3
Classificação Supervisionada

3.1.
Aprendizado de Máquina

A aprendizagem de máquina é uma área da inteligência artificial que estuda métodos computacionais, a fim de obter um determinado conhecimento específico através de experiências. Os algoritmos de aprendizado de máquina possuem o foco em métodos estatísticos e a aplicação prática de aprendizado de máquina inclui o processamento de linguagem natural, diagnósticos médicos, buscadores, entre outras.

O reconhecimento de padrões é um tópico de aprendizagem de máquina cujo objetivo é classificar informações baseadas ou em conhecimento a priori (supervisionado) ou em informações estatísticas extraídas dos padrões (não supervisionado). Um sistema de reconhecimento de padrões consiste em dois mecanismos, o primeiro relacionado à extração de informações (padrões), que processa as características semelhantes, e o segundo capaz de classificar as observações, dependendo das características extraídas anteriormente.

Nessa dissertação, focaremos apenas no reconhecimento de padrões supervisionado, que se utiliza de um conjunto dados de entrada, previamente rotulados, com objetivo principal de capacitar a máquina a aprender sobre o problema proposto.

A seguir, apresentaremos dois métodos de reconhecimento de padrão supervisionado, *Máxima Verossimilhança* e *Support Vector Machines*, mas para a classificação de imagens podemos utilizar outros métodos de aprendizado de máquina como Rede Neural, TBL, Árvore de Decisão entre outros.

3.2.
Máxima Verossimilhança

Considere a família D_θ de distribuições de probabilidades parametrizadas por um parâmetro θ desconhecido associado a uma função de densidade de

probabilidade (distribuição contínua) ou a uma função de massa de probabilidade (distribuição discreta), denominado f_θ.

Seja $x_1, x_2, \ldots, x_n$ um conjunto de n valores dessa distribuição, e usando f_θ para computar a densidade de probabilidade através dos dados observados. Assumindo que as amostras do conjunto $x_1, x_2, \ldots, x_n$ são obtidas de forma independente umas das outras, podemos utilizar um método de estimação de parâmetros conhecido como máxima verossimilhança para estimar o valor do parâmetro θ.

Dado a função de θ com $x_1, x_2, \ldots, x_n$ fixos, a função de verossimilhança é definida por:

$$L(\theta) = f_\theta(x_1, \ldots, x_n \mid \theta) \tag{3.2.1}$$

O método de máxima verossimilhança estima o parâmetro θ, encontrando o valor de $\hat{\theta}$ que maximiza $L(\theta)$, dada pela função abaixo, em que Θ representa o espaço de características [11]:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) \tag{3.2.2}$$

No caso da distribuição Gaussiana univariada, dois parâmetros devem ser estimados. Dessa maneira, o parâmetro θ da função de máxima verossimilhança passa a ser denotado pelo conjunto $\theta = \{\mu, \sigma^2\}$, em que μ e σ^2 representam a média e a variância, respectivamente. A representação resultante para a função de verossimilhança para uma distribuição Gaussiana é definida por:

$$L(\{\mu, \sigma^2\}) = f_{\{\mu,\sigma^2\}}(x_1, \ldots, x_n \mid \{\mu, \sigma^2\}) \tag{3.2.3}$$

3.3. Support Vector Machines

O modelo mais simples de Support Vector Machines (SVM), que também foi o primeiro a ser introduzido, é chamado Classificador de Margem Máxima. Ele trabalha apenas com dados linearmente separáveis, ficando restrito, portanto, a poucas aplicações práticas. Apesar dessa limitação, o Classificador de Margem Máxima apresenta propriedades importantes e é a pedra fundamental para a formulação de SVM mais sofisticadas.

A Figura 7 exibe um espaço de características linearmente separáveis para um conjunto de treinamento bidimensional e a Figura 8 ilustra um espaço linearmente inseparável. A linha escura presente em ambas as figuras, que separa os vetores de entrada de classes distintas, é chamada de superfície de decisão (ou separação). Em particular, na Figura 7, devido à linearidade da superfície de decisão, a mesma também é conhecida como hiperplano de separação.

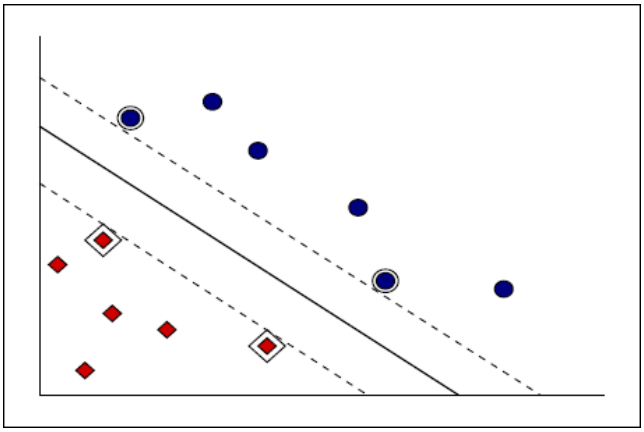

Figura 7 Espaço de características linearmente separável

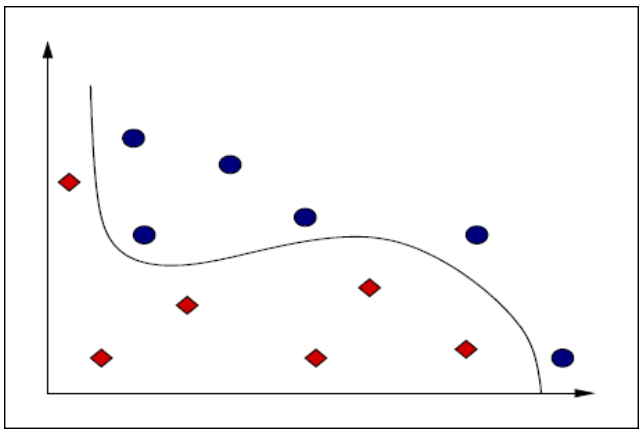

Figura 8 Espaço de características linearmente inseparável

O classificador de margem máxima otimiza limites no erro de generalização das máquinas lineares em termos da margem de separação entre as classes, a qual é determinada pelo hiperplano de separação. Essa estratégia envolve separar os dados com um tipo especial de hiperplano, o hiperplano de margem máxima ou de separação ótima, que é descrito na próxima seção.

3.3.1.
Hiperplanos de Separação Ótima

Um hiperplano é considerado de margem máxima (ou de separação ótima) se separa um conjunto de vetores sem erro e a distância entre os vetores (das classes opostas) mais próximos ao hiperplano é máxima [12]. A Figura 9(a) mostra um hiperplano com margem pequena e 9(b), um hiperplano de margem máxima, para um conjunto de treinamento bidimensional. Para o caso linearmente separável, o algoritmo de Support Vector Machines tem como objetivo encontrar esse hiperplano.

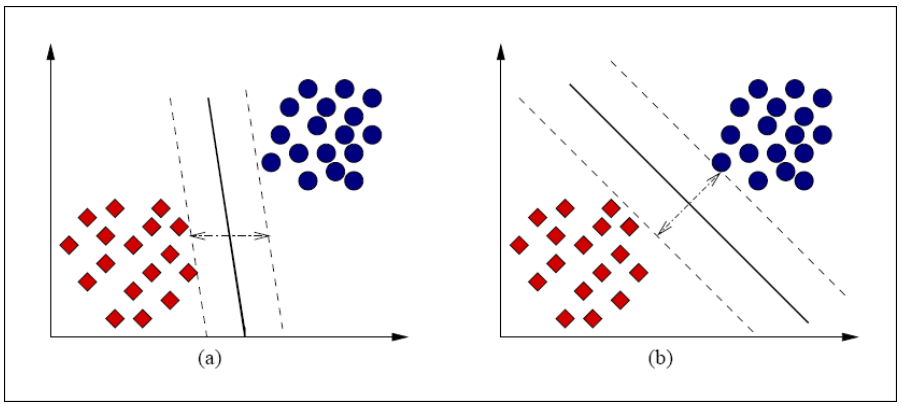

Figura 9 (a) Um hiperplano de separação com margem pequena. (b) Um Hiperplano de Margem Máxima

Considera um conjunto S de pontos de entrada $x_i \in R^N$ com $i = 1,2,\ldots,N$. Cada ponto x_i pertence a uma das duas classes, sendo fornecido, portanto, um rótulo $y_i \in \{-1,1\}$. Supõe-se que há um hiperplano que separa os exemplos positivos dos negativos. Os pontos x sobre o hiperplano satisfazem $w.x + b = 0$, em que w é normal (perpendicular) ao hiperplano, $|b|/\|w\|$ é a distância perpendicular do hiperplano à origem e $\|w\|$ é a norma Euclidiana de w. Seja d^+ a menor distância entre o hiperplano de separação e os pontos na fronteira da classe positiva $(+1)$, e d^- a menor distância entre o hiperplano de separação e os pontos mais próximos na fronteira da classe negativa (-1). A margem do hiperplano deve ser, dessa forma $d^+ + d^-$.

O algoritmo de SVM procura o hiperplano de separação com margem máxima que pode ser construído como segue. Assume-se que todos os dados de treinamento satisfazem as seguintes restrições:

$$x_i.w+b \geq +1, \text{ para } y_i = +1 \quad (3.3.1)$$

$$x_i.w+b \leq -1, \text{ para } y_i = -1 \quad (3.3.2)$$

Podemos combinar estas desigualdades e obter:

$$y_i(x_i.w+b)-1 \geq 0 \ \ \forall_i = 1,2,\ldots,N \quad (3.3.3)$$

Os pontos para os quais vale a igualdade na restrição (3.3.1) estão no hiperplano $H_1 : x_i.w+b=1$, que possui norma w e distância perpendicular à origem igual a $|1-b|/\|w\|$. Similarmente, os pontos para os quais vale a igualdade da restrição (3.3.2) estão no hiperplano $H_2 : x_i.w+b=-1$, com norma novamente w e distância perpendicular à origem igual a $|-1-b|/\|w\|$. Portanto, $d^+ = d^- = 1/\|w\|$ e a margem é simplesmente $2/\|w\|$.

Pode-se observar que H_1 e H_2 são paralelos (eles têm a mesma norma) e que não há pontos de treinamento entre eles. Assim, é possível encontrar o par de hiperplanos que geram a margem máxima, pela minimização de $\|w\|^2$, sujeito à restrição definida em 3.3.3.

A definição do hiperplano é a seguinte: dada uma amostra de treinamento linearmente separável representada da seguinte forma: $S = ((x_1, y_1),\ldots,(x_l, y_l))$, o hiperplano $w.x+b=0$ pode ser encontrado solucionando-se o problema de otimização:

$$\begin{aligned} &minimize_{w,b}(w.w), \\ &sujeito \ \ a \ \ y_i((x_i.w)+b) \geq 1, i = 1,\ldots,l \end{aligned} \quad (3.3.4)$$

A solução para um caso bidimensional típico deverá ter a forma representada na figura 10. Os pontos para os quais se aplica a igualdade na equação (3.3.3) (isto é, aqueles que estão em um dos hiperplanos H_1 ou H_2) e que, se forem removidos, devem alterar a solução encontrada, são chamados vetores de suporte. Esses pontos são indicados na Figura abaixo por círculos extras.

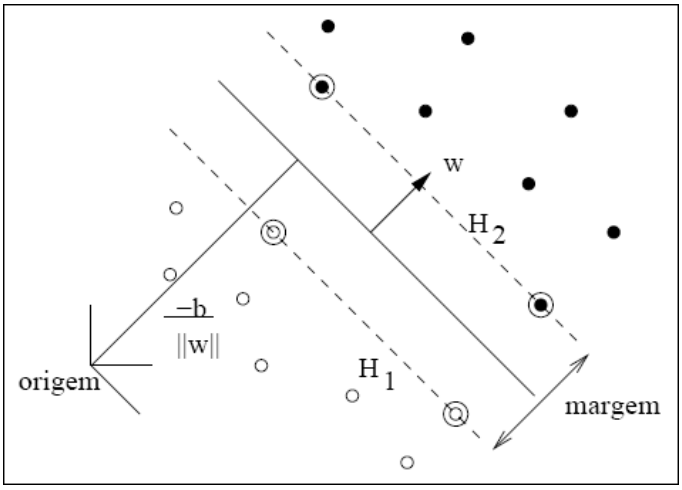

Figura 10 Hiperplano de separação para o caso linearmente separável. Os vetores de suporte estão circulados.

Para transformar o problema (3.3.4) em um problema quadrático, é necessário reescrevê-lo da forma abaixo:

$$\begin{aligned} &minimize_{w,b}\ \frac{1}{2}(w.w), \\ &sujeito\ \ a\ \ y_i((x_i.w)+b)\geq 1, i=1,\ldots,l \end{aligned} \tag{3.3.5}$$

Para solucionar esse problema de otimização é usado o método dos multiplicadores de Lagrange. Há duas principais razões para isso:

1. A restrição (3.3.3) é substituída por uma nova restrição, que é definida em função dos multiplicadores de Lagrange, os quais são mais fáceis de manusear computacionalmente;
2. Nessa reformulação do problema, os dados de treinamento apenas aparecem na forma de produto interno entre vetores. Essa é uma propriedade crucial que permite generalizar o procedimento para o caso não linear [13].

São introduzidos os multiplicadores de Lagrange $\alpha_i = 1,\ldots,l$. É realizado o produto entre a restrição (3.3.3) e os multiplicadores de Lagrange positivos e esse produto é subtraído da função objetivo para formar a funcional de Lagrange. Portanto, para solucionar o problema (3.3.5), deve-se encontrar o ponto de sela da seguinte funcional de Lagrange:

$$L_P = \frac{1}{2}(w.w)^2 - \sum_{i=1}^{l}\alpha_i y_i (x_i.w+b) + \sum_{i=1}^{l}\alpha_i \tag{3.3.6}$$

A derivada de L_P em relação a w e b, deve ser nula, isso corresponde ao fato de que no ponto ótimo, têm-se as seguintes equações de ponto de sela:

$$\frac{\partial L_P(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{l} y_i \alpha_i x_i = 0 \tag{3.3.7}$$

$$\frac{\partial L_P(w,b,\alpha)}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{3.3.8}$$

substituindo as relações obtidas, têm-se:

$$w = \sum_{i=1}^{l} y_i \alpha_i x_i \tag{3.3.9}$$

$$\sum_{i=1}^{l} y_i \alpha_i = 0 \tag{3.3.10}$$

Dadas essas restrições acima, pode-se substituí-las na Equação (3.3.6) e obter a formulação dual, uma vez que é computacionalmente mais eficiente encontrar o ponto de sela na formulação dual [14], a qual é definida da seguinte forma:

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{3.3.11}$$

É importante observar que L_P é o problema primal e L_D o problema dual. A solução é encontrada pela minimização de L_P ou maximização de L_D. Há um multiplicador de Lagrange para cada ponto de treinamento. Na solução, os pontos para os quais $\alpha_i > 0$ são chamados vetores de suporte e estão em um dos hiperplanos H_1 ou H_2. Para SVM, os vetores de suporte são os elementos críticos do conjunto de treinamento. Eles estão na fronteira, ou seja, mais próximos do hiperplano de decisão. Todos os outros pontos têm $\alpha_i = 0$. Se todos esses outros pontos forem removidos e o treinamento for repetido, o mesmo hiperplano deve ser encontrado [13].

Como os chamados vetores de suporte possuem α_i não nulos, eles são os únicos envolvidos na expressão do vetor peso w, logo, o vetor peso que representa o hiperplano de margem máxima é calculado na forma da combinação linear abaixo:

$$w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i \tag{3.3.12}$$

onde α^*, w^* pertencem a solução ótima. Pode, ainda, ser reescrito em função apenas dos vetores de suporte:

$$w^* = \sum_{vetores\ \ de\ \ \sup orte} y_i \alpha_i^* x_i \tag{3.3.13}$$

Reescrevendo o problema novamente, agora colocando a expressão para w^* na funcional de Lagrange dual, o problema quadrático de SVM torna-se o seguinte:

$$\begin{aligned} &\text{Maximize} \quad W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i . x_j) \\ &\text{Sujeito a} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \\ &\qquad\qquad \alpha_i \geq 0, i = 1, \ldots, l \end{aligned} \tag{3.3.14}$$

Dado $\alpha_0 = (\alpha_1^0, \ldots, \alpha_l^0)$ ser uma solução para esse problema, então, a norma do vetor w, que corresponde ao hiperplano ótimo é igual a:

$$\|w\|^2 = 2W(\alpha_0) = \sum_{vetores\ \ de\ \ \sup orte} \alpha_i^0 \alpha_j^0 (x_i . x_j) y_i y_j \tag{3.3.15}$$

A regra de separação, baseada no hiperplano ótimo, é a seguinte função indicadora:

$$f(x) = sign\left(\sum_{vetores\ \ de\ \ \sup orte} y_i \alpha_i^0 (x_i . x) - b_0 \right) \tag{3.3.16}$$

na qual x_i são os vetores de suporte, a_i^0 são os Coeficientes de Lagrange correspondentes e b_0 é um limiar constante:

$$b_0 = \frac{1}{2} \left[(w_0 . x^*(1)) + (w_0 . x^*(-1)) \right] \tag{3.3.17}$$

em que $x^*(1)$ corresponde à qualquer vetor de suporte pertencente à primeira classe e $x^*(-1)$, um vetor de suporte pertencente à segunda classe.

O classificador de margem máxima, quando aplicado a dados não separáveis linearmente, não encontra a solução desejada. Isso é evidenciado pela função objetivo (dual) que, aplicada a dados não linearmente separáveis, cresce arbitrariamente. O principal problema desse classificador é que ele sempre constrói hipóteses que se baseiam na inexistência de erros de treinamento. Entretanto, para dados com ruídos, que geralmente implica em separação não linear, o mínimo para o risco esperado não pode ser calculado dessa forma, pois pode causar *overfitting*. Essas desvantagens motivaram o

desenvolvimento de técnicas que permitem o tratamento de problemas não linearmente separáveis via SVM.

3.3.2. Support Vector Machines Não Lineares

Para tornar o método descrito na seção anterior capaz de manipular dados não linearmente separáveis, é necessário "relaxar" as restrições do problema. Diferentemente das restrições (3.3.1) e (3.3.2) da seção anterior que utilizam critérios rígidos, a estratégia apresentada nesta seção utiliza um critério mais relaxado. Isso pode ser feito introduzindo variáveis de folga ($\xi_i, i = 1, \ldots, N$) nas restrições, as quais se tornam:

$$x_i.w + b \geq +1 - \xi \text{, para } y_i = +1 \tag{3.3.18}$$

$$x_i.w + b \leq -1 + \xi \text{, para } y_i = -1 \tag{3.3.19}$$

$$\xi_i \geq 0 \text{ , } \forall i \tag{3.3.20}$$

Essa estratégia permite tolerar ruídos e *outliers* (pontos muito distantes das classes a que pertencem), considera mais pontos de treinamento, além dos que estão na fronteira, e permite a ocorrência de erros de classificação. Portanto, $\sum_{i=1}^{N} \xi_i$ é um limite superior para o número de erros de treinamento. Além disso, a fim de poder representar o custo extra para os erros, decorrente da adição das variáveis de folga, há a necessidade de mudar a função-objetivo a ser minimizada de $\frac{1}{2}\|w\|^2$ para:

$$\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{N} \xi_i\right)^k \tag{3.3.21}$$

na qual C é um parâmetro a ser escolhido pelo usuário. C é uma constante que atua como uma função de penalidade e prevenindo que *outliers* afetem o hiperplano ótimo [15]. $C > 0$ determina a relação entre o erro empírico e o termo de confiança. Um C maior corresponde a assumir uma penalidade maior para os erros. Por se tratar de um problema de programação convexa, o valor de k pode ser qualquer inteiro positivo em particular; se $k = 1$ ou $k = 2$, então, o problema também é de programação quadrática. Por razões computacionais, entretanto, uma escolha típica é $k = 1$. Esse caso corresponde ao menor dos $k > 0$ e tem a

vantagem de não ser necessário que ξ_i e seus multiplicadores de Lagrange, apareçam no problema dual. O problema, com essa alteração, torna-se:

$$\text{Maximize} \quad L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \left(x_i . x_j \right)$$

$$\text{Sujeito a} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad (3.3.22)$$

$$0 \le \alpha_i \le C, i = 1, \ldots, l$$

A solução é dada novamente por:

$$w = \sum_{vetores \ \ de \ \ \text{sup}orte} y_i \alpha_i x_i \qquad (3.3.23)$$

Assim, a única diferença do caso do hiperplano ótimo é que os α_i têm um limite superior em C. Essa situação é representada abaixo pela figura.

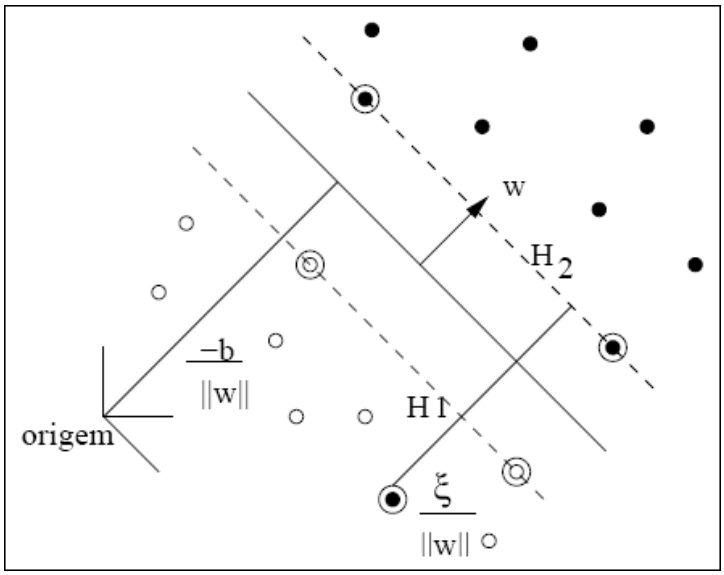

Figura 11 Hiperplano de separação para o caso linearmente inseparável.

Com o conteúdo que foi descrito nas seções anteriores, é possível, a partir deste ponto, descrever formalmente a construção da técnica SVM para uma tarefa de reconhecimento de padrões. A técnica SVM implementa basicamente duas operações matemáticas:

1. Mapeamento não-linear dos vetores de entrada x em um espaço de características Z com alta dimensão;
2. Construção de um Hiperplano de Margem Máxima no espaço de características.

A Figura 12 ilustra a estratégia do Support Vector Machines, de mapear o espaço real em um espaço de características.

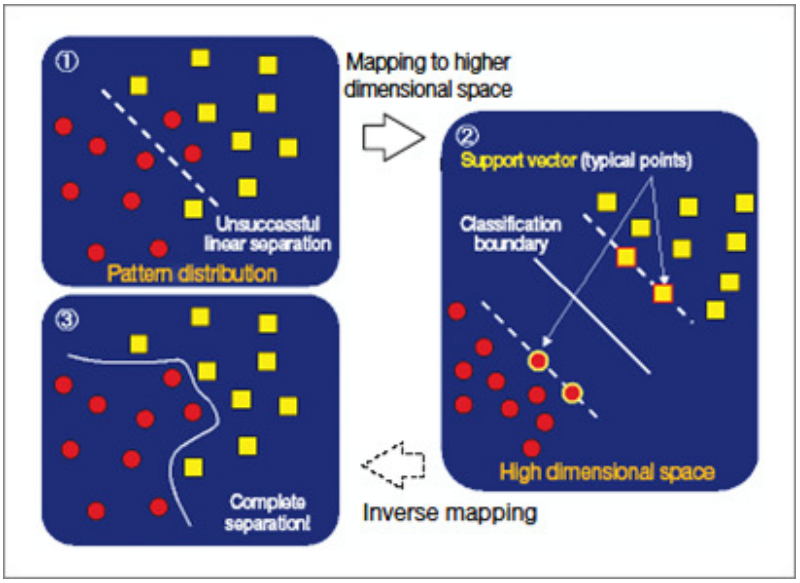

Figura 12 Ilustração da estratégia de Support Vector Machines

As representações Kernel trabalham com a projeção dos dados em um espaço de características com alta dimensão para permitir a classificação em espaços não linearmente separáveis. Trata-se, em primeira instância, de uma estratégia de pré-processamento que envolve mudar a representação dos dados da seguinte forma:

$$x = (x_1, \ldots, x_n) \longrightarrow \phi(x) = (\phi_1(x), \ldots, \phi_N(x)) \qquad (3.3.24)$$

Esse passo é equivalente ao mapeamento do espaço de entrada X em um novo espaço $Z = \{\phi(x) | x \in X\}$ chamado espaço de características em que ϕ_i são as funções Kernel. A Figura 13 ilustra um mapeamento de um espaço de entrada linearmente inseparável, para um espaço de características de maior dimensão, em que os dados podem ser separados linearmente.

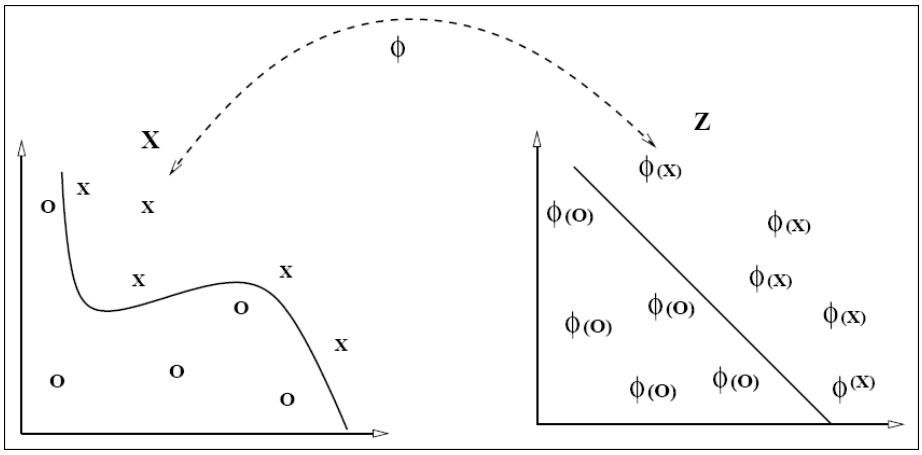

Figura 13 Mapeamento do espaço de entrada via função de Kernel

A escolha da função Kernel é de vital importância para SVM e não altera muito o problema de SVM, pelo menos não explicitamente. Com a introdução da função Kernel, para encontrar os coeficientes α_i, é necessário agora resolver o seguinte problema:

$$\begin{aligned} &\text{Maximize} \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i . x_j) \\ &\text{Sujeito a} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \\ &\qquad\qquad 0 \le \alpha_i \le C, i = 1, \ldots, l \end{aligned} \tag{3.3.25}$$

Essa funcional coincide com a funcional para encontrar o Hiperplano de Margem Máxima, exceto pela forma do produto interno, que anteriormente era representado unicamente por produto interno $(x_i \cdot x_j)$ e agora passa a ser representado pela função Kernel $K(x_i, x_j)$. Com a função Kernel, a função de decisão passa a ser representada como:

$$f(x) = sign\left(\sum_{vetores \ \ de \ \ \text{sup}orte} y_i \alpha_i K(x_i . x) - b \right) \tag{3.3.26}$$

O uso de diferentes funções Kernel $K(x_i, x_j)$ possibilita a construção de máquinas de aprendizagem com diferentes tipos de superfícies de decisão não-linear no espaço de entrada. Entre as funções Kernel mais usadas, destacam-se: Polinômios [13], Funções de Base Radial Gaussiana [16] e Rede Neural Sigmóide de duas camadas [17] [18].

3.3.3. Support Vector Machines Multiclasses

Embora o SVM separe os dados linearmente em duas classes, o reconhecimento de mais do que duas classes é possível utilizando a estratégia de decomposição do problema multiclasses em subproblemas binários. Existem algumas técnicas que são utilizadas para resolver o problema de multiclasses, as mais conhecidas são a *One-Against-All* e *One-Against-One*.

Seja N o número de classes, a técnica *One-Against-All* consiste em separar uma classe A e agrupar as $N-1$ classes restantes em uma classe B, a partir da separação encontraremos o hiperplano que separa a classe A da classe B. O processo de separação da classe é realizado N vezes para cada

classe pertencente ao conjunto de classes, logo, encontraremos N hiperplanos que separam as classes.

Seja N o número de classes, a técnica *One-Against-One* consiste em separar duas classes A e B do conjunto de classes e encontrar um hiperplano que separe esse par de classes. O processo de separação é realizado para cada par de classes pertencente ao conjunto de classes, logo encontraremos N hiperplanos que separam as classes.

Abaixo, apresentamos duas figuras que representam as duas técnicas de decomposição, *One-Against-All* e *One-Against-One*, respectivamente [19], [20].

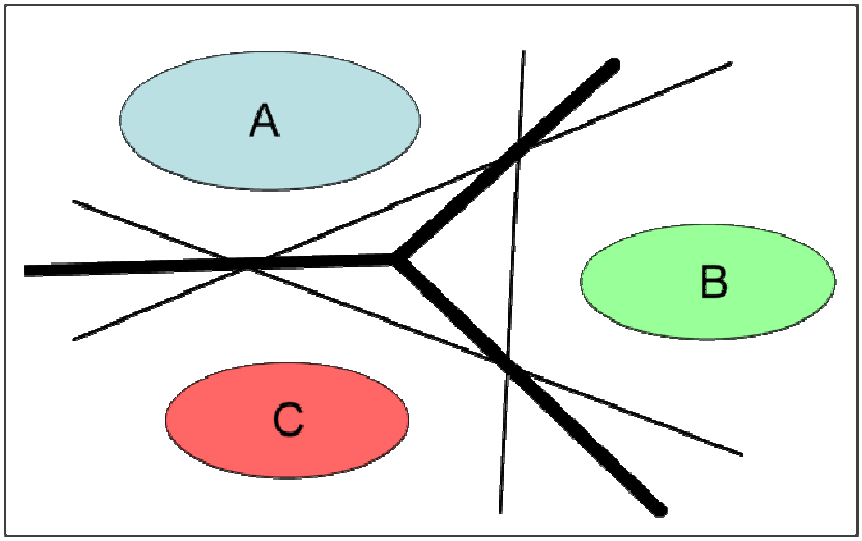

Figura 14 Exemplifica a técnica de decomposição *One-Against-All* para três classes

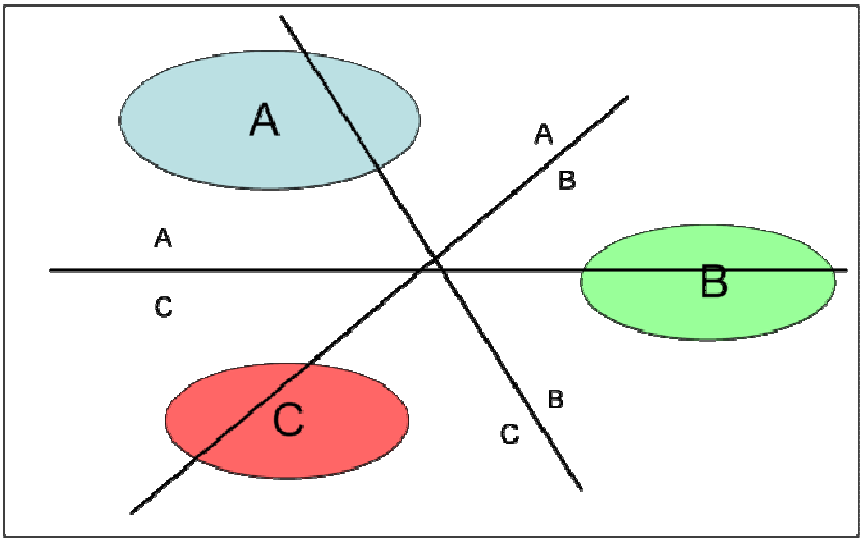

Figura 15 Exemplifica a técnica de decomposição *One-Against-One* para três classes