_j) = \langle x_i, x_j \rangle$
- Kernel Polinomial Homogêneo de grau p: $K(x_i, x_j) = (\langle x_i, x_j \rangle)^p$
- Kernel Polinomial não Homogêneo de grau p:
- $K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^p$ onde c é uma constante
- Kernel Sigmoidal: $K(x_i, x_j) = tanh(\langle bx_i, x_j \rangle + c)$ onde b é um coeficiente e c é uma constante negativa
- Kernel Gaussiano (ou Função de Base Radial): $K(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}}$ onde

σ é um parâmetro

SVC NÃO LINEAR

ASSIM, O PROBLEMA SE TORNA:

$$maximizar \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

$$sujeito\ a: \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad \forall i = 1, \dots, l$$

ONDE α_i SÃO OS MULTIPLICADORES DE LAGRANGE.

SVC NÃO LINEAR

AS CONDIÇÕES DE KKT SÃO AS MESMAS DO CASO LINEARMENTE SEPARÁVEL, ENTRETANTO O VETOR DE PESOS JÁ NÃO PODE SER MAIS DESCRITO EXPLICITAMENTE, POIS SEU CÁLCULO ENVOLVE A FUNÇÃO $\phi(x)$ QUE NEM SEMPRE É CONHECIDA.

$$w^* = \sum_{i=1}^{l} y_i \alpha_i^* \phi(x_i)$$

O *BIAS* É ENCONTRADO POR

$$b^* = -\frac{1}{2}\sum_{i=1}^{l} \alpha_i y_i [K(x_i, x_r) + K(x_i, x_s)]$$

E O HIPERPLANO SEPARADOR ÓTIMO, NO ESPAÇO CARACTERÍSTICO, É

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^*$$

SVC NÃO LINEAR

VALORES DIFERENTES PARA A CONSTANTE DE REGULARIZAÇÃO C E DIFERENTES FUNÇÕES KERNEL (E SEUS PARÂMETROS) VÃO CONDUZIR A RESULTADOS DIFERENTES.

O MÉTODO PODE SOFRER *OVERFITTING* – ESPECIALIZAÇÃO NOS DADOS DE TREINAMENTO E RESULTADOS RUINS NO CONJUNTO DE TESTE.

SVC NÃO LINEAR

EXEMPLO: SEJA O CONJUNTO DE 5 PONTOS:

$x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 5$ E $x_5 = 6,$

QUE DEVEM SER SEPARADOS EM 2 CLASSES:

$y_1 = 1, y_2 = 1, y_3 = -1, y_4 = -1$ E $y_5 = 1.$

USE A FUNÇÃO KERNEL $K(a, b) = (1 + a^T b)^2$ PARA SEPARAR OS DADOS.

SUPPORT VECTOR REGRESSION (SVR)

SVR

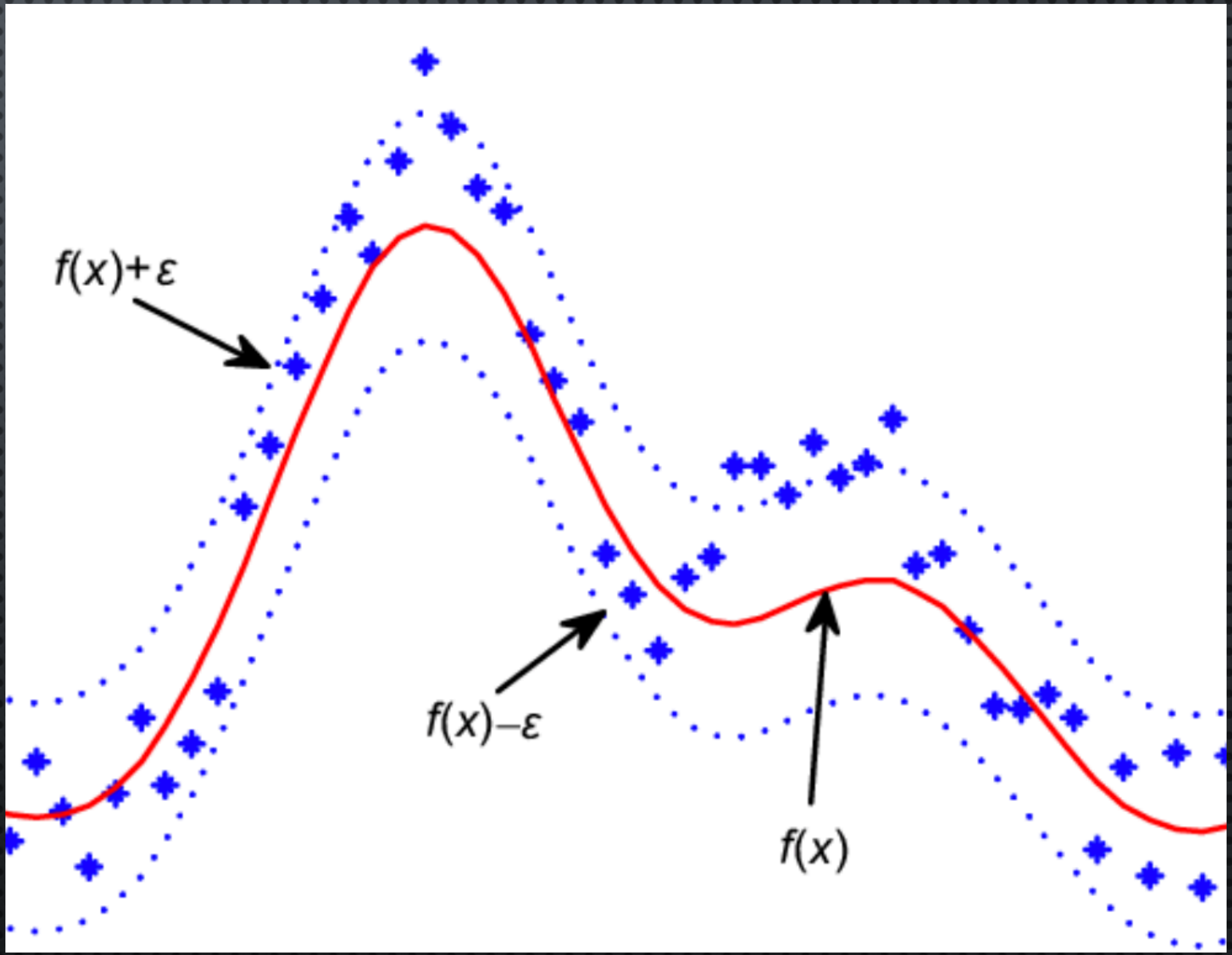
Seja T um conjunto de treinamento dado por:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \subseteq (X \times Y)^l$$

onde l é o número de pontos de treinamento, $x_i \subseteq \mathbb{R}^n$ são as entradas e $y_i \subseteq \mathbb{R}$ são as saídas desejadas (não mais binárias como no SVC), $i = 1, \dots, l$.

SVR

O OBJETIVO DO SVR É ENCONTRAR UMA FUNÇÃO $f(x) = \langle w, x \rangle + b$ QUE APRESENTE NO MÁXIMO UM DESVIO ε EM RELAÇÃO AOS VALORES ALVO y_i, OU SEJA, PROCURA-SE UMA FUNÇÃO COM UMA MARGEM DE ERROS CARACTERIZADA PELO INTERVALO $[y_i - \varepsilon, y_i + \varepsilon]$. ASSIM, DESVIOS SÃO PERMITIDOS DESDE QUE NÃO ULTRAPASSEM A MARGEM ESPECIFICADA.



SVR

Visto que a função $f(x) = \langle w, x \rangle + b$ deve satisfazer as restrições de erro $|f(x_i) - y_i| \leq \varepsilon$, $\forall i = 1, \dots, l,$ modela-se o problema de otimização primal como:

$$
\begin{aligned}
&minimize && \frac{1}{2}\|w\|^2 \\
&sujeito\ a: && y_i - \langle w, x_i \rangle - b \leq \varepsilon \quad \forall = 1, \dots, l \\
& && \langle w, x_i \rangle + b - y_i \leq \varepsilon \quad \forall = 1, \dots, l
\end{aligned}
$$

onde $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$ são as incógnitas do problema.

SVR

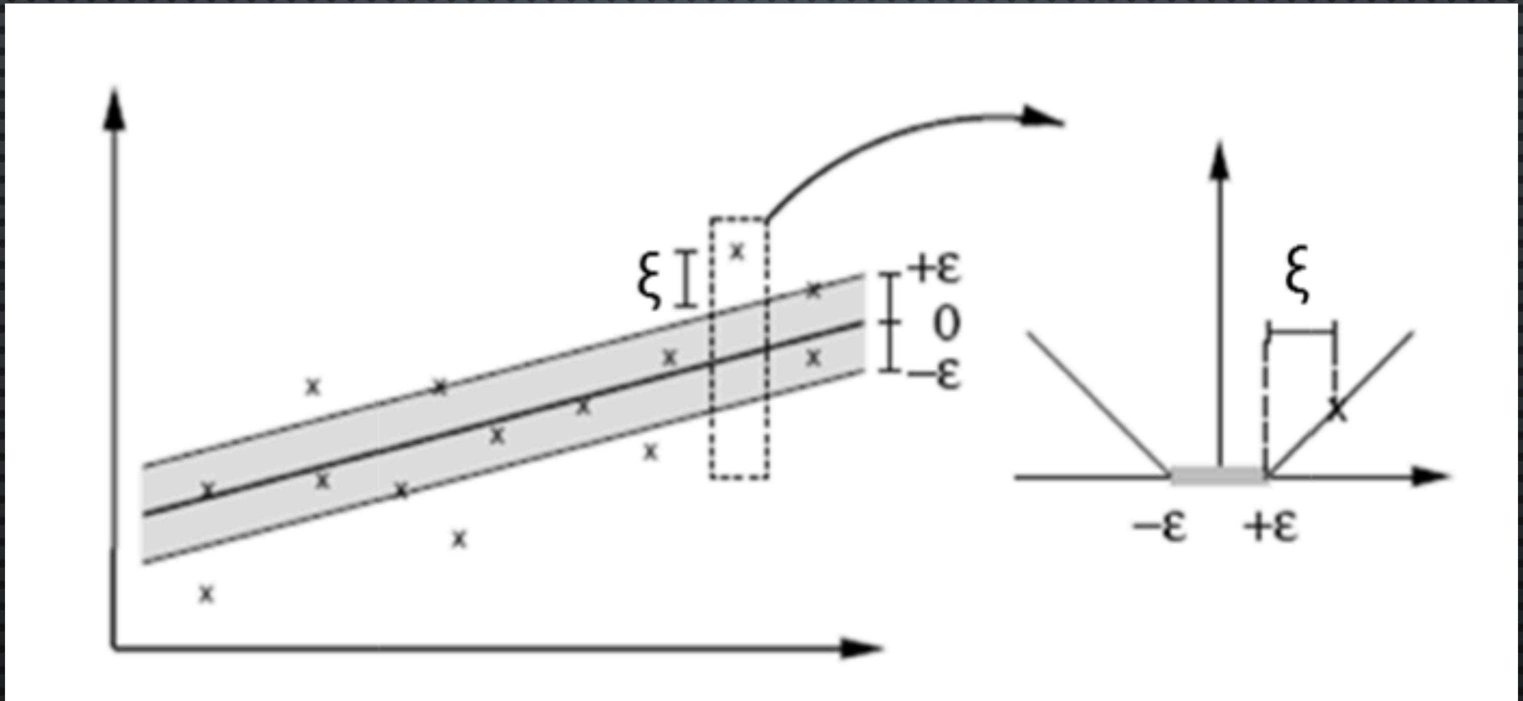
Nem sempre é possível garantir a viabilidade do problema, visto que existem pontos que violam suas restrições. Assim, estuda-se uma função de perda, que introduz variáveis de folga não negativas ξ e $\dot{\xi}$, cuja finalidade é penalizar dados que se situem fora da margem $|f(x_i) - y_i| \leq \varepsilon$.

A função perda, denominada $\varepsilon - Insensitive$, é descrita como:

$$|\xi|_{\varepsilon} = \begin{cases} 0 & se|\xi| \leq \varepsilon \\ |\xi| - \varepsilon & caso\ contrario \end{cases}$$

SVR

As variáveis de folga ξ_i e $\dot{\xi}_i$ estão associadas, respectivamente, aos dados localizados abaixo da margem inferior e acima da margem superior.



Função de perda $\varepsilon - Insensitive$

SVR

O PROBLEMA PRIMAL PODE SER ESCRITO COMO:

$$
\begin{aligned}
&minimize \qquad && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \dot{\xi}) \\
& && y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\
&sujeito\ a: && \langle w, x_i \rangle + b - y_i \leq \varepsilon + \dot{\xi}_i \\
& && \xi_i, \dot{\xi}_i \geq 0, \qquad \forall i = 1, \dots, l
\end{aligned}
$$

ONDE $w \in \mathbb{R}^n$ E $b \in \mathbb{R}$ SÃO AS INCÓGNITAS DO PROBLEMA E C É DENOMINADA CONSTANTE DE REGULARIZAÇÃO (MEDE O GRAU DE IMPORTÂNCIA DADO ÀS VARIÁVEIS DE FOLGA, CARACTERIZANDO A QUANTIDADE DE ERROS QUE SERÃO PERMITIDOS – DESVIOS ACIMA DE ε).

SVR

NOVAMENTE RECORRENDO AO PROBLEMA DUAL E UTILIZANDO A TEORIA LAGRANGEANA, TEM-SE:

$$maximize \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \dot{\alpha}_i)(\alpha_j - \dot{\alpha}_j)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \dot{\alpha}_i) + \sum_{i=1}^{l} y_i(\alpha_i - \dot{\alpha}_i)$$

$$sujeito\ a: \quad \sum_{i=1}^{l}(\alpha_i - \dot{\alpha}_i) = 0$$

$$\alpha_i, \dot{\alpha}_i \in [0, C], \qquad \forall i = 1, \dots, l$$

ONDE α E $\dot{\alpha}$ SÃO OS MULTIPLICADORES DE LAGRANGE.

SVR

As condições de KKT são:

$$\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) = 0$$

$$\dot{\alpha}_i\left(\varepsilon + \dot{\xi}_i + y_i - \langle w, x_i \rangle - b\right) = 0$$

$$(C - \alpha_i)\xi = 0$$

$$(C - \alpha_i)\dot{\xi}_i = 0$$

$$\alpha_i \dot{\alpha}_i = 0$$

$$\xi_i \dot{\xi}_i = 0$$

para todo $i = 1, \dots, l$.

SVR

Observações:

- Por $\alpha_i \dot{\alpha}_i = 0$, nota-se que nunca existirá um conjunto de variáveis duais α_i e $\dot{\alpha}_i$ em que ambos são nulos;
- Das duas primeiras equações, observa-se que apenas os pontos em que $|f(x_i) - y_i| \geq \varepsilon$ estão associados a multiplicadores de Lagrange $\neq 0$. Tais pontos estão localizados sobre as margens $+\varepsilon$ e $-\varepsilon$ ou fora da região delimitada pelas margens. Estes dados são os únicos a serem utilizados no cálculo do vetor de pesos w e por isso são chamados de vetores suporte;
- Assim com no SVC com margens flexíveis, os vetores suporte podem ser limitados ou não. Aqueles que são limitados, cujo $\alpha_i = C$ ou $\dot{\alpha}_i = C$, tem respectivamente variáveis de folga ξ ou $\dot{\xi}$ não nulas (pois $(C - \alpha_i)\xi = 0$ e $(C - \alpha_i)\dot{\xi}_i = 0$). Ou seja, são pontos fora das margens e correspondem a erros do modelo.

SVR

O vetor de pesos é calculado por:

$$w^* = \sum_{i=1}^{l} (\alpha_i - \dot{\alpha}_i) x_i$$

O *bias* é calculado por:

$$b^* = -\frac{1}{2} \langle w^*, (x_r + x_s) \rangle$$

onde x_r e x_s são quaisquer vetores suporte de cada classe, satisfazendo $\alpha_r, \alpha_s > 0$.

E a função de decisão é:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \dot{\alpha}_i) \langle x_i, x \rangle + b$$

SVR

Até agora, foram consideradas regressões lineares no espaço de entrada. Representar $f(x)$ por uma função linear é conveniente, porém em algumas aplicações práticas as relações lineares tornam-se simplistas demais.

Assim, a formulação dual do problema SVR permite trabalhar com um espaço de alta dimensionalidade denominado F (espaço das características). Com isso, é realizado um mapeamento não linear dos dados de entrada para um espaço de dimensão maior, onde a regressão linear torna-se possível.

SVR

NOVAMENTE UTILIZA-SE A ABORDAGEM BASEADA EM FUNÇÕES KERNEL. SUBSTITUINDO $\langle x_i, x_j \rangle$ POR $K(x_i, x_j)$ (SIGNIFICANDO QUE A REGRESSÃO LINEAR É REALIZADA NO ESPAÇO DAS CARACTERÍSTICAS), TEM-SE:

$$maximize \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \dot{\alpha}_i)(\alpha_j - \dot{\alpha}_j)K(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \dot{\alpha}_i) + \sum_{i=1}^{l} y_i(\alpha_i - \dot{\alpha}_i)$$

$$sujeito\ a\text{:} \quad \sum_{i=1}^{l}(\alpha_i - \dot{\alpha}_i) = 0$$

$$\alpha_i, \dot{\alpha}_i \in [0, C], \qquad \forall i = 1, \dots, l$$

ONDE α E $\dot{\alpha}$ SÃO OS MULTIPLICADORES DE LAGRANGE.

SVR

O vetor de pesos passa a não ser mais descrito explicitamente, pois envolve a função $\phi(x)$ em seu cálculo:

$$w^* = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\phi(x_i)$$

No entanto a função de decisão é descrita por:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)K(x_i, x) + b$$

onde o *bias* é determinado por:

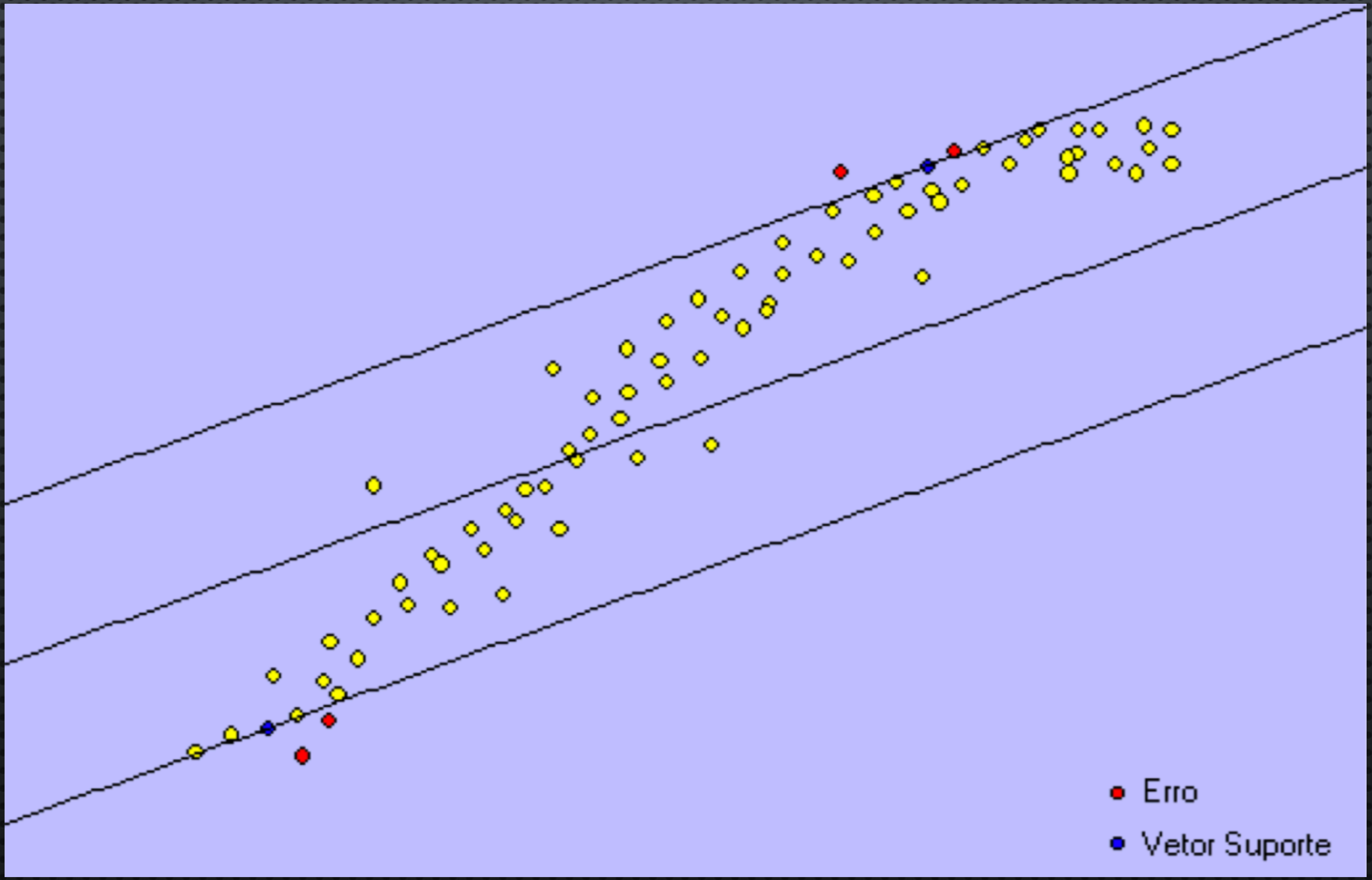
$$b^* = -\frac{1}{2}\sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(K(x_i, x_r) + K(x_i, x_s))$$

SVR

Os resultados do SVR dependem significativamente dos valores da constante de regularização C, margem ε, do tipo de função kernel e seus respectivos parâmetros.

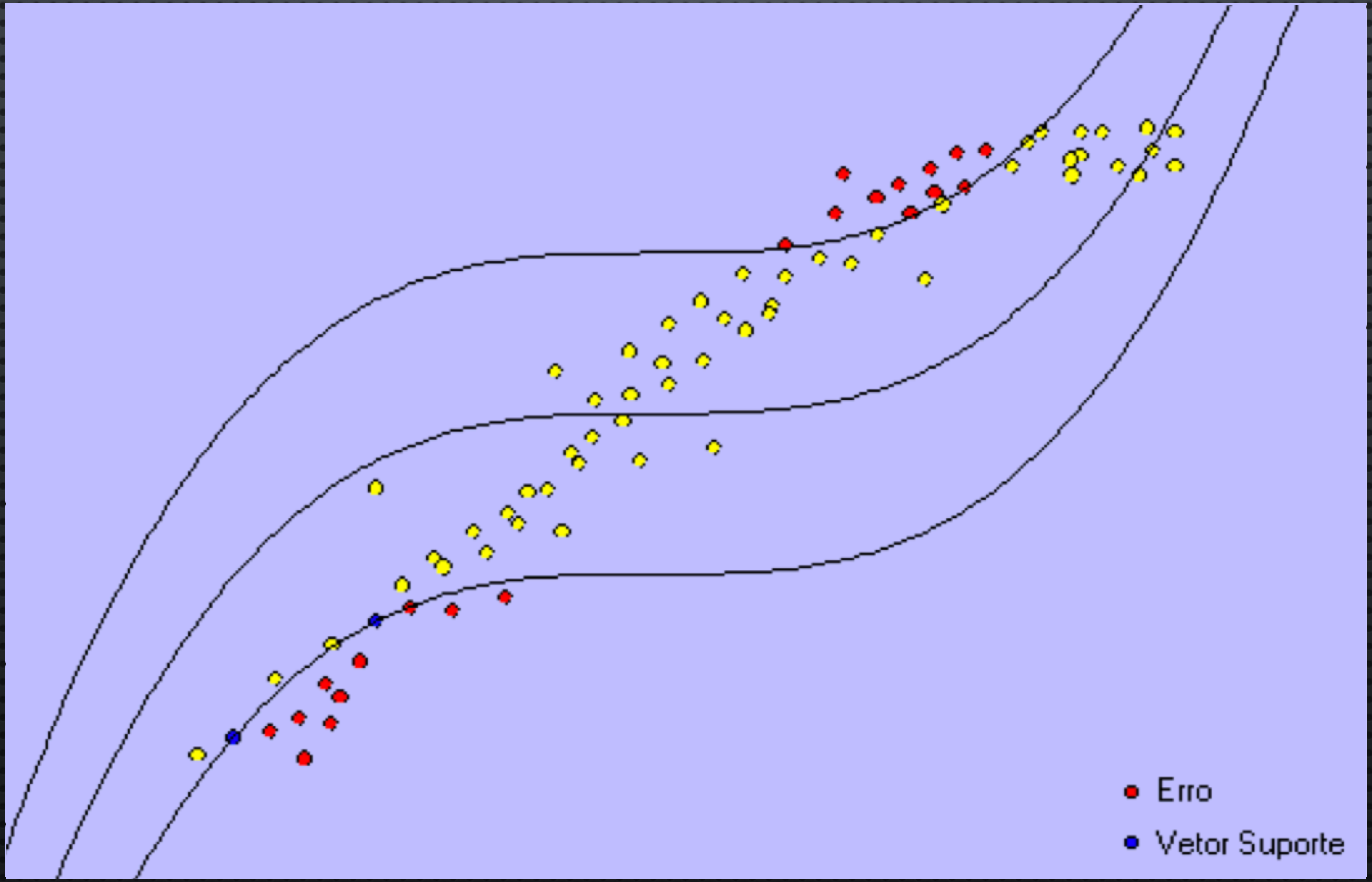
SVR

$C = 10, \varepsilon = 20$ E KERNEL LINEAR



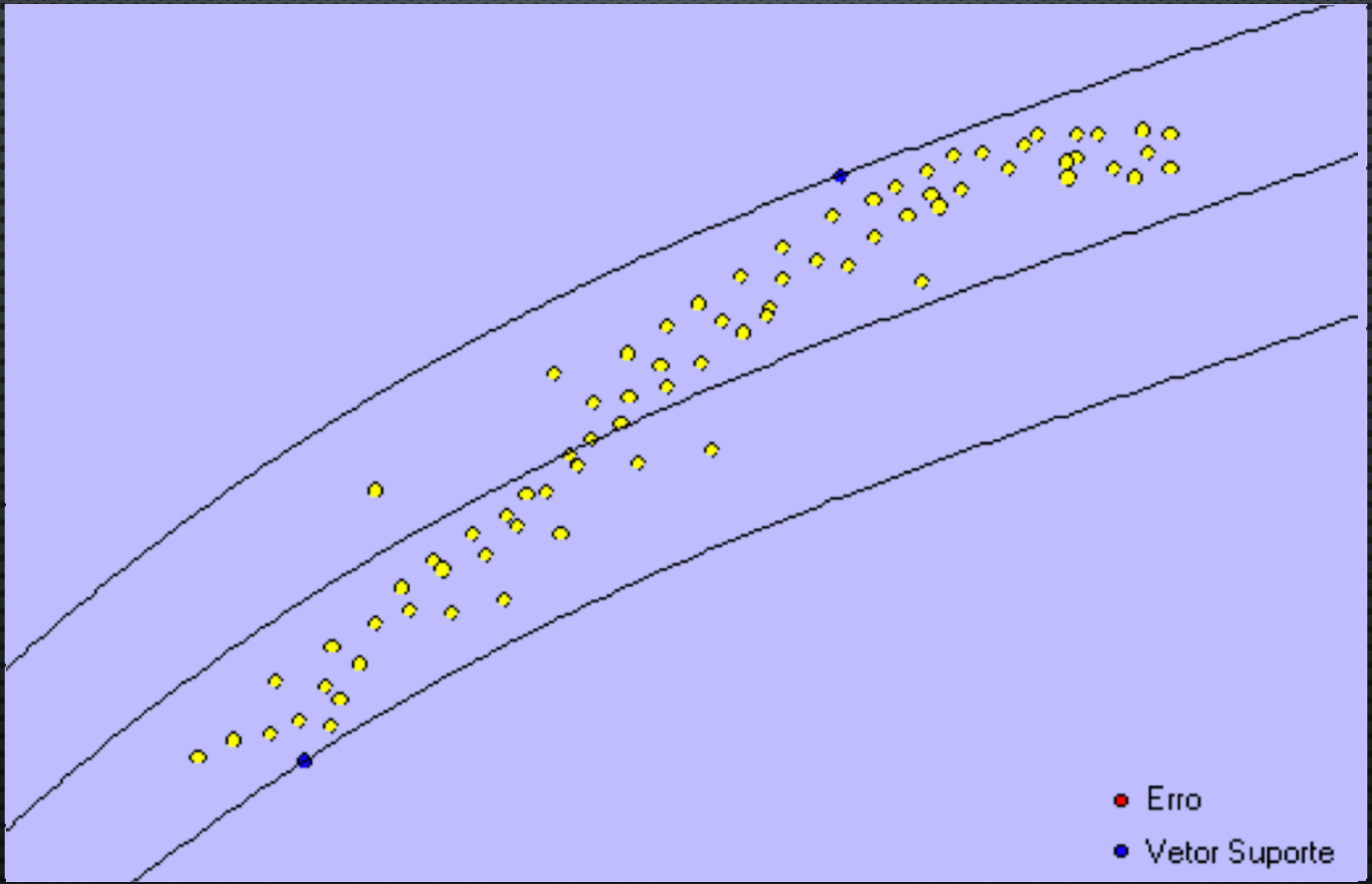
SVR

$C = 10, \varepsilon = 20$ E KERNEL POL. HOMOGÊNEO COM PARÂMETRO $p = 3$



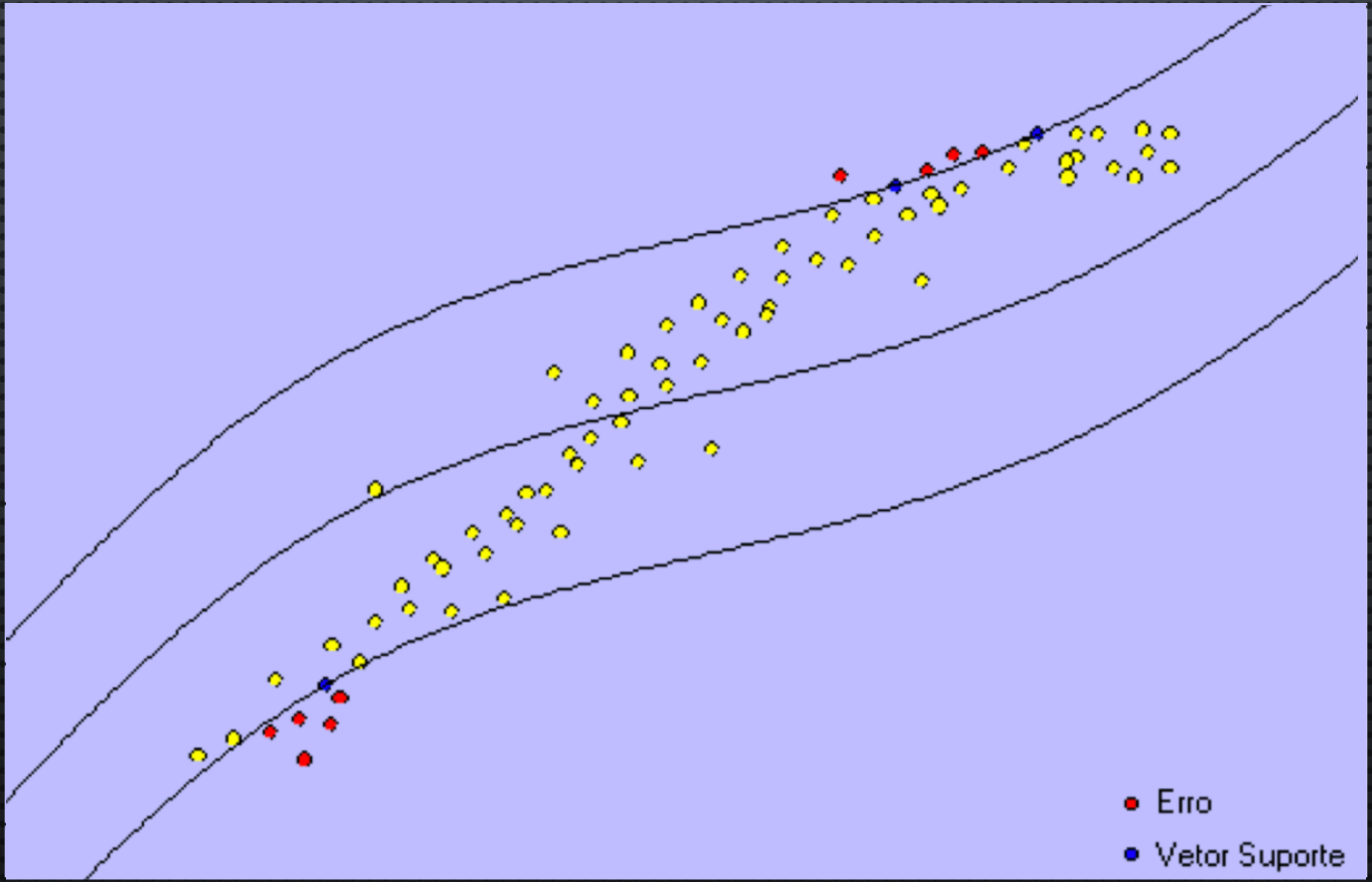
SVR

$C = 10, \varepsilon = 20$ E KERNEL POL. NÃO HOMOGÊNEO COM PARÂMETROS $p = 4$ E $c = 2$



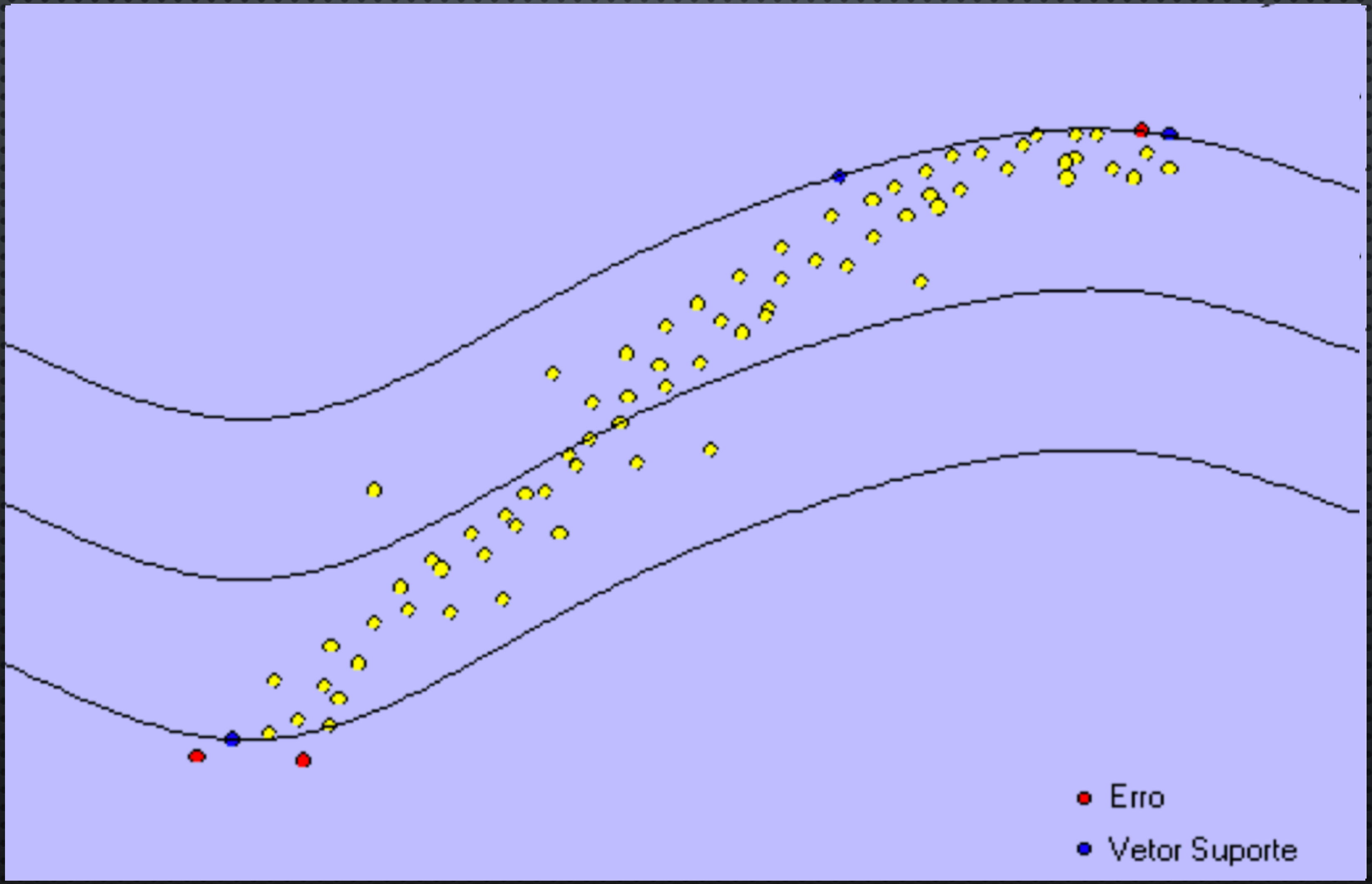
SVR

$C = 10, \varepsilon = 20$ e kernel sigmoidal com parâmetros $b = 2$ e $c = -1{,}7$



SVR

$C = 10, \varepsilon = 20$ E KERNEL GAUSSIANO COM PARÂMETRO $\sigma = 0,4$



REFERÊNCIAS

- BELTRAMI, M. Precificação de opções sobre ações por modelos de support vector regression. Dissertação de mestrado UFPR, 2009. Disponível em: https://www.acervodigital.ufpr.br/bitstream/handle/1884/27334/Dissertacao_Monica%20Beltrami.pdf?sequence=1&isAllowed=y
- LORENA, A. C. e CARVALHO, A. C. P. L. F. Uma introdução às support vector machines. RITA, vol. 10, 2007. Disponível em: https://seer.ufrgs.br/rita/article/viewFile/rita_v14_n2_p43-67/3543
- KOUTSOULERIS, N et al. Use of Neuroanatomical Pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch psychiatry. Vol.66, n. 7, 2009. Disponível em: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/483126
- Building a Student Intervention System. Disponível em: https://github.com/Tsmith5151/Student-Intervention
- Huang, H. J., Shi, Z. e Ding, S. Primal least squares twin support vector regression. Jounal of Zhejiang University. Vol. 14(9), p. 722-732 , 2013. Disponível em: https://www.researchgate.net/publication/269383070_Primal_least_squares_twin_support_vector_regression