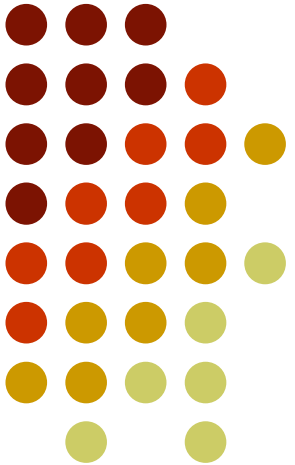
PGCA009 –Inteligência Computacional
Aula 7 SVM

Prof. Angelo Loula

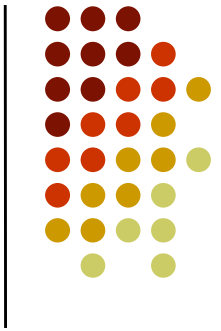
Mestrado em Computação Aplicada (UEFS)



SVM

- Support Vector Machines
- SVMs foram propostas por Vladimir Vapnik e colaboradores, 1992
- Muitos desenvolvimentos depois disso
- Um instância de Kernel Machines (tipo de algoritmo de aprendizado de máquina)
- Muito utilizado em problemas de classificação

Classificação

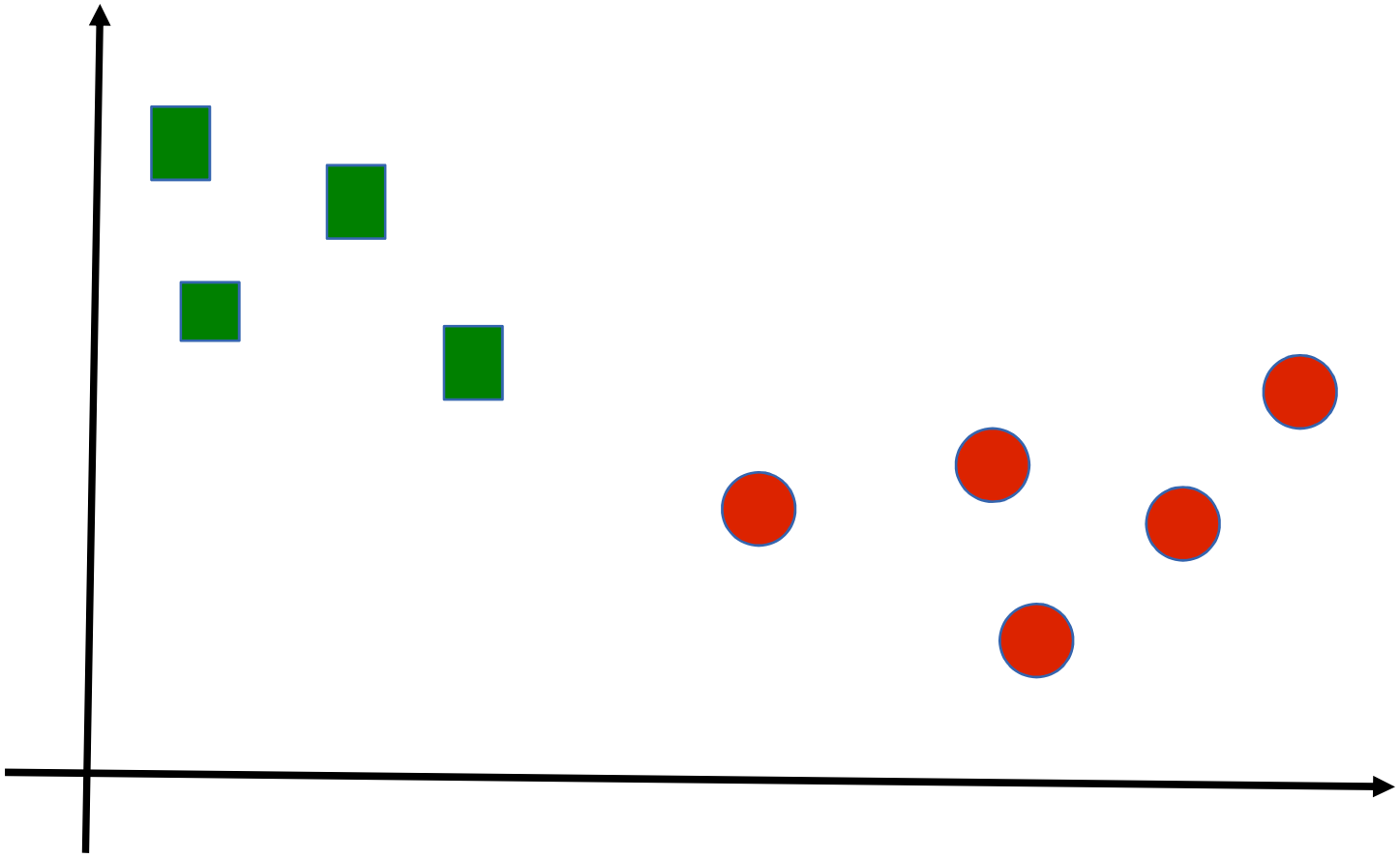


- Problema de Classificação:
- Determinar a classe (categoria) y para uma dada entrada x
- Podemos estimar uma função $f:\Re^n \rightarrow \{0,1\}$ a partir de exemplos

$$(x_1,y_1),\ldots,(x_m,y_m) \in \Re^n \rightarrow \{0,1\}$$

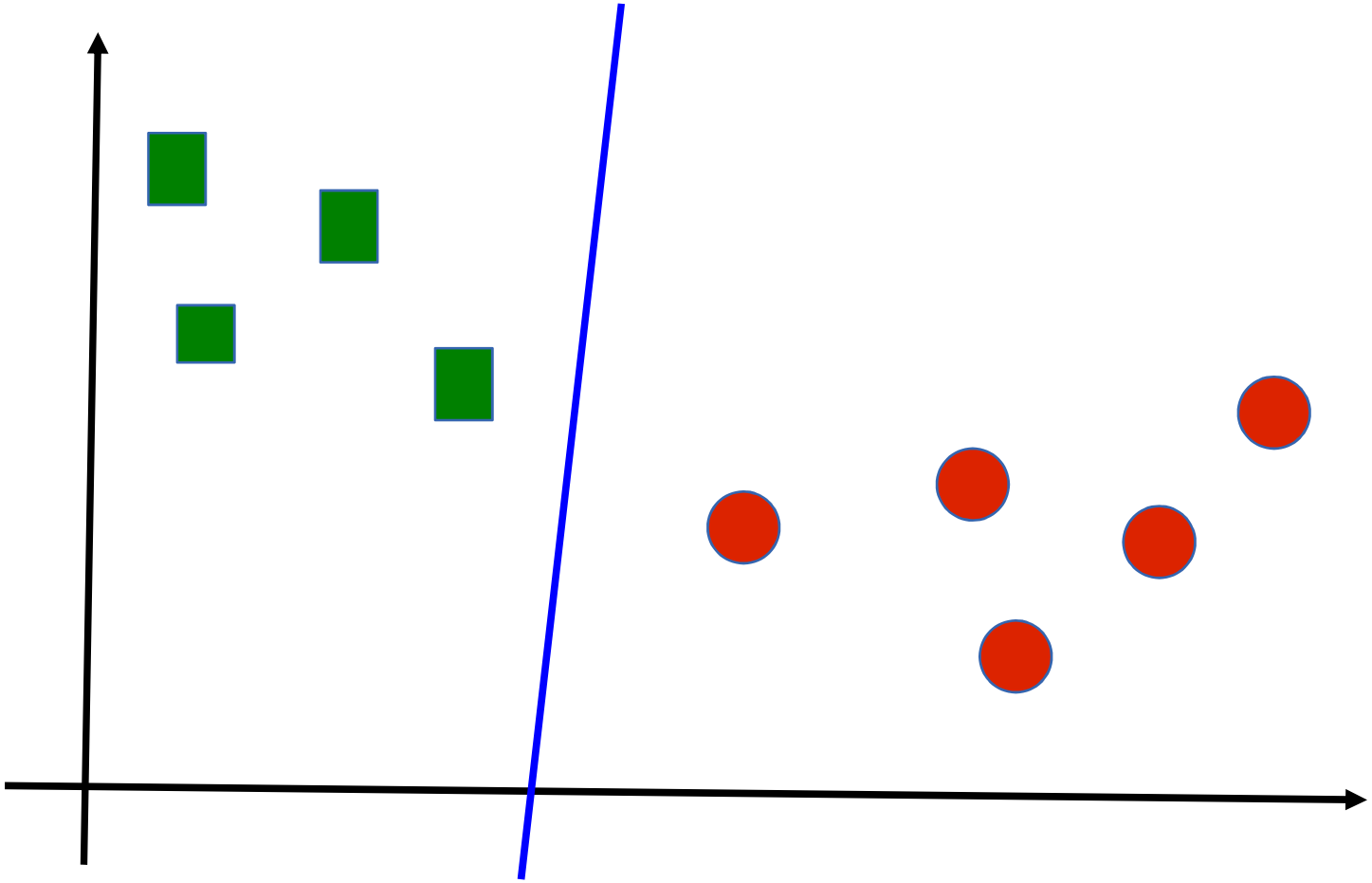
- Mas como generalizar? Como saber o que ocorre entre os exemplos?

Classificação



Como classificar?

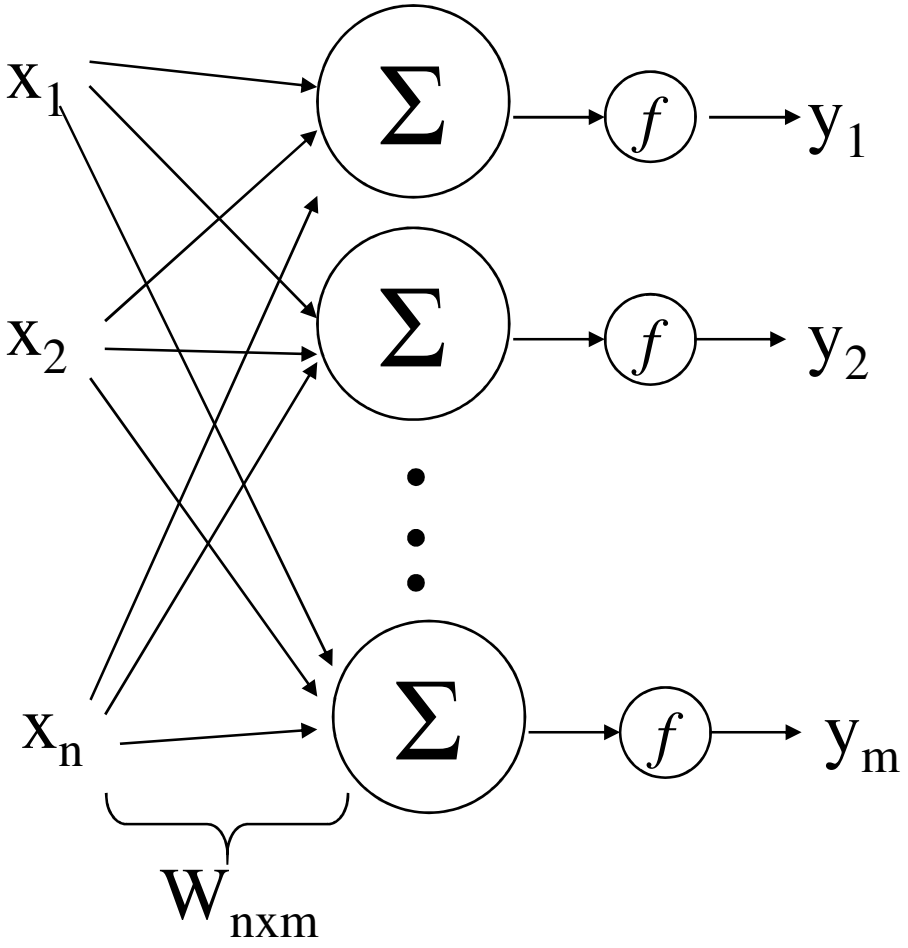
Classificação



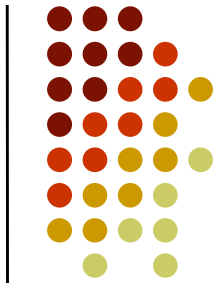
Um classificador linear?

Classificação

- Uma rede perceptron (1 camada) pode realizar o classificador linear

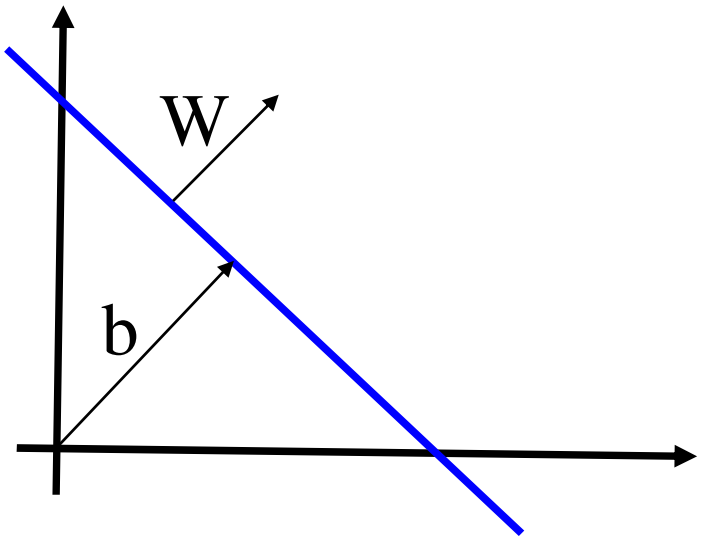


Classificação



- Uma rede perceptron (1 camada) pode realizar o classificador linear

$$X = (x_1, x_2, x_3)$$
$$W = (w_1, w_2, w_3)$$
$$a = X \cdot W + b$$
$$y = \text{sign}(a)$$



Redes Neurais Artificiais

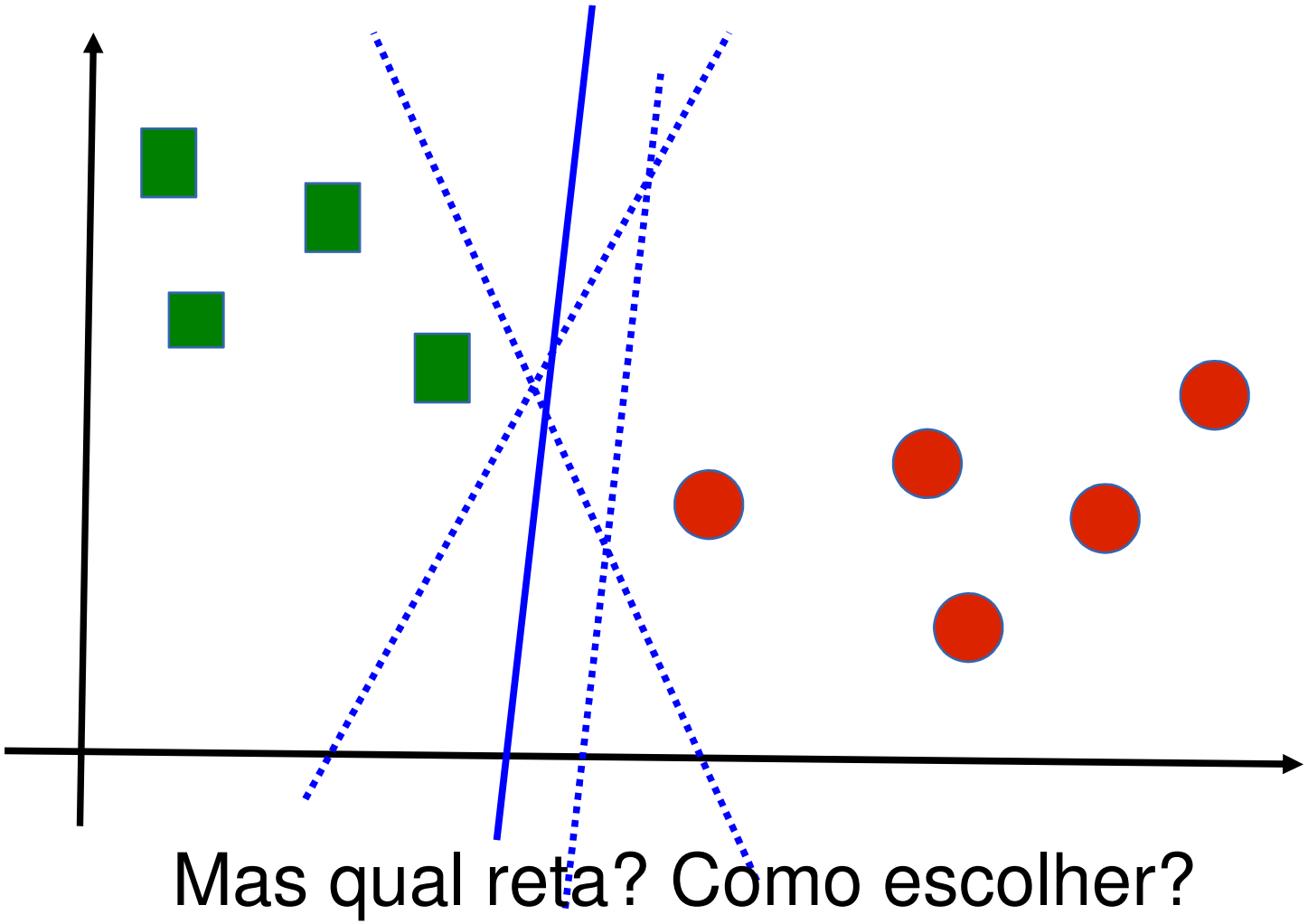
- Regra de Aprendizado Perceptron
  - Para neurônio com função de transferência hardlim, o ajuste será dado por:

$$W_{novo} = W_{anterior} + \rho \cdot erro \cdot X$$

repete-se até critério de parada (por exemplo, erro = 0)

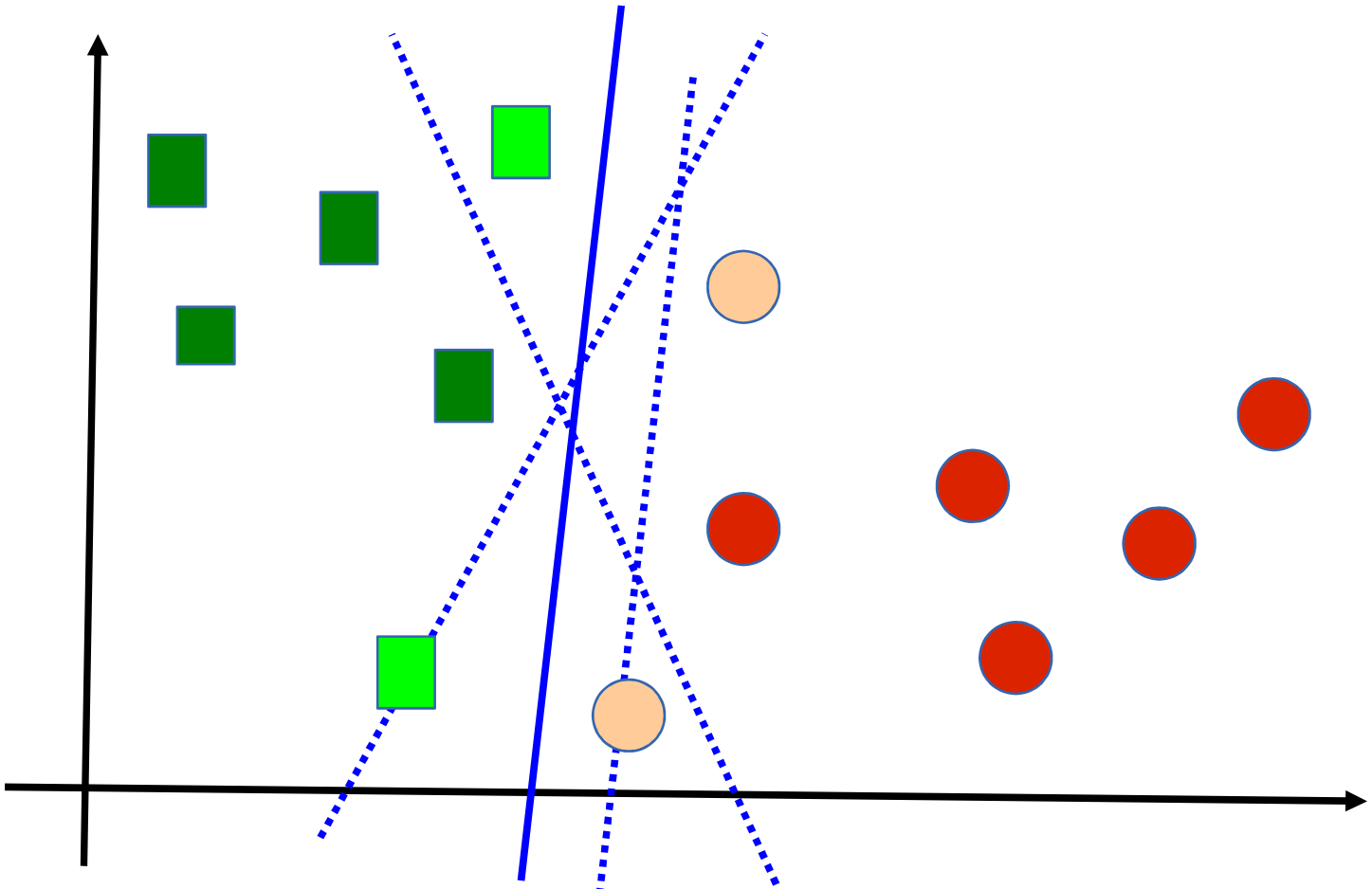
  - Mas onde vai ficar a reta de decisão?

Classificação



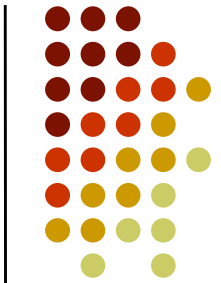
Mas qual reta? Como escolher?

Classificação



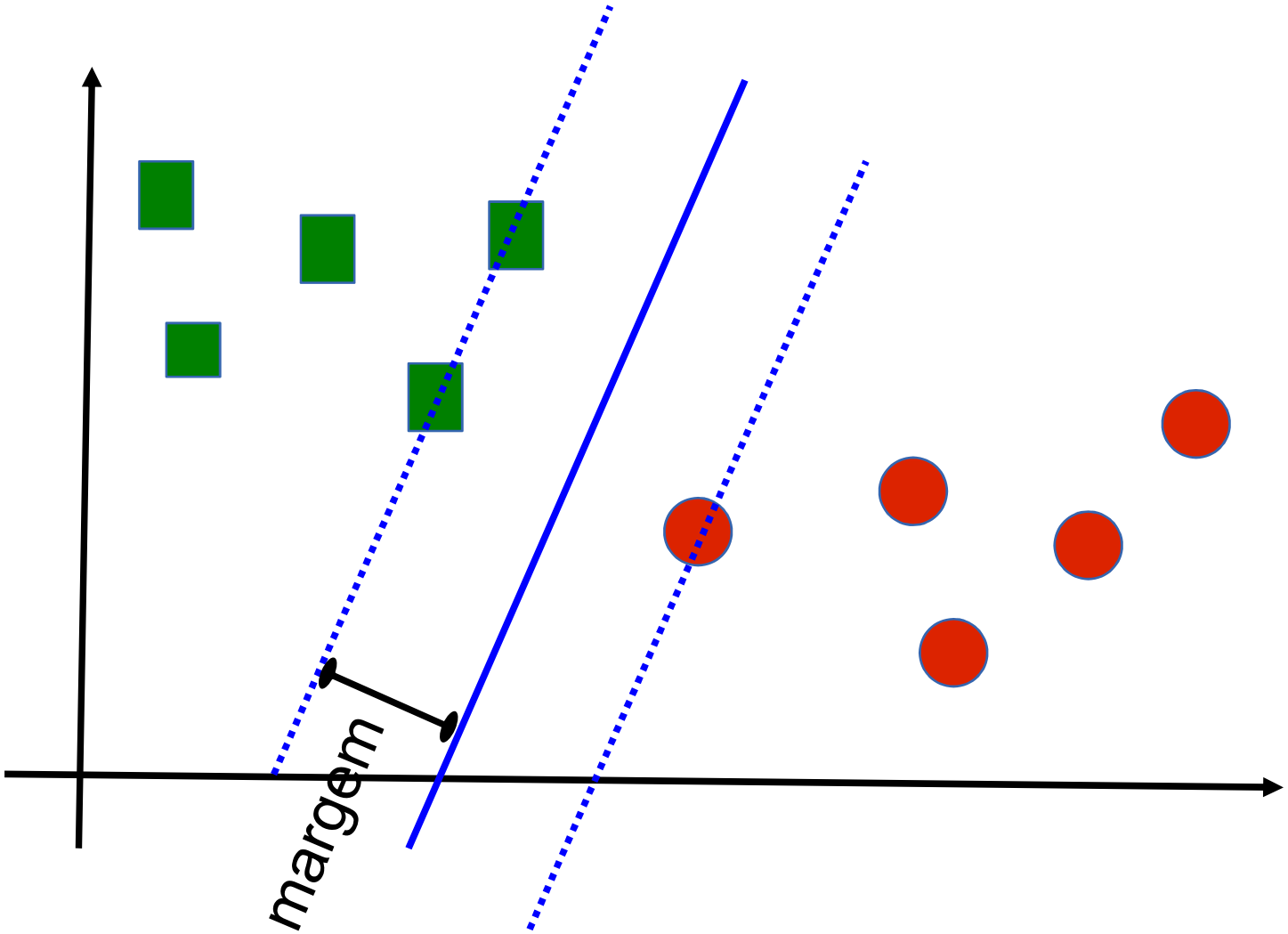
Qual o melhor para generalizar para dados desconhecidos?

SVM

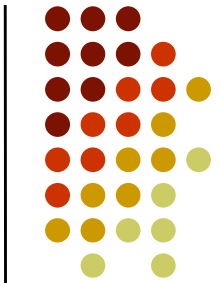


- Intuitivamente, esperamos que uma reta que passe ao meio da região sem pontos seja melhor
- SVM define um critério para definir a posição da fronteira de decisão
  - A distância do reta(plano) de decisão para o ponto mais próximo determina a margem do classificador.
  - A margem tem que ser a maior possível!
  - Então somente alguns pontos definem o plano: os 'support vectors'

Classificação



SVM



- Maximizar a margem permite que o plano de decisão esteja longe dos dados, aumentando a robustez e generalização.
- Um pequeno erro de medida/ruído pode ser tolerado sem falha de classificação.
- Essa margem também reduz a capacidade do modelo, diminuindo as possibilidades de posição.

SVM

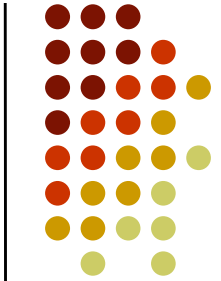
- A reta/plano/hiperplano que define a fronteira de decisão do SVM pode ser definida como:

  $W \cdot (X - X_0) = 0$, ou $X \cdot W + b = 0$
  W: vetor normal, perpendicular ao hiperplano
  b: deslocamento do hiperplano em relação a origem

- Para SVM, as classes são sempre -1 e +1

  $y = \text{sign}(X \cdot W + b)$
  ou seja, $y = +1$ se $X \cdot W + b >= 0$ e $y = -1$ se $X \cdot W + b < 0$

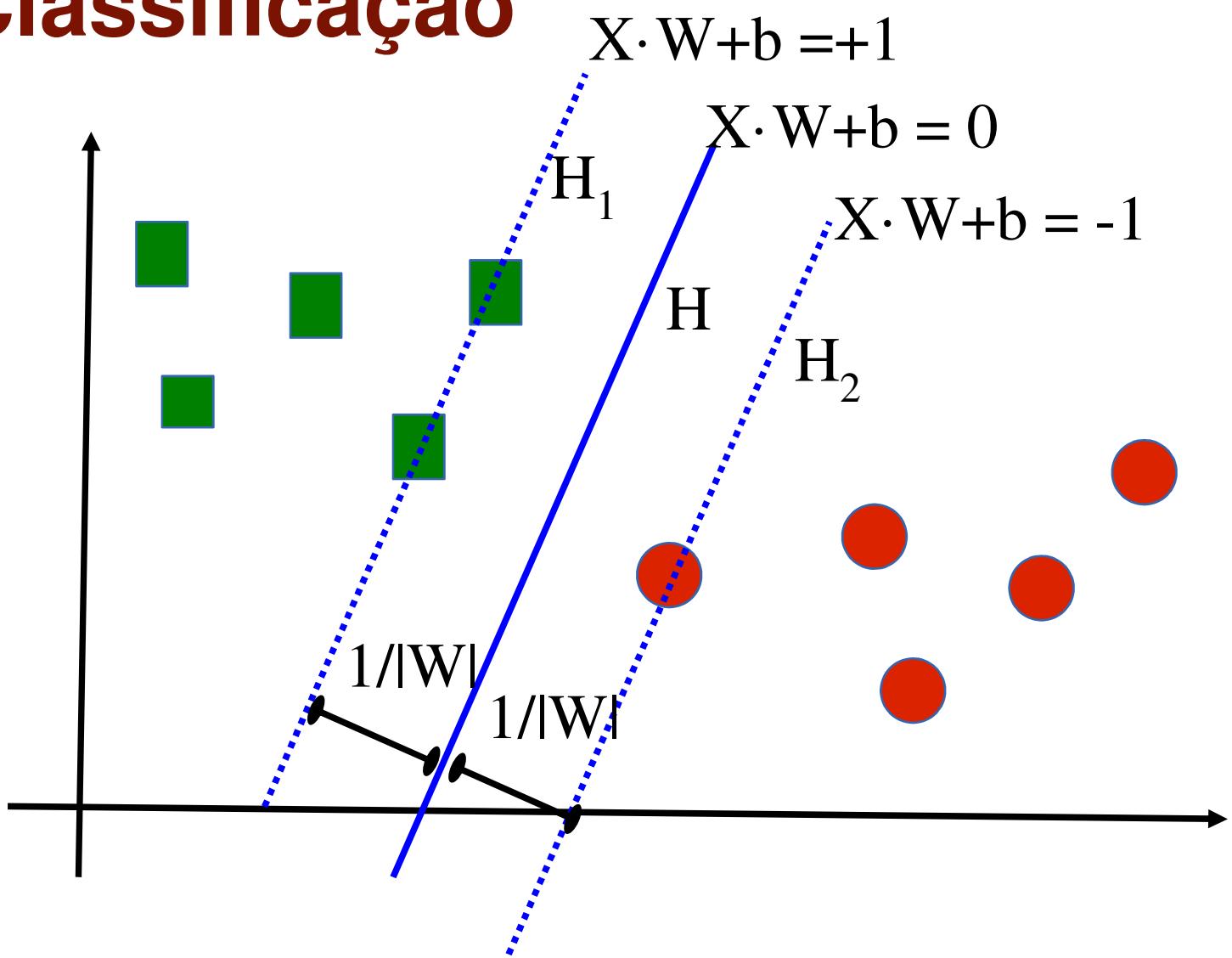
SVM



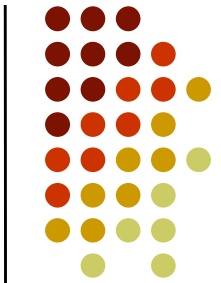
- Considerando uma margem para posicionar o hiperplano equidistante dos pontos mais próximos das classes diferentes, para cada exemplo (X_i, y_i) fornecido:

$$\text{se } y_i = +1, \text{ então } X_i \cdot W + b >= +1$$
$$\text{se } y_i = -1, \text{ então } X_i \cdot W + b <= -1$$
$$\text{ou ainda,}$$
$$y_i(X_i \cdot W + b) >= +1$$
$$y_i(X_i \cdot W + b) - 1 >= 0$$

Classificação



SVM



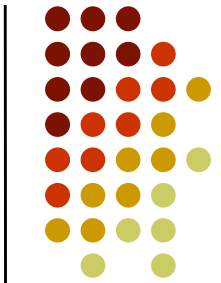
- A distância do hiperplano H até a origem é

$$D = \frac{|b|}{\|W\|}$$

- A distância dos hiperplanos H1 e H2 até a origem são

$$D_1 = \frac{|b-1|}{\|W\|} \qquad D_2 = \frac{|b+1|}{\|W\|}$$

SVM



- A margem é a distância entre os planos H1 e H2, dada por

$$\frac{2}{\| W \|}$$

- A distância do hiperplano H1 para H ou de H2 para H é dado por

$$\frac{1}{\| W \|}$$

SVM

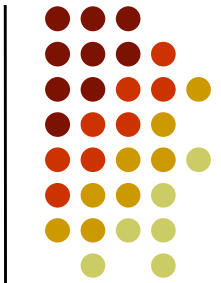
- Precisamos encontrar então W e b em

$$y_i(X_i \cdot W + b) >= +1$$

- Tal que seja maximizado o valor de $\frac{2}{\| W \|}$

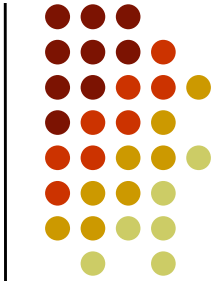
- Ou, de forma equivalente, minimizando ||W||, ou ainda, minimizando $||W||^2$
- (é importante normalizar/escalonar os dados nas diversas dimensões!)

SVM



- SVM constroi um hiperplano que define uma fronteira de decisão maximizando distâncias para os pontos de exemplo mais próximos da fronteira de decisão.
- Segue aprendizado supervisionado utilizando exemplos de treinamento, pares de vetores X_i associados com classes (saídas) y_i
- Somente os 'support vectors' serão realmente utilizados para cálculo do hiperplano.

SVM



- Para construir o classificador SVM:

$$\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{sujeito a} \quad y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1.$$

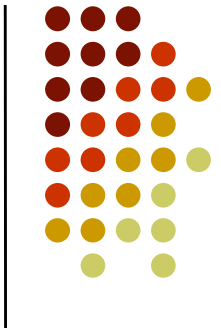
que pode ser re-escrito como (usando multiplicadores de Lagrange)

$$\arg\min_{\mathbf{w},b}\max_{\boldsymbol{\alpha}\geq 0}\left\{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x_i} - b) - 1]\right\}$$

a solução pode ser escrita na forma $\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i}.$

Mas somente alguns α_i serão diferentes de zero, aqueles associados com x_i de pontos de support vectors

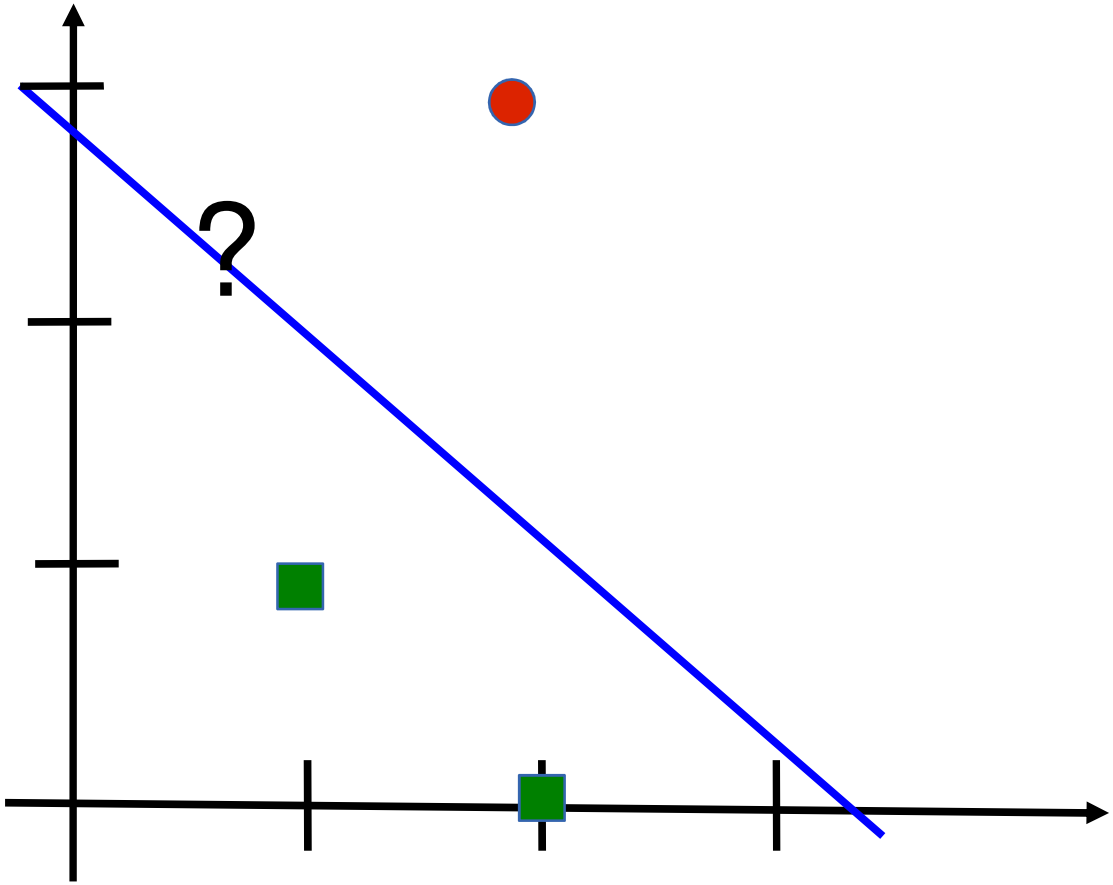
SVM



- Para solucionar o problema de encontrar os parâmetros do SVM, é preciso resolver um problema de programação quadrática.
- Trata-se de problema conhecido de otimização, existem vários algoritmos para resolver.
- E existem alguns especializados no problema de otimização do SVM.
- Usar bibliotecas especializadas!

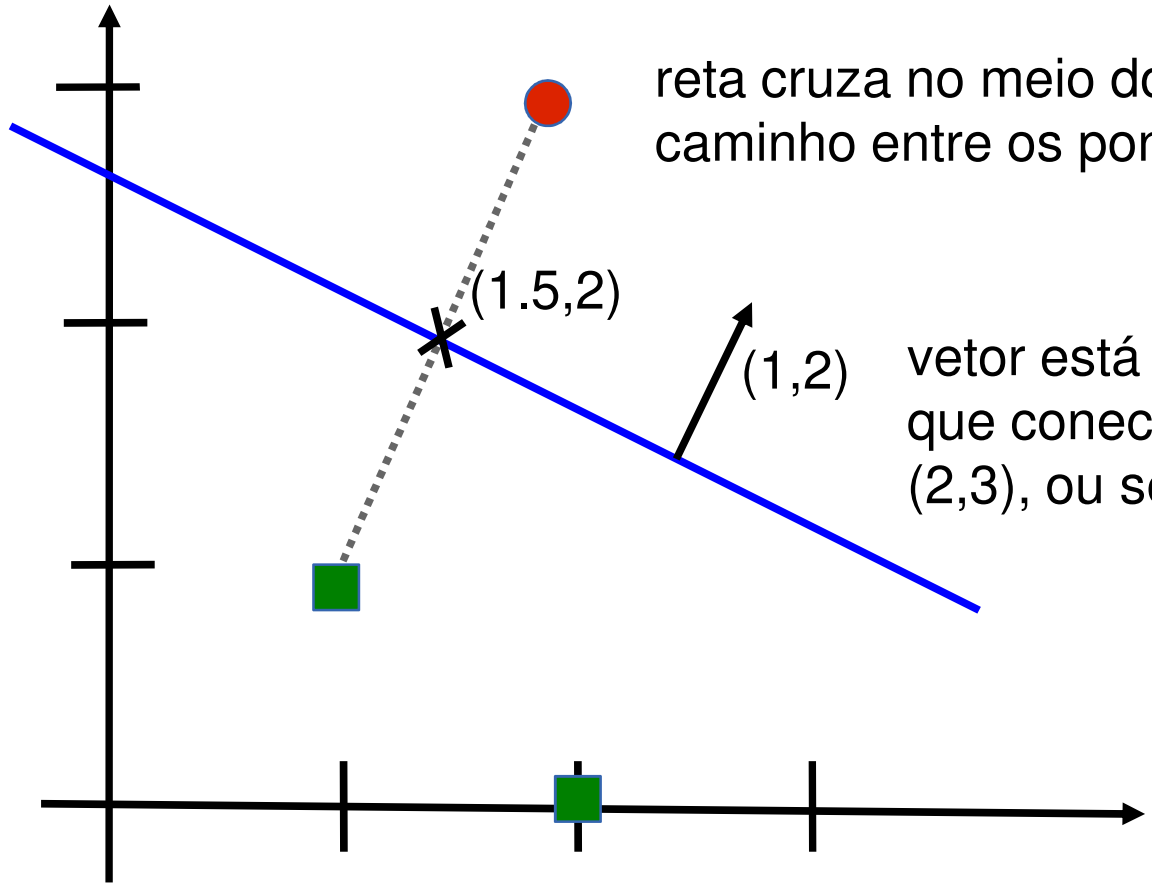
SVM

- Exemplo:



SVM

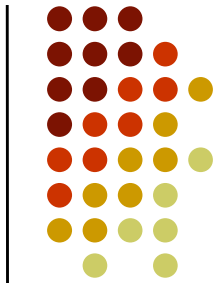
- Exemplo:



reta cruza no meio do caminho entre os pontos

vetor está na direção da linha que conecta os pontos (1,1) e (2,3), ou seja, (2,3)-(1,2)

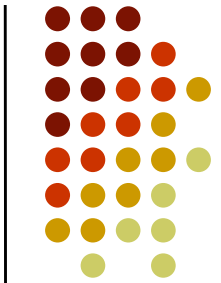
SVM



- Temos então que W = (1,2) e X_0=(1.5,2), então

$$W \cdot (X - X_0) = 0$$
$$(1,2) \cdot [(x_1,x_2) - (1.5,2)] = 0$$
$$x_1 + 2x_2 - 5.5 = 0$$

SVM



- Podemos também fazer a resolução algébrica:

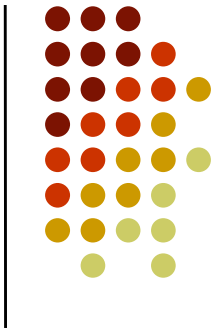
sabendo que $y_i(X_i \cdot W+b) >= +1$ alçanca a equalidade para (1,1) e (2,3), temos que

$(1,1) \cdot W+b = -1$

$(2,3) \cdot W+b = +1$

como W está na direção da reta entre (1,1) e (2,3), assim W será (a,2a) para algum a

SVM



$(1,1)\cdot(a,2a)+b = -1$; $(2,3)\cdot(a,2a)+b = +1$

$a+2a+b = -1$; $2a+6a+b = +1$

$a = 2/5$

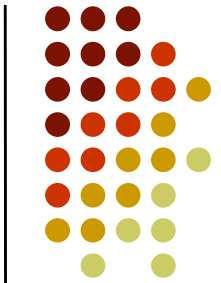
$b = -11/5$

$W = (2/5, 4/5)$

$\text{margem} = 2/\|W\| = \sqrt{5}$

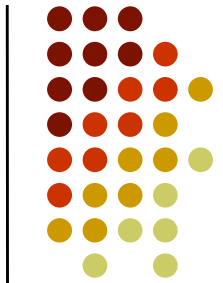
(obs.:porque W ficou diferente em relação à solução geométrica?)

SVM



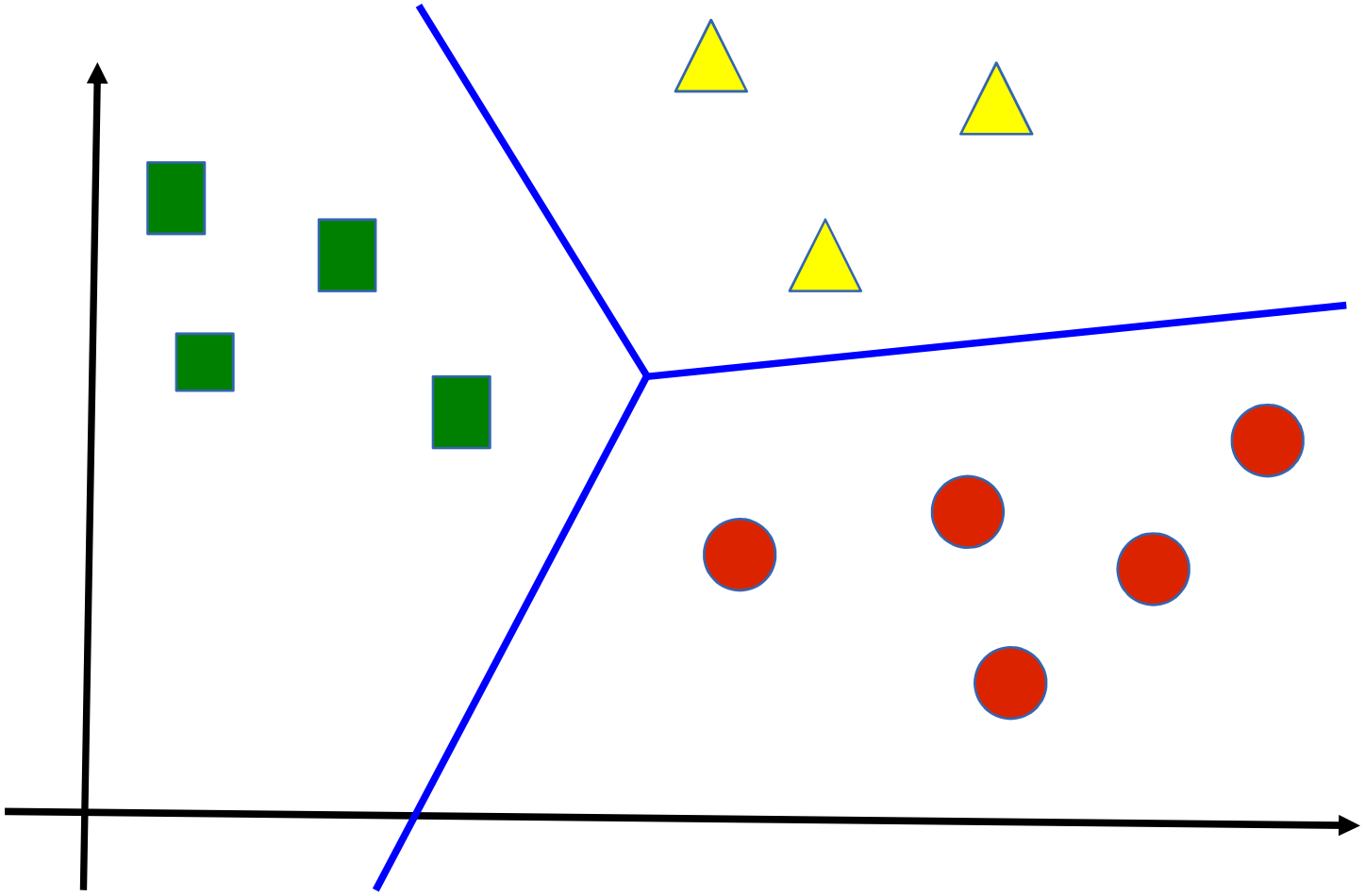
- E se houver exemplos inconsistentes ou não são linearmente separáveis?
- Soft-margin SVM
- Acrescenta uma tolerância a falhas de classificação $y_i(X_i \cdot W+b) >= 1 - \xi$
- Temos então uma otimização que precisa maximizar a margem $2/\|W\|$ e minimizar ξ

SVM

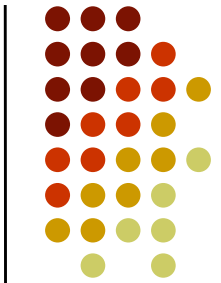


- E se tivermos mais do que 2 classes?
- Multi-class SVM
- No SVM só podemos tratar de 2 classes, para mais classes crie vários classificadores do tipo um-versus-todos para cada classe
- Escolha como classe para cada entrada nova X_k, a classe com maior saída dentres os classificadores

Classificação



SVM



- Mas e se os exemplos não são linearmente separáveis?
- Mapeamos os pontos em um espaço de dimensão muito mais alta por uma transformação Φ, usando um 'kernel trick'.
- Neste novo espaço, os exemplos podem ser linearmente separáveis.
- São usadas funções de kernel que garantam computar o produto interno facilmente.

SVM

- SVM não-linear
- O 'truque' está em escolher uma função de kernel $K(x_i,x_j)$ que substitua o produto interno $x_i \cdot x_j$, transformando x_i em $\Phi(x_i)$ e assim $x_i \cdot x_j$ torna-se $\Phi(x_i) \cdot \Phi(x_j)$
- Mas não seja necessário achar cada $\Phi(x_i)$ mas somente computar $K(x_i,x_j)$ pois é a única operação afetada

SVM

- SVM não-linear
- antes $y = \text{sign}(\sum_i \alpha_i y_i (X_i \cdot X) + b)$
- Com kernel $y = \text{sign}(\sum_i \alpha_i y_i K(X_i, X) + b)$
- funções de kernel mais comuns:
  - RBF (ex. Gaussiana)

    formato radial de features
  - polinominais

    conjunções de features em pares (quadrático)
    ou em triplas (cúbico)