SVM
SUPPORT VECTOR MACHINE

Prof. André Backes | @progdescomplicada

Definição

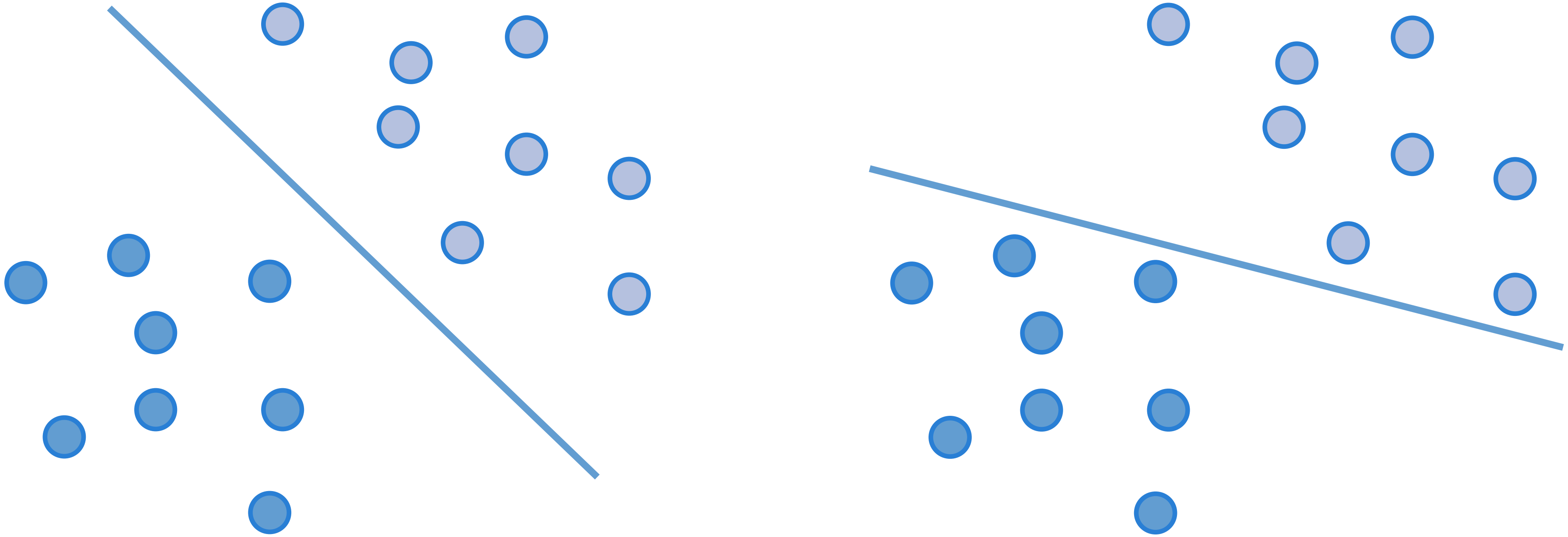
- Máquinas de Vetores Suporte (*Support Vector Machines* - SVMs)
  - Proposto em 79 por Vladimir Vapnik
  - Um dos mais importantes acontecimentos na área de reconhecimento de padrões nos últimos 15 anos.
  - Tem sido largamente utilizado com sucesso para resolver diferentes problemas.

Definição

- Máquinas de Vetores Suporte (*Support Vector Machines* - SVMs)
  - É uma técnica de classificação supervisionada
  - Trata-se de um classificador linear binário não-probabilístico
    - Classifica os dados sempre em apenas 2 classes
  - Resultados comparáveis aos obtidos por outros algoritmos de aprendizado
    - Redes Neurais Artificiais (RNAs)

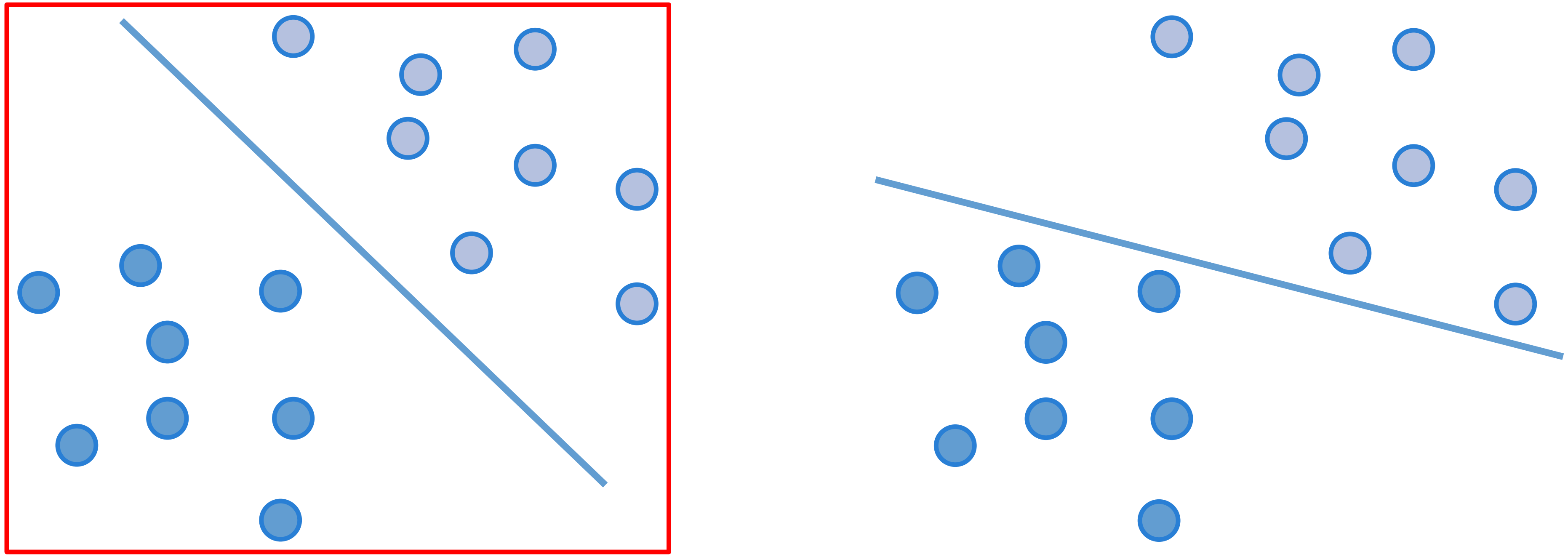
Definição

- Ideia geral
  - Perceptron é capaz de construir uma fronteira se os dados forem linearmente separáveis
  - Mas qual fronteira é a melhor?



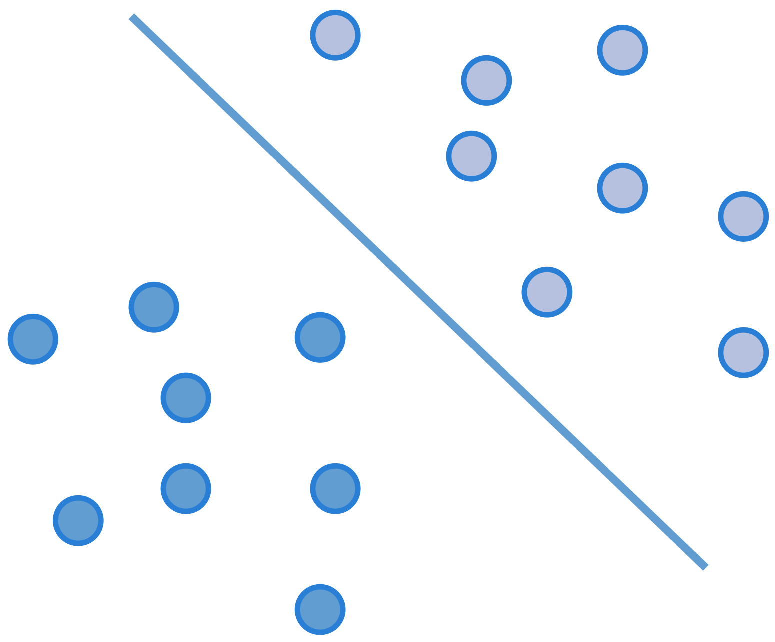
Definição

- Ideia geral
  - SVM trabalha com a maximização da margem
  - A fronteira mais distante dos dados de treinamento é a melhor



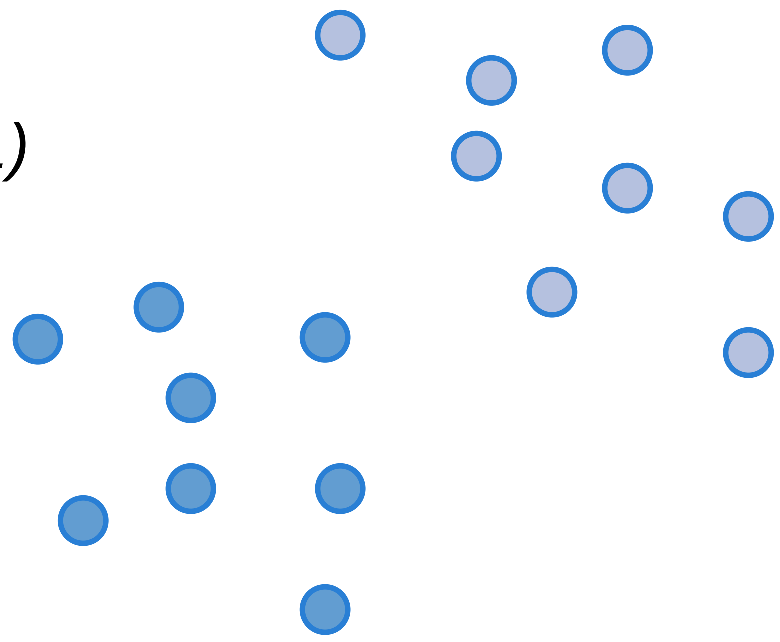
SVM Linear

- SVM linear com margens rígidas
  - Define uma fronteira linear a partir de dados linearmente separáveis
    - Separam os dados por meio de um hiperplano
  - Conjunto de dados contendo somente duas classes
    - -1 e +1



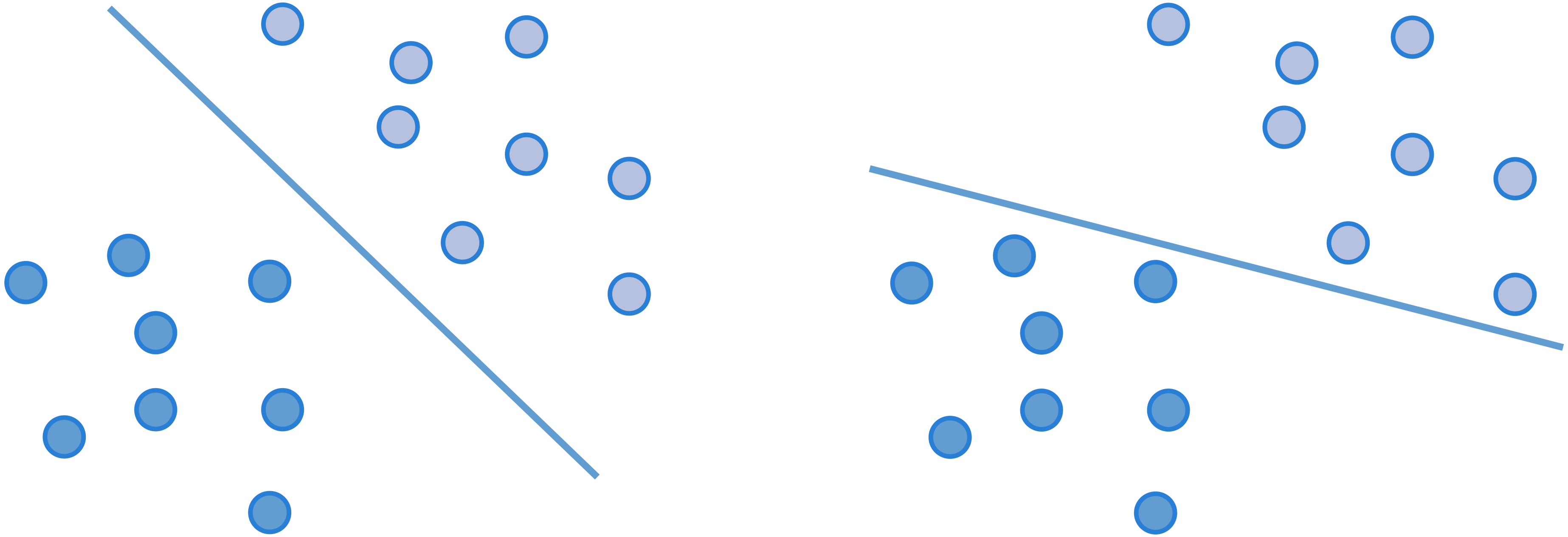
SVM Linear

- Um dado é visto como um ponto num espaço de p dimensões
  - Queremos saber se podemos separar esses pontos com um hiperplano de *(p - 1)* dimensões
  - Problema Linearmente Separável



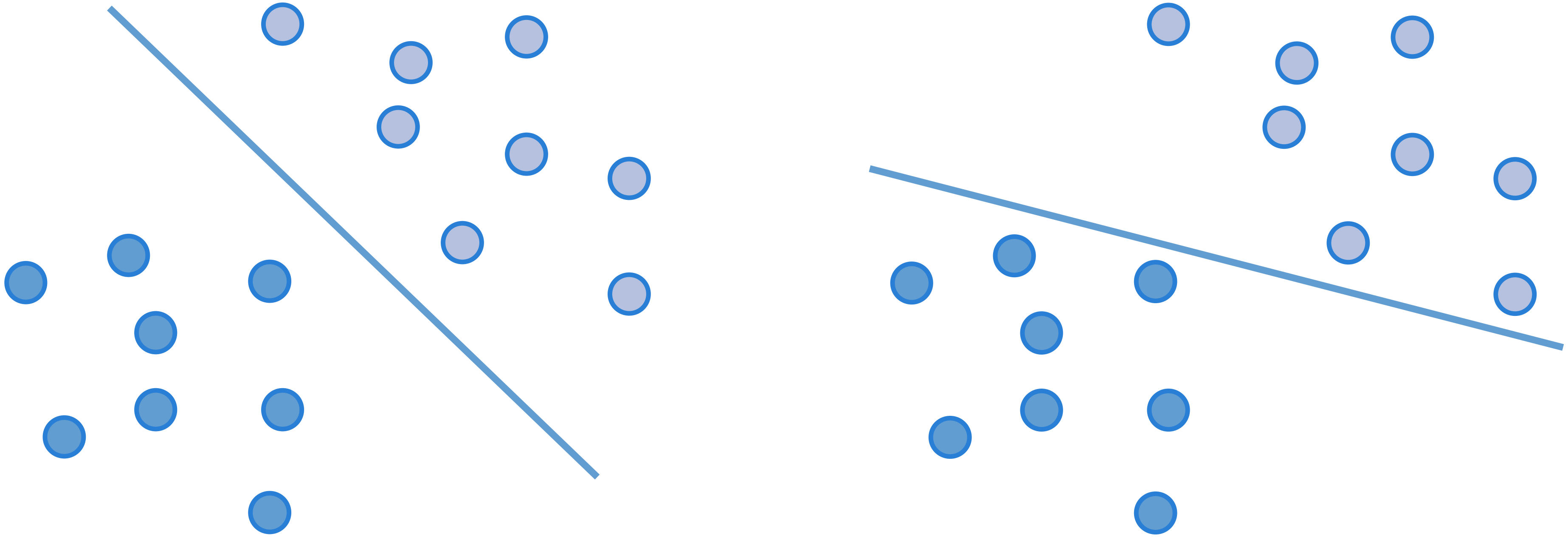
SVM Linear

- Existem muitos hiperplanos possíveis
  - Duas possíveis soluções (mas existem outras...)
  - Qual é o melhor hiperplano?



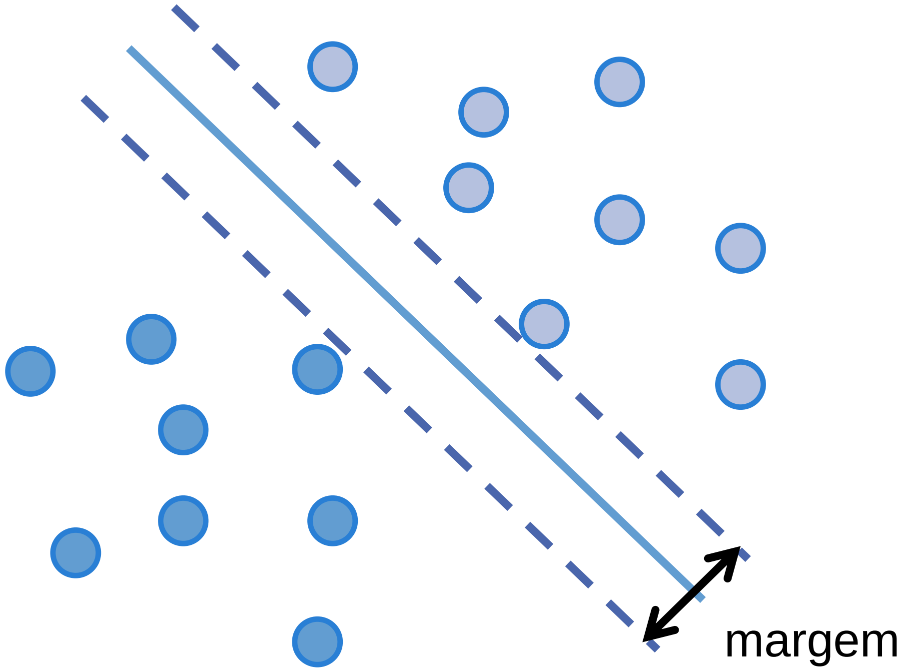
SVM Linear

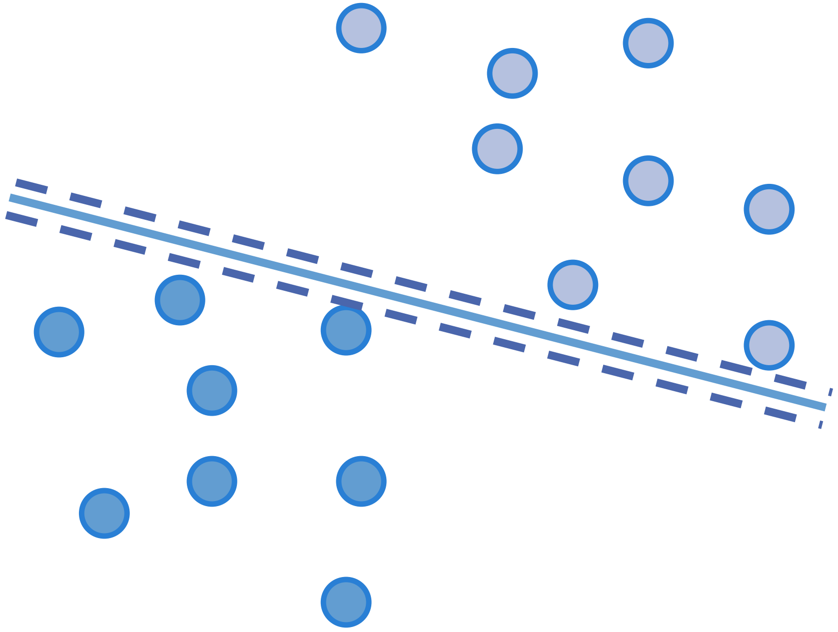
- Existem muitos hiperplanos possíveis
  - SVM busca o hiperplano máximo
  - Maior separação, ou margem, entre as duas classes



SVM Linear

- Solução
  - Encontrar o hiperplano que maximiza a margem do limiar de decisão



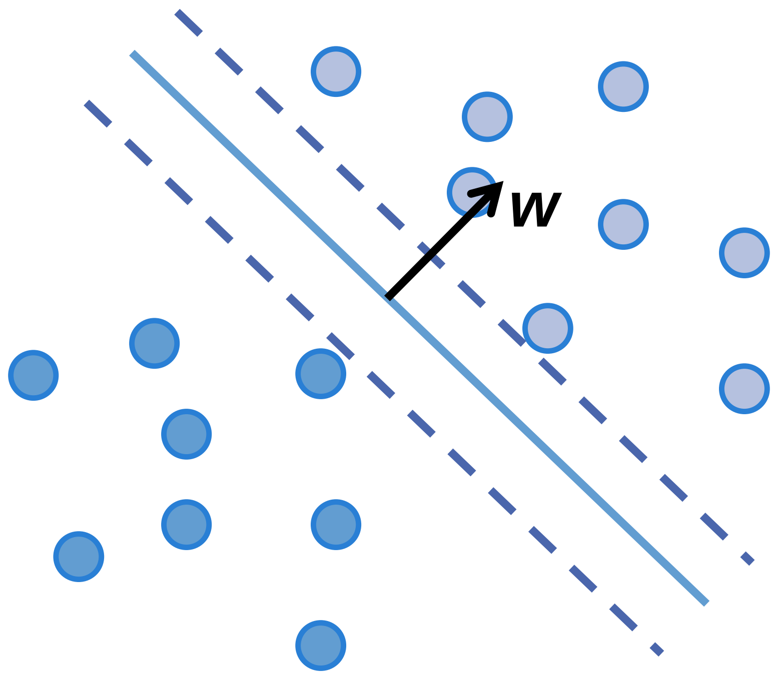


SVM Linear

- O hiperplano de separação é dado pela equação

$$f(x) = wx + b = 0$$

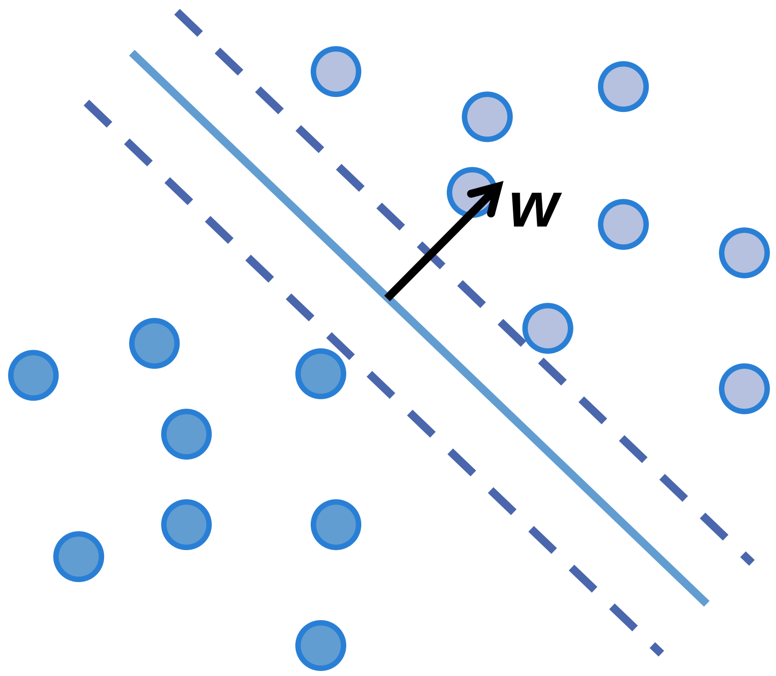
- Onde *w* é o vetor de pesos (mesma dimensão das amostras) perpendicular ao hiperplano de separação e *b* é um escalar.



SVM Linear

- A equação divide o espaço duas regiões
  - $wx + b > 0$
  - $wx + b < 0$

- Apenas o sinal é necessário para fazer a classificação
  - $y(x) = \begin{cases} +1, se\ wx + b > 0 \\ -1, se\ wx + b < 0 \end{cases}$

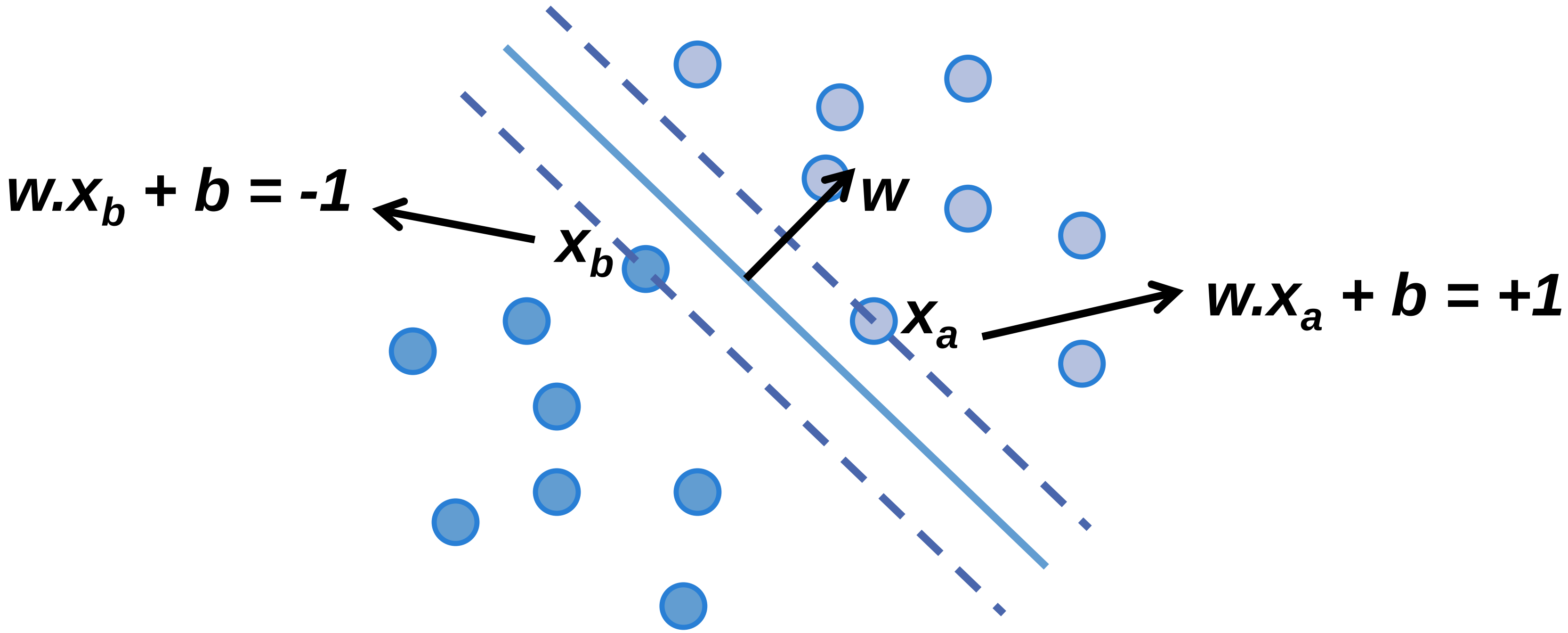


SVM Linear

- Essa equação permite obter um número infinito de hiperplanos equivalentes. Qual escolher?
  - Selecionar *w* e *b* de forma que os exemplos mais próximos ao hiperplano satisfaçam
    - $|wx + b| = 1$
  - Assim temos que
    - $\begin{cases} wx + b \geq +1 \; se \; y = +1 \\ wx + b \leq -1 \; se \; y = -1 \end{cases}$

SVM Linear

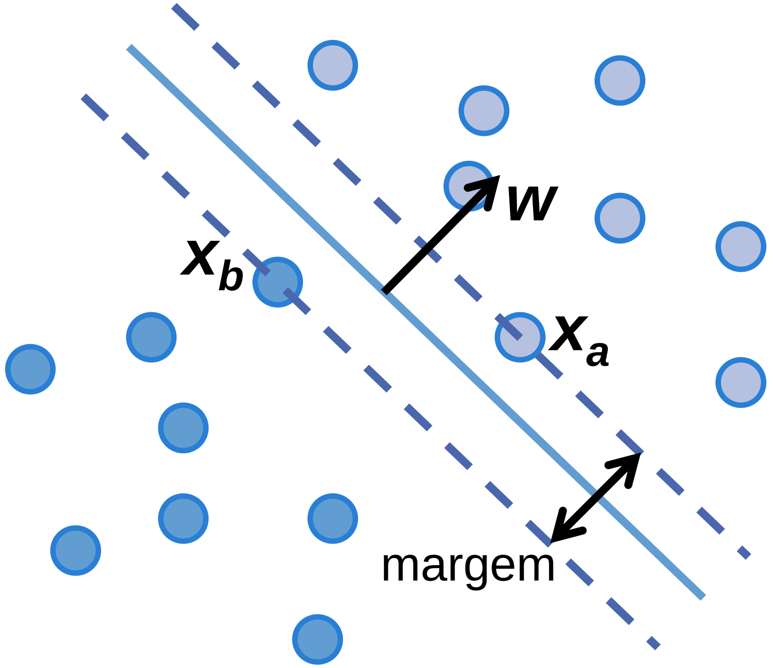
- Como escolher *w* e *b*?
- Seja dois pontos x_a e x_b
  - $wx_a + b = +1$
  - $wx_b + b = -1$



SVM Linear

- A diferença entre as equações
  - $wx_a + b = +1$
  - $wx_b + b = -1$

- Fazendo a diferença entre os hiperplanos de x_a e x_b
  - $w(x_a - x_b) = 2$



SVM Linear

- Calculando a margem
  - É a distância entre os hiperplanos x_a e x_b
  - Diferença entre hiperplanos
    - $w(x_a - x_b) = 2$

  - Margem é o comprimento do vetor diferença projetado na direção de w
    - $\|x_a - x_b\| = \frac{2}{\|w\|}$
    - $margem = \frac{2}{\|w\|}$

SVM Linear

- Temos então que
  - $margem = \frac{2}{||w||}$
- A distância mínima entre o hiperplano separador e os dados é dada por $w(x_a - x_b) = 2$
  - $\frac{1}{||w||}$
- Logo, maximizar a margem envolve minimizar
  - $||w||$

SVM Linear

- Minimizar $||w||$ é um problema difícil de resolver
  - Depende da norma de ***w***, a qual envolve uma raíz

- Solução
  - Substituir o termo $||w||$ por $\frac{1}{2}||w||^2$
  - Removemos o cálculo da raiz e acrescentamos uma constante por conveniência matemática

SVM Linear

- Temos agora um problema de otimização de uma função quadrática

- Vamos maximizar a margem do limiar de decisão em função do vetor de pesos ***w*** (***forma primal***)
  - $\min_{w,b} \frac{1}{2} ||w||^2$

SVM Linear

- Problema de otimização sujeito a seguinte restrição
  - $y_i(w.x_i + b) \geq 1, i = 1, \dots n$

- Devemos lembrar que
  - $x_i, i = 1, \dots n$, conjunto de padrões
  - $y_i = \{-1, +1\}, i = 1, \dots n$, respectivas classes

SVM Linear

- Problema de otimização
  - Trata-se de um problema quadrático com restrições lineares
  - Função objetivo é convexa, logo há somente um mínimo, que é o global
  - Solução global ótima é encontrada usando métodos numéricos

SVM Linear

- Solução com *função de Lagrange*
  - Problemas desse tipo podem ser solucionados com a introdução de uma *função de Lagrange*, que engloba as restrições à função objetivo, associadas a parâmetros denominados multiplicadores de Lagrange $\alpha_i \geq 0$

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i (w x_i + b) - 1)$$

  - A *função de Lagrange* deve ser minimizada, o que implica em maximizar as variáveis α_i e minimizar ***w*** e ***b***

SVM Linear

- Podemos aplicar os mesmos princípios de resolução encontrando o gradiente para a *função de Lagrange*
  - Essa abordagem também permite chegar à ***forma dual*** do problema. Isso envolve encontrar
  - $\frac{\partial L}{\partial b} = 0$
  - $\frac{\partial L}{\partial w} = 0$

SVM Linear

- Resolvendo
  - $\frac{\partial L}{\partial b} = 0$
  - $\frac{\partial L}{\partial w} = 0$

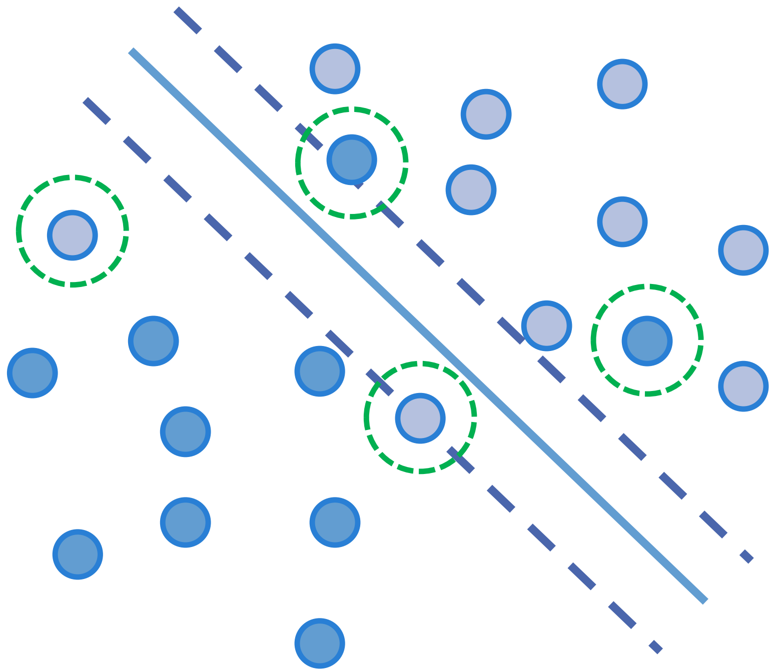
- Chegamos, respectivamente, a
  - $\sum_{i=1}^{n} \alpha_i y_i = 0$
  - $w = \sum_{i=1}^{n} \alpha_i y_i \, x_i$

SVM Linear

- Substituindo os termos anteriores a ***forma primal***, o problema passa a ser otimizar
  - $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$
- Sujeito a
  - $\begin{cases} \alpha_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$

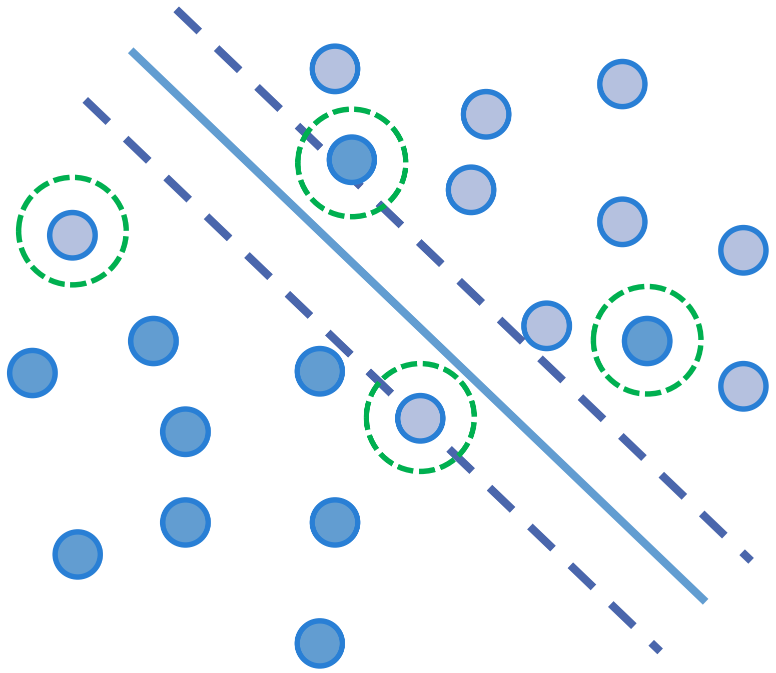
SVM Linear com Margens Suaves

- E se o problema for não linearmente separável ?
  - Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis.
  - Presença de ruído e exemplos inconsistentes (*outliers*)



SVM Linear com Margens Suaves

- Necessidade de uma nova abordagem
  - É empregada quando não há um hiperplano que divida os exemplos em +1 e 1
  - Permite-se que alguns dados possam violar a restrição



SVM Linear com Margens Suaves

- Nova abordagem
  - Flexibilizar as restrições de otimização utilizando variáveis de relaxamento do problema
  - Essas variáveis são conhecidas como “variáveis de folga”
  - São utilizadas para medir o grau de classificação errônea no conjunto de treinamento

SVM Linear com Margens Suaves

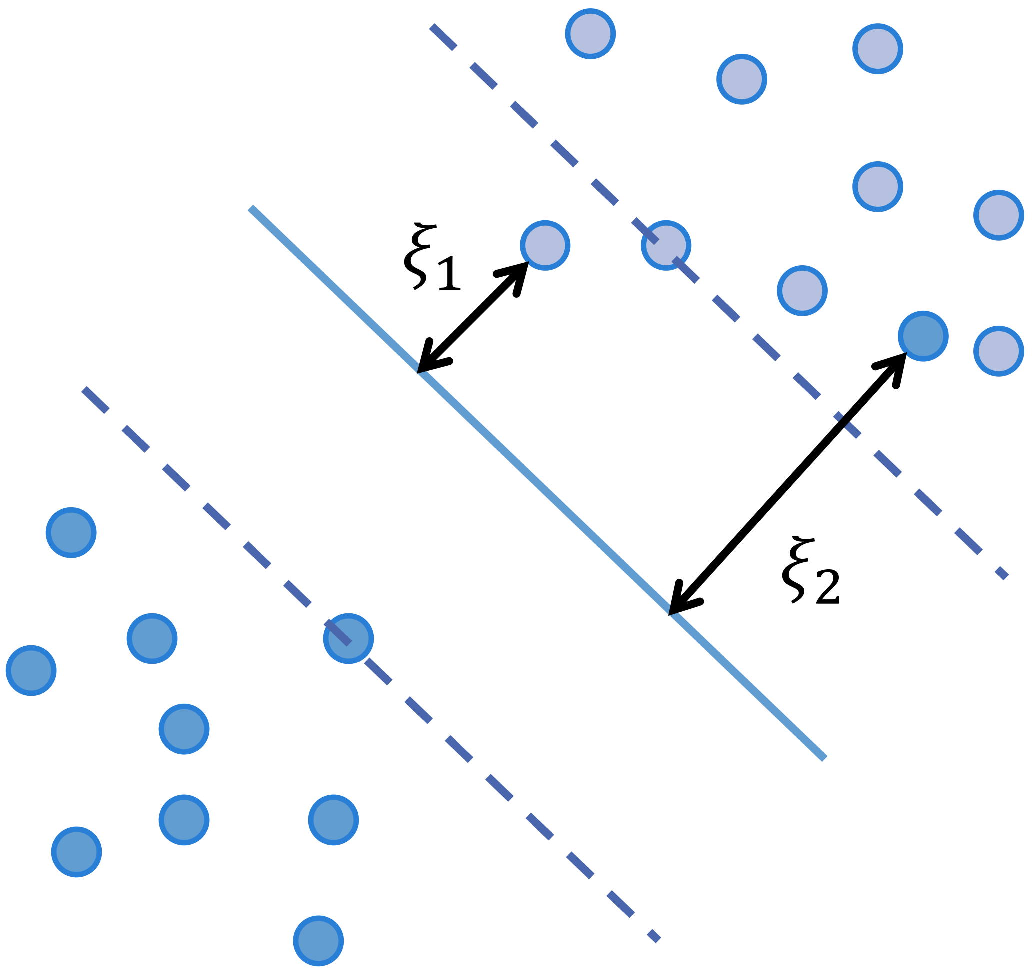
- Nova abordagem
  - Essas variáveis relaxam as restrições impostas ao problema de otimização na forma primal

- SVM Linear com Margens Rígidas
  - $y_i(wx_i + b) \geq 1$
- SVM Linear com Margens Suaves
  - $y_i(wx_i + b) \geq 1 - \xi_i$

SVM Linear com Margens Suaves

- Interpretação geométrica
  - As variáveis de folga, ξ_i, medem onde se encontram as amostras em relação as margens de separação
  - Se seu valor for 0, a amostra está fora da região entre estes hiperplanos e é classificada corretamente
  - Se for positivo, mede a distância da amostra em relação aos mesmos
  - Quando o dado é classificado erroneamente, a variável de folga, ξ_i, assume valor maior do que 1

SVM Linear com Margens Suaves



SVM Linear com Margens Suaves

- Problema desta abordagem
  - Não há restrições sobre o número de classificações incorretas
  - O algoritmo tentará a maximizar a margem do limiar de decisão indefinidamente relaxando as restrições o quanto for necessário

SVM Linear com Margens Suaves

- Solução
  - Inserir uma penalidade *C* sobre os relaxamentos
  - C é uma constante que impõe um peso diferente para o treinamento em relação à generalização e deve ser determinada empiricamente

SVM Linear com Margens Suaves

- Temos agora um novo problema de otimização
  - Queremos obter o menor número possível de erros no treinamento e maximizar a margem de separação entre as classes
  - $\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$

- Sujeito a seguinte restrição
  - $y_i(w.x_i + b) \geq 1, i = 1, \ldots n$

SVM Linear com Margens Suaves

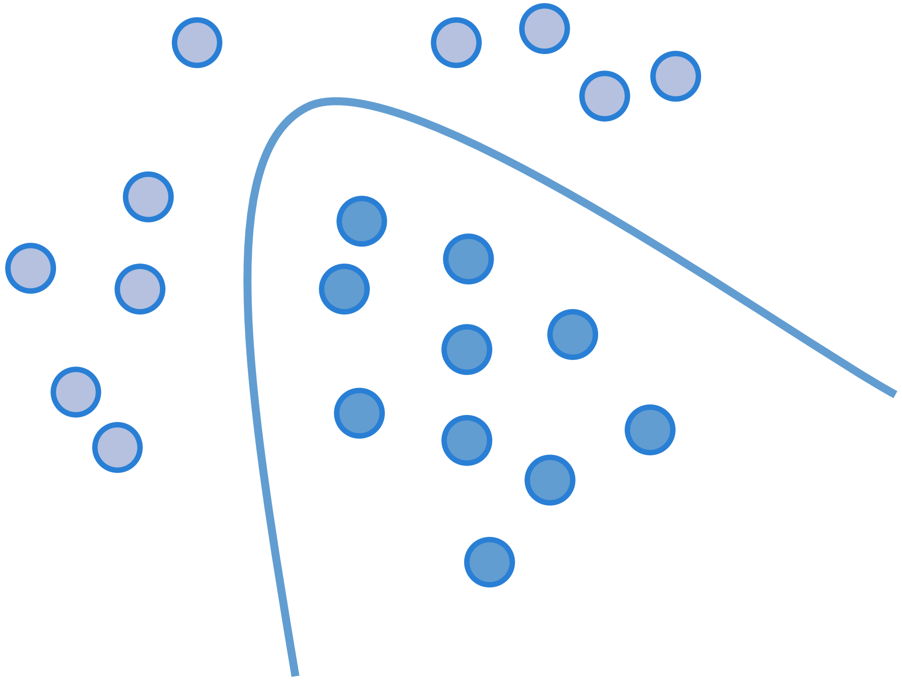
- Problema de otimização
  - Como na SVM linear
    - É um problema quadrático com restrições lineares
    - É um problema convexo
    - Solução global ótima é encontrada usando métodos numéricos
  - Parâmetro *C* pode ser escolhido experimentalmente
    - Com base no desempenho do classificador em dados de validação

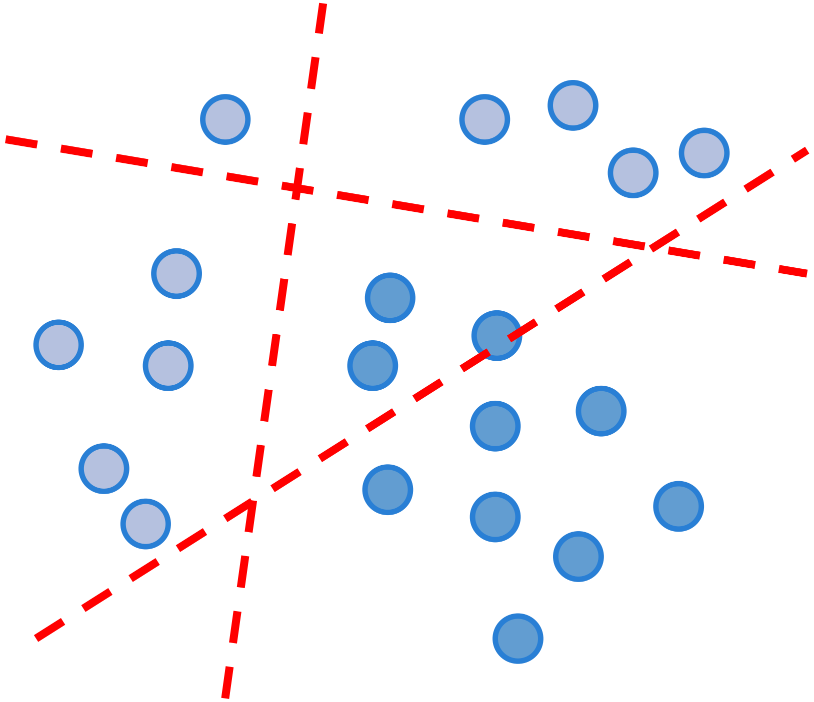
SVM Linear com Margens Suaves

- Problema convexo
  - Implica na otimização de uma função quadrática, que possui apenas um mínimo global
  - Trata-se de um vantagem sobre, por exemplo, as Redes Neurais Artificiais
    - Presença de mínimos locais na função objetivo a ser minimizada

SVM Não Linear

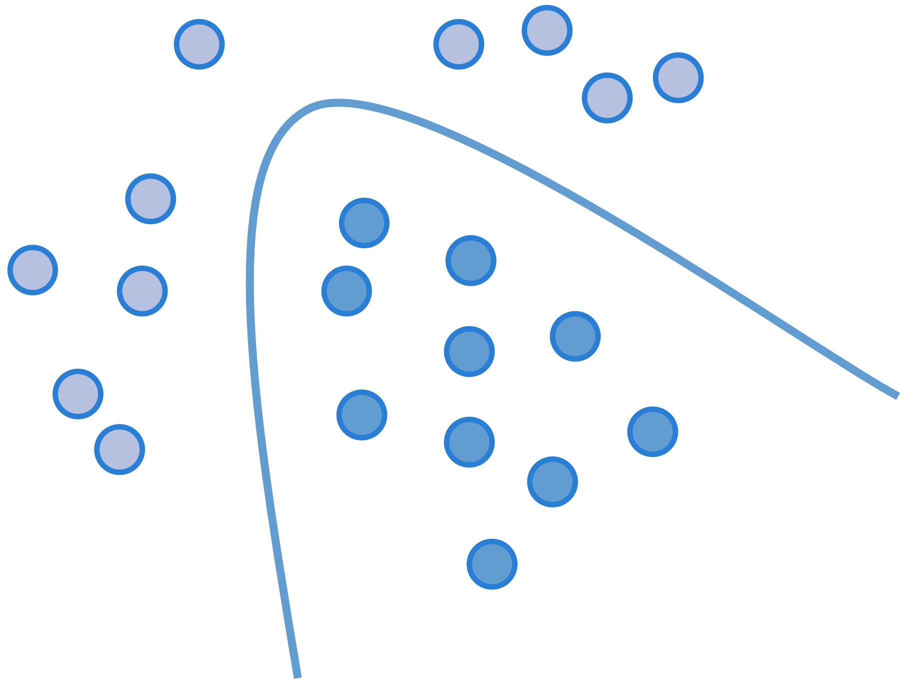
- E se o problema de classificação não for linear?
  - Há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento por um hiperplano





SVM Não Linear

- Solução
  - Mapear o conjunto de treinamento de seu espaço original (não linear) para um novo espaço de maior dimensão, denominado espaço de características (*feature space*), que é linear



SVM Não Linear

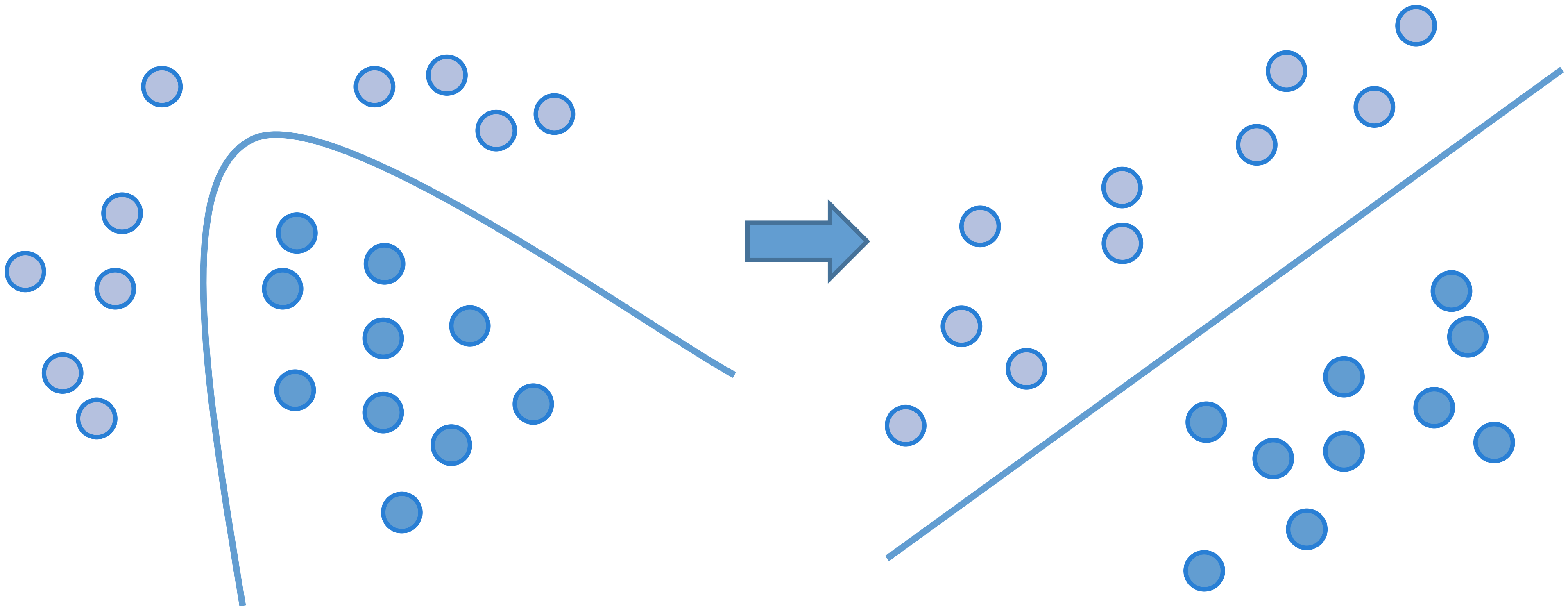
- Para isso, precisamos
  - Encontrar uma transformação não linear
  - $\varphi(x) = [\phi_1(x), \ldots, \phi_m(x)]$
    - Essa transformação mapeia o espaço original dos padrões para um novo espaço de atributos m-dimensional
    - Nesse novo espaço, os padrões ***x*** passam a ser linearmente separáveis
  - ***m*** pode ser muito maior que a dimensão do espaço original

SVM Não Linear

- Exemplo de transformação
  - Dado de entrada (amostra)
    - $\boldsymbol{x = [x_1, x_2]}$
  - Função de transformação
    - ***m = 2 (número de dimensões é igual neste caso)***
    - $\boldsymbol{\phi(x) = [\phi_1(x), \phi_2(x)]}$
    - $\boldsymbol{\phi_1(x) = x_1}$
    - $\boldsymbol{\phi_2(x) = (x_1 + x_2)^4}$

SVM Não Linear

- Exemplo de transformação



SVM Não Linear

- Com a função de transformação, nosso problema de otimização recai pra uma SVM linear
  - $\min_{w,b} \frac{1}{2} \|w\|^2$

- **SVM Linear** é sujeita a seguinte restrição
  - $y_i(w.x_i + b) \geq 1, i = 1, \ldots n$

- **SVM Não Linear** é sujeita a seguinte restrição
  - $y_i(w.\varphi(x_i) + b) \geq 1, i = 1, \ldots n$

SVM Não Linear

- Ou seja, apenas substitui-se x_i *por* $\varphi(x_i)$
  - Mas isso se a transformação for conhecida
- Problema
  - Qual a transformação $\boldsymbol{\phi(x)}$ que torna linearmente separável um determinado conjunto de N padrões $\boldsymbol{x_1, \ldots, x_n}$?
  - Podemos contornar esse problema utilizando uma formulação equivalente do problema de otimização
    - Formulação via multiplicadores de Lagrange

SVM Não Linear

- Formulação via multiplicadores de Lagrange
  - Multiplicadores de Lagrange são muito utilizados em problemas de otimização
    - Permitem encontrar extremos (máximos e mínimos) de uma função de uma ou mais variáveis suscetíveis a uma ou mais restrições
    - É uma ferramenta importante em restrições de igualdade

SVM Não Linear

- Formulação via multiplicadores de Lagrange
  - Solução depende apenas do produto $\varphi(x_i).\varphi(x_j)$ para cada par de padrões $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$, e não dos termos individuais
- Isso é obtido com o uso de funções denominadas *Kernels*
  - $K\left(x_i, x_j\right) = \varphi(x_i).\varphi(x_j)$

SVM Não Linear

- O *Kernel* realiza a transformações de espaço
  - É comum empregar a função *Kernel* sem conhecer o mapeamento φ, que é gerado implicitamente: matriz *Kernel*
    - Nosso objetivo é determinar essa matriz de produtos sem precisar conhecer a transformação φ

  - A utilidade dos *Kernels* está, portanto, na simplicidade de seu cálculo e em sua capacidade de representar espaços abstratos

SVM Não Linear

- Teorema de Mercer
  - Garante que, para algumas classes de *Kernels* $K(x_i, x_j)$, sempre existe uma transformação φ
  - O teorema não garante nada sobre a dimensão ***m*** do espaço transformado φ (pode até ser infinita!)
    - Depende da classe de *Kernels* e dos ***N*** padrões
    - Utilizar *Kernels* pode evitar trabalhar diretamente nesse espaço

SVM Não Linear

- Em termos de Lagrange, a forma Dual da SVM é dada por
  - Minimizar $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$
  - Sujeito a $\begin{cases} \alpha_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$

- E usando um *Kernel*, temos
  - Minimizar $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
  - Sujeito a $\begin{cases} \alpha_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$

SVM Não Linear

- Alguns *Kernels* muito utilizados
  - Polinomial
    - $K(x_i, x_j) = (\delta(x_i . x_j) + k)^d$
  - Gaussianos ou RBF (*Radial-Basis Function*)
    - $K(x_i, x_j) = \exp(-\sigma . \|x_i - x_j\|^2)$
  - Sigmoidal
    - $K(x_i, x_j) = \tanh(\delta(x_i . x_j) + k)$

SVM Não Linear

- Kernel RBF e SVM
  - Quando usamos um kernel RBF em uma SVM, temos que o problema recai exatamente em uma rede neural do tipo RBF
  - Nesse caso, os centros e o número de neurônios da rede são dados automaticamente pelos vetores suporte

SVM Não Linear

- *Overfitting*
  - Maximizar a margem no espaço transformado pelo SVM não-linear não garante a inexistência de *overfitting* no classificador
  - Sempre existe um número de dimensões suficientemente grande que separa s dados de treinamento
    - Exemplo: 1 *Kernel RBF* para cada padrão
    - ***N*** padrões = ***N*** vetores suporte
  - Como controlar o *overfitting*?
    - Técnica de relaxamento já descrita para SVMs lineares

Vantagens e desvantagens

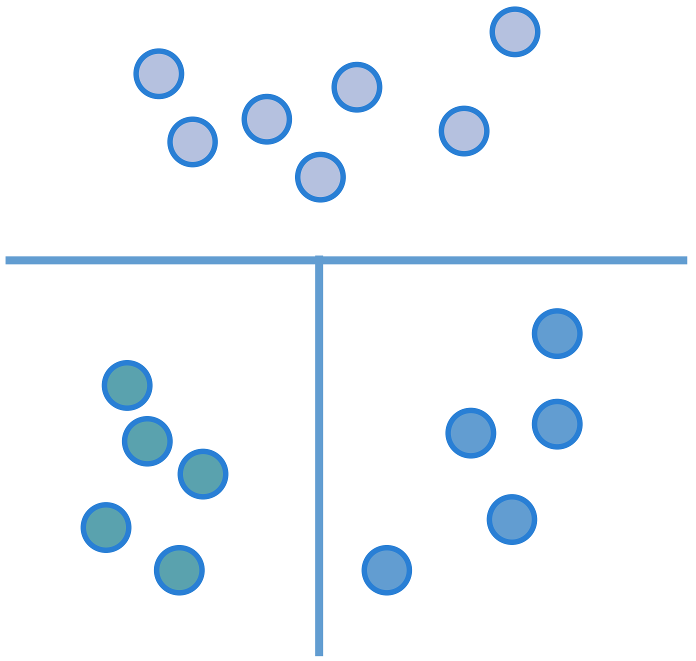
- Vantagens
  - Sempre encontram a melhor solução possível para o problema de otimização em questão
  - Um dos mais eficientes classificadores para problemas de elevada dimensionalidade (muitos atributos)
  - Sua técnica de relaxamento minimiza o risco de *overfitting*
    - Problema crítico em dados com grande dimensionalidade (dados esparsos), e presença de ruído
  - Podem ser adaptados e/ou estendidos para problemas de regressão

Vantagens e desvantagens

- Desvantagens
  - São classificadores do tipo “caixa-preta”, ou seja, não permitem interpretação da estratégia de decisão como as árvores
  - Voltados apenas para atributos numéricos
    - Necessidade de conversão para trabalhar com atributos discretos
  - Possuem complexidade mínima $\boldsymbol{O(N^2)}$, usualmente $\boldsymbol{O(N^3)}$, onde $\boldsymbol{N}$ é o número de padrões de treinamento
    - Se torna crítico a partir de uma certa quantidade de dados de treinamento

SVMs Multi-Classes

- SVMs são classificadores binários
  - Discriminar entre 2 classes possíveis
- O que fazer quando se tem mais de 2 classes de dados?
  - Problema multi-classes
    - padrões de várias classes {1, 2, ..., n}
    - Classes mutuamente excludentes



SVMs Multi-Classes

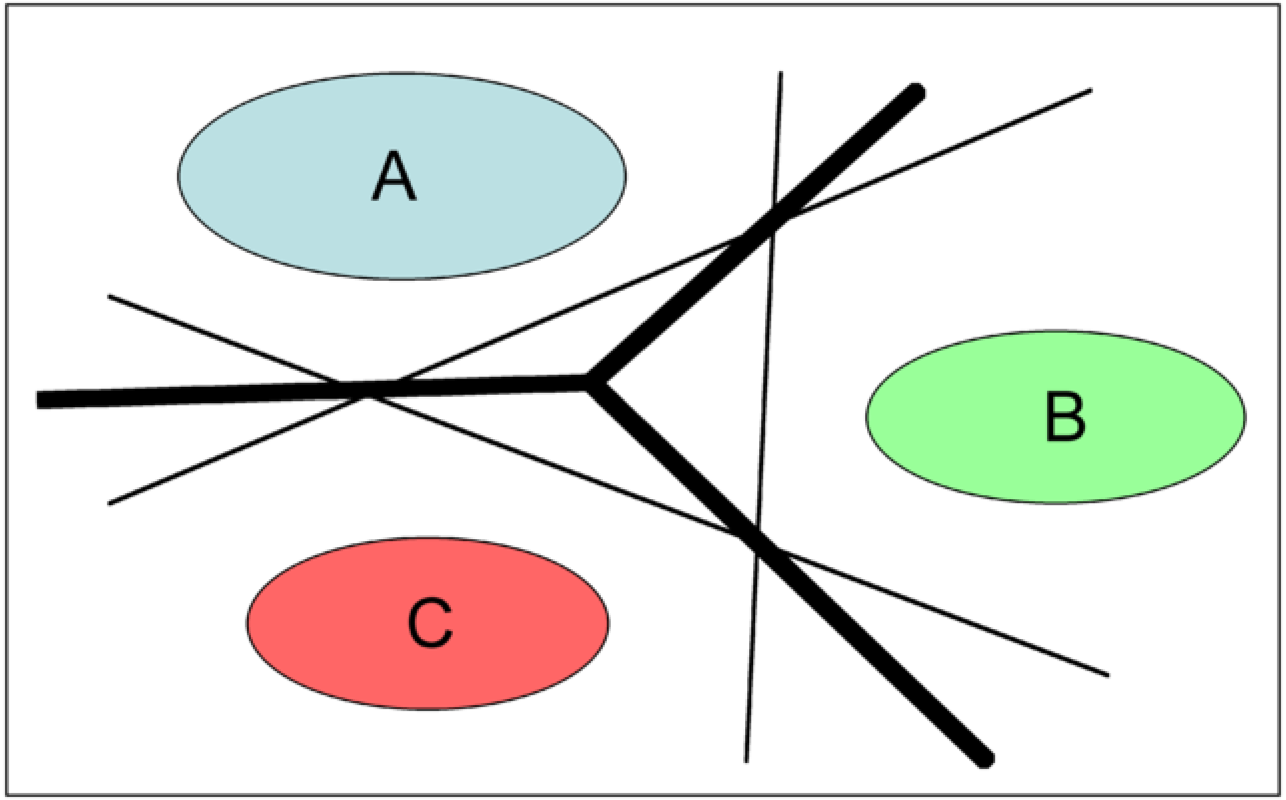
- Nesse caso, precisamos de múltiplos SVMs binários para construir um classificador multi-classes
- Duas alternativas possíveis
  - Decomposição 1-de-n
  - Decomposição 1-1

SVMs Multi-Classes

- Decomposição 1-de-n
  - n classificadores binários
  - Cada classificador identifica uma classe das demais (n-1) classes restantes
  - Essa decomposição simplifica o problema
    - É mais simples distinguir entre 2 classes
    - Empates podem ser resolvidos utilizando alguma medida de confiabilidade das classificações

SVMs Multi-Classes

- Decomposição 1-de-n

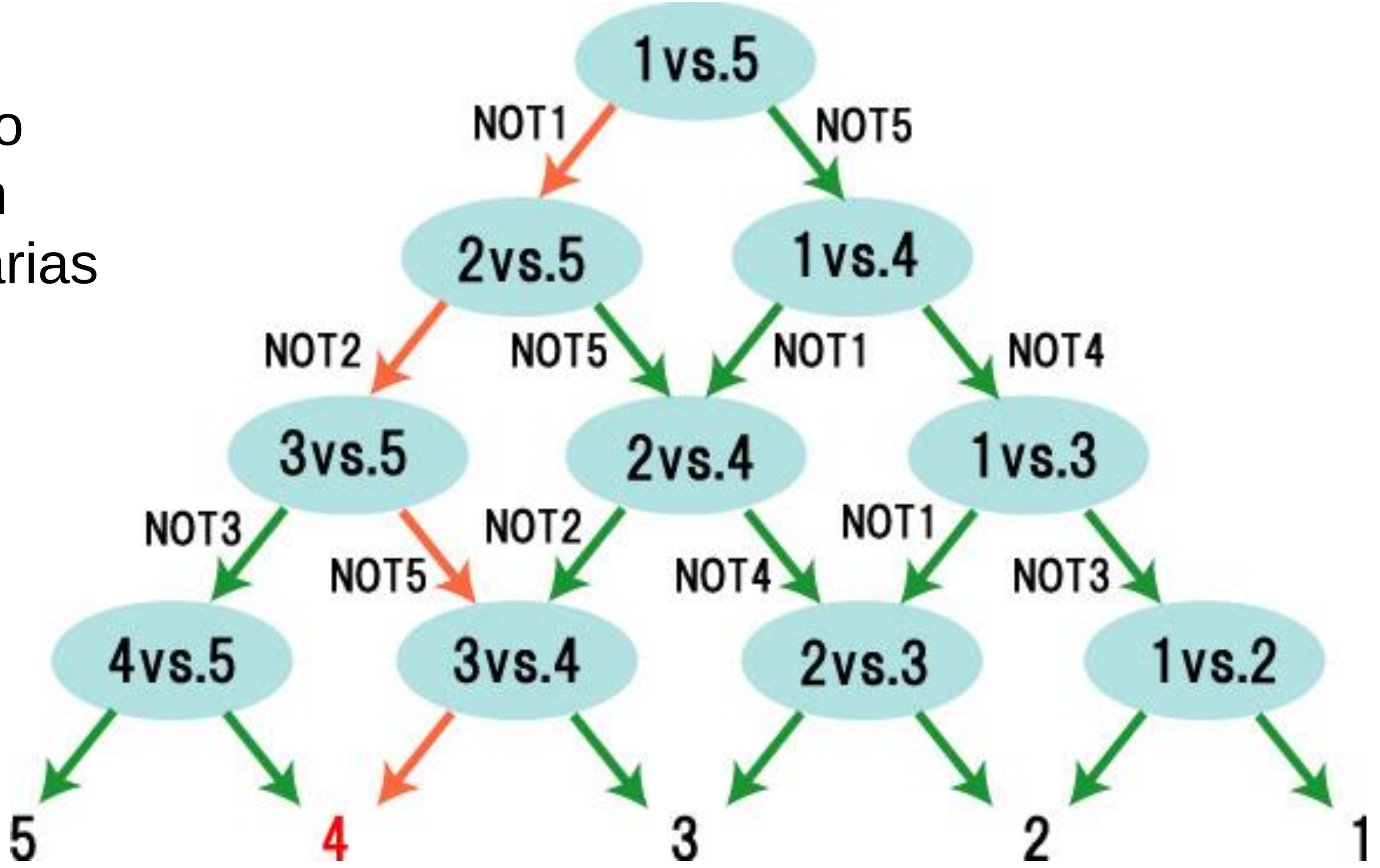


SVMs Multi-Classes

- Decomposição 1-1
  - n*(n – 1)/2 classificadores binários
  - Cada classificador classifica uma amostra dentre um par de classes possíveis
  - No treinamento, padrões que não pertençam as 2 classes envolvidas são ignorados
  - Utiliza mais classificadores que abordagem 1-de-n
  - Classificação
    - Amostra passa por todos os classificadores
    - Classe com maior número de votos é escolhida
    - Menor susceptibilidade a erros

SVMs Multi-Classes

- Decomposição 1-1
  - Grafo direcionado acíclico: o problema é decomposto em diversas classificações binárias em cada nó do grafo



Agradecimentos

- Agradeço ao professor
  - Prof. Ricardo J. G. B. Campello – ICMC/USP
  - Prof. Rodrigo Fernandes de Mello – ICMC/USP
- pelo material disponibilizado