Universidade Federal do Ceará
Departamento de Estatística e Matemática Aplicada
Introdução à Mineração de Dados (CC0101), período 2019/2
Professor: Tibérius O. Bonates (tb@ufc.br).

Roteiro de Aula sobre Máquinas de Vetores Suportes (Support Vector Machines (SVM))
(Estas não são notas de aula. Por gentileza, não distribua este material. Ele é para uso exclusivo na disciplina identificada neste documento.)

Este material é fracamente baseado nos livros *Learning with Kernels* e *Kernel Methods for Pattern Analysis*.

Assumimos que o conjunto de treinamento $S = \{(\mathbf{x}^i, y^i) : i = 1, \ldots, m\} \subset \mathbb{R}^n \times \{-1, +1\}$ possui duas classes ($y^i \in \{+1, -1\}$, para todo $i = 1, \ldots, m$) e que existe pelo menos um elemento de cada classe em S.

1 Hiperplano Separador e Margem

O classificador linear que conhecemos é dado por:

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0,$$

com $\mathbf{w} \in \mathbb{R}^n$ e $w_0 \in \mathbb{R}$.

Desejamos encontrar $(\mathbf{w}, w_0)$ tal que

$$h(\mathbf{x}^i) \geq 0, \qquad \text{se } y^i = +1 \tag{1}$$
$$h(\mathbf{x}^i) \leq 0, \qquad \text{se } y^i = -1, \tag{2}$$

para todo i. Por enquanto, assumiremos que os dados sao linearmente separáveis, o que significa que as desigualdades podem ser estritas. Como o valor de $h(\mathbf{x})$ pode ser tornado arbitrariamente grande em valor absoluto com o ajuste de $\mathbf{w}$, iremos utilizar a noção de *hiperplano separador canônico* relativo a S, que consiste em um hiperplano $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0 = 0$, satisfazendo (1) e (2), mas também satisfazendo

$$\min\{|h(\mathbf{x}^i)| : i = 1, \ldots, m\} = 1.$$

Assim, podemos substituir as condições (1) e (2) por

$$h(\mathbf{x}^i) \geq 1, \qquad \text{se } y^i = +1 \tag{3}$$
$$h(\mathbf{x}^i) \leq -1, \qquad \text{se } y^i = -1, \tag{4}$$

para todo i. (3) e (4) podem ser reescritas compactamente como

$$y^i h(\mathbf{x}^i) \geq 1, \forall i.$$

Como $\mathbf{w}$ é um vetor normal ao hiperplano $h(\mathbf{x}) = 0$, podemos reescrever qualquer vetor $\hat{\mathbf{x}} \in \mathbb{R}^n$, com $h(\hat{\mathbf{x}}) > 0$ como $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{p} + r \cdot \dfrac{\mathbf{w}}{\|\mathbf{w}\|}$, com $\hat{\mathbf{x}}_\mathbf{p}$ sendo a projeção de $\hat{\mathbf{x}}$ sobre o hiperplano $h(\mathbf{x}) = 0$, com r sendo a distância entre $\hat{\mathbf{x}}$ e $h(\mathbf{x}) = 0$, e $\dfrac{\mathbf{w}}{\|\mathbf{w}\|}$ sendo o vetor unitário correspondente à direção de $\mathbf{w}$.

Aplicando h a $\hat{\mathbf{x}}$, e levando em conta que $h(\hat{\mathbf{x}}_\mathbf{p}) = 0$ (por definição, $\hat{\mathbf{x}}_\mathbf{p}$ reside exatamente sobre

o hiperplano), obtemos

$$\begin{aligned} h(\hat{\mathbf{x}}) = h\left(\hat{\mathbf{x}}_\mathbf{p} + r \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|}\right) &= \mathbf{w}^\top \left(\hat{\mathbf{x}}_\mathbf{p} + r \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= \mathbf{w}^\top \hat{\mathbf{x}}_\mathbf{p} + w_0 + r \cdot \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|} \\ &= r \cdot \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} = r \cdot \|\mathbf{w}\|, \end{aligned}$$

que nos permite escrever $r = \dfrac{h(\hat{\mathbf{x}})}{\|\mathbf{w}\|}$.

Se tivéssemos tomado $\hat{\mathbf{x}}$ com $h(\hat{\mathbf{x}}) < 0$, teríamos obtido $r = -\dfrac{h(\hat{\mathbf{x}})}{\|\mathbf{w}\|}$.

O valor de r está associado à ideia de *margem do separador* $h(\mathbf{x}) = 0$, que é definida como duas vezes a menor distância entre um ponto do conjunto de treinamento e o hiperplano $h(\mathbf{x}) = 0$, correspondendo a uma faixa delimitada por dois hiperplanos, $h(\mathbf{x}) = r$ e $h(\mathbf{x}) = -r$, em torno do hiperplano separador propriamente dito. Dentro desta faixa não existe nenhum ponto do conjunto de treinamento. Note que a largura de cada banda desta faixa é $r = \dfrac{1}{\|\mathbf{w}\|}$.

2 O Problema de Otimização Associado

A capacidade de generalização (classificação de observações do conjunto de teste) está associada à largura da margem de separação, ou seja, ao valor $\dfrac{2}{\|\mathbf{w}\|}$. Uma maneira de perceber essa relação é levar em conta que, se o conjunto de teste segue a mesma distribuição do conjunto de treinamento, então podemos esperar que observações do conjunto de teste que sejam próximas a uma observação do conjunto de treinamento pertençam à mesma classe desta. Em particular, se para toda observação de teste a distância entre ela e uma observação do conjunto de treinamento, de mesma classe, for limitada superiormente por um valor τ, então um hiperplano que separe corretamente o conjunto de treinamento, com margem $r > \tau$, classificará corretamente todas as observações de teste. Embora a situação descrita acima seja muito peculiar, a intuição que se deriva dela é válida como uma das motivações para se preferir hiperplanos com margem grande (ou larga).

Uma vez que estamos interessados em encontrar um hiperplano de margem maior possível, podemos tentar resolver o seguinte problema de otimização:

$$\max_{\mathbf{w}, w_0} \quad \frac{1}{\|\mathbf{w}\|} \tag{5}$$

$$\text{sujeito a} \quad y^i \left(\mathbf{w}^\top \mathbf{x}^i + w_0\right) \geq 1, \forall i. \tag{6}$$

Este é um problema de difícil solução, em particular devido à função objetivo. No entanto, maximizar $\dfrac{1}{\|\mathbf{w}\|}$ equivale a minimizar $\|\mathbf{w}\|$, que, por sua vez, equivale a minimizar $\|\mathbf{w}\|^2$. Assim, podemos optar por resolver o problema quadrático

$$(\mathrm{P}) \min_{\mathbf{w}, w_0} \quad \frac{1}{2} \cdot \|\mathbf{w}\|^2 \tag{7}$$

$$s.a \quad y^i \left(\mathbf{w}^\top \mathbf{x}^i + w_0\right) \geq 1, \forall i. \tag{8}$$

(P) é estritamente convexo e, portanto, possui mínimo único. (P) é o problema *primal* do método de SVM, em contraposição à versão *dual* deste problema, que veremos adiante e que é a versão tipicamente utilizada em implementações computacionais do método.

Uma maneira alternativa de lidar com o problema (P) é incorporar as restrições (8) na própria função. Isso é feito por meio de uma penalização das restrições (8) com o uso de novas variáveis $\alpha_i \geq 0,\ \forall i$:

$$L(\mathbf{w}, w_0, \alpha) = \frac{1}{2} \cdot \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y^i(\mathbf{w}^\top \mathbf{x}^i + w_0) - 1 \right). \tag{9}$$

Perceba que, para valores de $\mathbf{w}$ e w_0 que satisfaçam as restrições (8), temos que a melhor escolha de valores para as variáveis α_i é $\alpha_i = 0,\ \forall i$. A função (9) deve ser minimizada com relação a $\mathbf{w}$ e w_0 e maximizada com relação às variáveis α_i. No intuito de encontrar valores que realizem essa tarefa, temos

$$\frac{\partial L}{\partial w_0} = 0 \ \Rightarrow \sum_{i=1}^{m} \alpha_i y^i = 0 \tag{10}$$

$$\frac{\partial L}{\partial w_j} = 0 \ \Rightarrow w_j = \sum_{i=1}^{m} \alpha_i y^i x_j^i$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \ \Rightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y^i \mathbf{x}^i. \tag{11}$$

Perceba que agora temos, por meio de (11), uma definição de $\mathbf{w}$ em função das observações $\mathbf{x}^i$. Uma observação $\mathbf{x}^i$ cujo coeficiente correspondente α_i é não-nulo nesta expressão é um dos *vetores suportes* do hiperplano.

Substituindo em (9), temos

$$\begin{aligned} L(\mathbf{w}, w_0, \alpha) &= \frac{1}{2} \cdot \left(\sum_{i=1}^{m} \alpha_i y^i \mathbf{x}^i \right)^\top \left(\sum_{j=1}^{m} \alpha_j y^j \mathbf{x}^j \right) - \sum_{i=1}^{m} \alpha_i \left[y^i \left(\sum_{j=1}^{m} \alpha_j y^j {\mathbf{x}^j}^\top \mathbf{x}^i + w_0 \right) - 1 \right] \\ &= -\frac{1}{2} \cdot \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i y^i \alpha_j y^j {\mathbf{x}^i}^\top \mathbf{x}^j \right) - w_0 \sum_{i=1}^{m} \alpha_i y^i + \sum_{i=1}^{m} \alpha_i \\ &= -\frac{1}{2} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle + \sum_{i=1}^{m} \alpha_i. \end{aligned}$$

Isso nos dá uma formulação dual do problema (P), que, embora também seja um problema de otimização quadrática, é mais conveniente de resolver do que (P):

$$\text{(Q)} \quad \max \qquad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle \tag{12}$$

$$s.a \qquad \sum_{i=1}^{m} \alpha_i y^i = 0 \tag{13}$$

$$\alpha_i \geq 0,\ i = 1, \ldots, m. \tag{14}$$

De fato, note que os vetores correspondentes às observações só aparecem em (Q) em expressões envolvendo produtos internos entre eles mesmos. Desta forma, não são os valores propriamente ditos de $\mathbf{x}^i$ que estão desempenhando um papel central na determinação da solução de (Q), mas apenas a relação entre eles por meio da operação de produto interno.

Para classificar um dado ponto $\mathbf{x}$, utilizamos a expressão de $\mathbf{w}$ dada em (11) e reescrevemos $\mathbf{w}^\top \mathbf{x} + w_0$ como $\sum_{i=1}^{m} \alpha_i y^i {\mathbf{x}^i}^\top \mathbf{x} + w_0$ e, assim, podemos definir a classificação de $\mathbf{x}$ como

$$c(\mathbf{x}) = \text{sinal}\left(\sum_{i=1}^{m} \alpha_i y^i \langle \mathbf{x}^i, \mathbf{x} \rangle + w_0\right). \tag{15}$$

A solução do problema (Q) nos fornece os multiplicadores α_i, mas não nos fornece o valor de w_0 diretamente. Para determinar w_0, podemos utilizar algum dos vetores suportes do hiperplano. A teoria da dualidade de programação quadrática nos diz que quando temos uma solução ótima de (Q), as observações cujos valores correspondentes α_i são diferentes de zero são exatamente aquelas que satisfazem $y^i(\mathbf{w}^\top \mathbf{x}^i + w_o) = 1$. Em outras palavras, as observações $\mathbf{x}^i$ para as quais $\alpha_i \neq 0$ são os vetores suportes do hiperplano.

Assim, seja $(\mathbf{x}^j, y^j) \in S$ tal que $\alpha_j \neq 0$. A observação $\mathbf{x}^j$ é um dos vetores suportes do hiperplano e, portanto, satisfaz $h(\mathbf{x}^j) = 1$. Consequentemente, temos $\mathbf{w}^\top \Phi(\mathbf{x}^j) + w_o = 1$. Usando o fato de que $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y^i \mathbf{x}^i$, podemos escrever

$$\begin{aligned}
\left(\sum_{i=1}^{m} \alpha_i y^i \mathbf{x}^i\right)^\top \mathbf{x}^j + w_0 &= 1 \\
\sum_{i=1}^{m} \alpha_i y^i {\mathbf{x}^i}^\top \mathbf{x}^j + w_0 &= 1 \\
\sum_{i=1}^{m} \alpha_i y^i \langle \mathbf{x}^i, \mathbf{x}^j \rangle + w_0 &= 1
\end{aligned} \tag{16}$$

Resolvendo (16) para w_0, obtemos o valor que faltava para determinarmos completamente a função de classificação (15).

3 O "Truque" do *Kernel*

Diante do fato de que apenas o cálculo do produto interno entre observações é necessário para a obtenção do classificador – em vez dos valores das observações em termos de suas características individuais –, podemos substituir as expressões $\langle \mathbf{x}^i, \mathbf{x}^j \rangle$ por $\langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle$, onde Φ é uma função que mapeia observações de S para um espaço de maior dimensão, que contém mais características do que o espaço em que S está imerso originalmente.

O interesse em fazer esse mapeamento reside no fato de que é possível que exista um hiperplano separador neste espaço aumentado com margem maior do que aquela do hiperplano ótimo no espaço original de S. Dependendo do mapeamento Φ adotado, isso permite a construção de superfícies de separação que, apesar de lineares no espaço aumentado, são não-lineares quando projetadas no espaço original de S. Este espaço de maior dimensão é comumente chamado de *espaço de características* (*feature space*, em inglês).

É muito importante destacar que a operação $\langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle$ pode, em muitos casos, ser realizada sem que o mapeamento Φ seja necessariamente realizado de forma explícita. Em outras palavras, não é necessário realmente mapear cada observação para o *feature space*. De fato, se tivermos como calcular o produto interno no espaço aumentado em função das observações originais – isto é, sem efetuar o mapeamento dos vetores envolvidos – podemos adotar um mapeamento para um espaço de dimensão arbitrariamente elevada, sem arcar com o custo decorrente do mapeamento e do cálculo do produto interno naquele espaço de maior dimensão. Este fato é chamado de *truque do kernel*, devido ao termo utilizado para designar a função que calcula $\langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle$, comumente denotada por $\kappa(\mathbf{x}^i, \mathbf{x}^j)$.

Assim, a mesma ferramenta que foi desenvolvida para criar um classificador linear no espaço pode ser utilizada para calcular um classificador não-linear, sem alto custo adicional e com maior capacidade de separar os dados. A superfície construída possui os vetores suportes assumindo o papel de pontos de controle.

A classificação de um novo ponto $\mathbf{x}$ é dada de maneira similar àquela da equação (15), com a diferença de que substituimos $\langle \mathbf{x}^i, \mathbf{x}^j \rangle$ por $\kappa(\mathbf{x}^i, \mathbf{x}^j)$:

$$c(\mathbf{x}) = \text{sinal}\left(\sum_{i=1}^{m} \alpha_i y^i \kappa(\mathbf{x}^i, \mathbf{x}^j) + w_0 \right). \tag{17}$$

Exemplo: $\Phi : \mathbf{x} = (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$, um mapa de $\mathbb{R}^2$ para $\mathbb{R}^3$. Desenvolvendo $\langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle$, temos

$$\begin{aligned} \langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle &= \left\langle ({x_1^i}^2, {x_2^i}^2, \sqrt{2}x_1^i x_2^i), ({x_1^j}^2, {x_2^j}^2, \sqrt{2}x_1^j x_2^j) \right\rangle \\ &= {x_1^i}^2 {x_1^j}^2 + {x_2^i}^2 {x_2^j}^2 + 2\, x_1^i x_2^i\, x_1^j x_2^j \\ &= \left(x_1^i x_1^j + x_2^i x_2^j \right)^2 = \langle \mathbf{x}^i, \mathbf{x}^j \rangle^2. \end{aligned}$$

Isto é, a função *kernel* $\kappa(\mathbf{x}^i, \mathbf{x}^j) = \langle \mathbf{x}^i, \mathbf{x}^j \rangle^2$ calcula o produto interno no espaço de maior dimensão em função apenas de $\mathbf{x}^i$ e $\mathbf{x}^j$, sem necessidade de efetuar o mapeamento explicitamente.

Outras funções kernel são comumente utilizadas. Como exemplo, o *kernel* polinomial de grau d e parâmetro c,

$$\kappa(\mathbf{x}^i, \mathbf{x}^j) = \left(\langle \mathbf{x}^i, \mathbf{x}^j \rangle + c \right)^d,$$

e o *kernel* de função de base radial (*RBF kernel*),

$$\kappa(\mathbf{x}^i, \mathbf{x}^j) = \exp\left(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{2\sigma^2} \right) = \exp\left(-\gamma \|\mathbf{x}^i - \mathbf{x}^j\|^2 \right),$$

com parâmetro σ ou γ, estão entre os mais frequentemente utilizados. Ambos estão disponíveis, por exemplo, no classificador `functions.SMO` do pacote WEKA, e no classificador `svm.SVC` do pacote scikit-learn, que implementam o algoritmo de SVM.

4 Dados Não Linearmente Separáveis

Caso o conjunto de treinamento não seja linearmente separável, é necessário ajustar o problema de otimização (P). A mudança deve permitir que exemplos estejam "do lado errado" do hiperplano, ou

seja, a violação das restrições (8) deve ser permitida, embora o ideal seja que poucas violações deste tipo aconteçam.

Para conseguir isso, acrescentamos ao problema uma variável de decisão ϵ_i, para cada observação i. A presença destas variáveis nas restrições, permite que $y^i\left(\mathbf{w}^\top \mathbf{x}^i + w_0\right) < 1$. Um valor positivo para ϵ_i compensaria este fato, fazendo com que $y^i\left(\mathbf{w}^\top \mathbf{x}^i + w_0\right) + \epsilon_i \geq 1$. Para evitar que muitas observações estejam do lado errado do hiperplano, um termo de penalização é introduzido na função a ser otimizada, no intuito de tornar não atrativa a atribuição de valores não-nulos às variáveis ϵ_i.

O problema primal passa a ser dado por:

$$\text{(P')} \quad \min_{\mathbf{w},w_0,\epsilon} \quad \frac{1}{2}\cdot\|\mathbf{w}\|^2 \;+\; C\sum_{i=1}^{m}\epsilon_i \tag{18}$$

$$s.a \quad y^i\left(\mathbf{w}^\top \mathbf{x}^i + w_0\right) + \epsilon_i \geq 1, \;\; \forall i \tag{19}$$

$$\epsilon_i \geq 0, \;\; \forall i. \tag{20}$$

Se os dados forem linearmente separáveis, não é necessário atribuir um valor positivo a nenhuma variável ϵ_i. No entanto, para dados não linearmente separáveis, é necessário que uma ou mais variáveis ϵ_i assumam valores diferente de zero.

O valor C é um hiperparâmetro que define a ênfase dada ao fator de penalização, em relação ao critério de otimização original, que envolve a largura da margem. Valores altos de C farão com que a resolução do problema de otimização tenda a fornecer um hiperplano com poucos erros (e talvez uma margem pequena). Por outro lado, valores pequenos de C farão com que o problema de otimização seja tolerante com relação a erros (permita várias observações no lado errado do hiperplano), ao mesmo tempo em que dá ênfase à maximização da margem. C é comumente chamado de *parâmetro de complexidade*.

Um desenvolvimento similar ao que fizemos para (P) nos leva a uma versão dual de (P'), similar a (Q). O valor de C juntamente com o tipo de função *kernel* utilizado (e seus parâmetros associados) constituem os principais hiperparâmetros para calibração deste algoritmo.

5 Dados com Duas ou Mais Classes

O aprendizado de dados com múltiplas classes usando SVM é feito por meio de múltiplos classificadores. Uma maneira de fazer isso é com o esquema *um-contra-todos*, em que vários classificadores SVM de duas classes são criados, um para cada classe y. Isso é feito definindo-se problemas de classificação entre duas classes: classe y e uma classe não-y, que consiste de todas as outras classes que não são y.

Ao classificar uma nova observação, o classificador prevê a classe como sendo aquela para a qual um dos classificadores forneceu a saída de maior valor do argumento da função sinal na Equação (15).

Outra forma de lidar com múltiplas classes é fazer um esquema *um-contra-um*, em que cada par de classes dá origem a um classificador. Ao classificar uma nova observação, a classe com mais vitórias é a classe prevista da observação. Esta é a versão implementada, por exemplo, no pacote WEKA (classificador `functions.SMO`).

6 Desvantagens

Apesar de ter rápido tempo de treinamento, alta precisão, e ser amplamente utilizado, com extensões para outras tarefas, como agrupamento, regressão, descoberta de novidade, etc, a técnica de SVM possui algumas desvantagens, dentre as quais apontamos:

- Não lida naturalmente com valores faltantes;
- Só lida com separação de duas classes. Para dados com múltiplas classes, exige a criação de vários classificadores;
- A função de decisão pode ser arbitrariamente complexa, a depender da função *kernel* utilizada. Isso torna impossível interpretar o classificador, em geral. De fato, o classificador nem mesmo é definido em função das características dos dados, e sim de produtos internos com alguns dos dados originais, o que torna muito difícil qualquer tipo de explicação sobre a função de classificação.

7 Um Exemplo Simples de Construção do Classificador SVM

Este seção contém um exemplo de construção de um classificador do tipo SVM para um conjunto de dados bem pequeno, com o intuito apenas de demonstrar o processo.

Recapitulando, descrevemos o seguinte problema de otimização

$$(\mathrm{Q}) \quad \max_{\alpha \in \mathbb{R}^m_+} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle \tag{21}$$

$$\text{sujeito a} \quad \sum_{i=1}^{m} \alpha_i y^i = 0. \tag{22}$$

Vamos considerar um conjunto de dados bem pequeno, composto por dois exemplos apenas:

$$D = \{((2,4),+1),((6,1),-1)\}.$$

O primeiro exemplo é $\mathbf{x}^1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ e pertence à classe $y^1 = +1$, enquanto o segundo exemplo é $\mathbf{x}^2 = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$ e pertence à classe $y^2 = -1$.

Para formular o problema de otimização (Q), teremos duas variáveis de decisão: $\alpha_1, \alpha_2 \geq 0$.

Escrevemos explicitamente a função objetivo (21) do problema (Q) da seguinte forma:

$$\begin{aligned} & \alpha_1 + \alpha_2 - \frac{1}{2} \cdot \left[\alpha_1 \alpha_1 y^1 y^1 \langle \mathbf{x}^1, \mathbf{x}^1 \rangle + \alpha_1 \alpha_2 y^1 y^2 \langle \mathbf{x}^1, \mathbf{x}^2 \rangle + \alpha_2 \alpha_1 y^2 y^1 \langle \mathbf{x}^2, \mathbf{x}^1 \rangle + \alpha_2 \alpha_2 y^2 y^2 \langle \mathbf{x}^2, \mathbf{x}^2 \rangle \right] = \\ = \quad & \alpha_1 + \alpha_2 - \frac{1}{2} \cdot \left[\alpha_1^2 \langle \mathbf{x}^1, \mathbf{x}^1 \rangle - 2\, \alpha_1 \alpha_2 \langle \mathbf{x}^1, \mathbf{x}^2 \rangle + \alpha_2^2 \langle \mathbf{x}^2, \mathbf{x}^2 \rangle \right]. \end{aligned}$$

Os valores dos produtos internos são dados por:

$$\langle \mathbf{x}^1, \mathbf{x}^1 \rangle = 20, \quad \langle \mathbf{x}^2, \mathbf{x}^2 \rangle = 37, \quad \langle \mathbf{x}^1, \mathbf{x}^2 \rangle = 16.$$

Agora, podemos escrever o problema (Q) por completo:

$$\begin{array}{lll} (\mathrm{Q}) & \max\limits_{\alpha_1,\alpha_2\in\mathbb{R}} & \alpha_1+\alpha_2-\dfrac{1}{2}\cdot\left(20\alpha_1^2-32\ \alpha_1\alpha_2+37\alpha_2^2\right) \\ & \text{sujeito a} & \alpha_1-\alpha_2=0. \end{array} \tag{23}$$

Como (Q) impõe $\alpha_1 = \alpha_2$, vamos substituir α_1 e α_2 por α apenas e reescrever a função (23) como

$$\begin{aligned} g(\alpha) &= \alpha+\alpha-10\alpha^2+16\ \alpha^2-\frac{37}{2}\alpha^2= \\ &= 2\alpha+6\ \alpha^2-\frac{37}{2}\alpha^2 \\ &= 2\alpha-\frac{25}{2}\alpha^2, \end{aligned}$$

que é uma função do segundo grau, cujo ponto de máximo podemos determinar, fazendo:

$$g'(\alpha)=2-25\alpha=0 \quad \therefore \quad \alpha=\frac{2}{25}.$$

Ainda segundo o roteiro de aula, temos $\mathbf{w}=\sum_{i=1}^{m}\alpha_i y^i\mathbf{x}^i$. Note que conhecemos todos os valores do lado direito dessa equação, o que nos permite obter $\mathbf{w}$. Para recuperarmos o hiperplano $\mathbf{w}^\top\mathbf{x}+w_0$, fazendo

$$\mathbf{w}^\top\mathbf{x}+w_0=\sum_{i=1}^{m}\alpha_i y^i\langle\mathbf{x}^i,\mathbf{x}\rangle+w_0 \tag{24}$$

precisamos apenas determinar o valor de w_0. Para isso, podemos notar que $h(\mathbf{x}^1)=1$ e $h(\mathbf{x}^2)=-1$. Ou seja, o hiperplano que passa exatamente sobre $\mathbf{x}^1$ é $h(\mathbf{x}^1)=1$, e podemos usar este fato para determinar o valor de w_0:

$$\begin{aligned} h(\mathbf{x}^1) &= 1 \\ \sum_{i=1}^{m}\alpha_i y^i\langle\mathbf{x}^i,\mathbf{x}^1\rangle+w_0 &= 1 \\ \frac{2}{25}\langle\mathbf{x}^1,\mathbf{x}^1\rangle+\frac{2}{25}(-1)\langle\mathbf{x}^1,\mathbf{x}^2\rangle+w_0 &= 1 \\ \frac{2}{25}20-\frac{2}{25}16+w_0 &= 1 \\ \frac{8}{25}+w_0 &= 1 \\ w_0 &= \frac{17}{25}. \end{aligned}$$

De forma similar, poderíamos ter usado o hiperplano que passa sobre $\mathbf{x}^2$ para obter:

$$
\begin{aligned}
h(\mathbf{x}^2) = \sum_{i=1}^{m} \alpha_i y^i \langle \mathbf{x}^i, \mathbf{x}^2 \rangle + w_0 &= -1 \\
\frac{2}{25}\langle \mathbf{x}^1, \mathbf{x}^2 \rangle + \frac{2}{25}(-1)\langle \mathbf{x}^2, \mathbf{x}^2 \rangle + w_0 &= -1 \\
\frac{2}{25}16 - \frac{2}{25}37 + w_0 &= -1 \\
-\frac{42}{25} + w_0 &= -1 \\
w_0 &= \frac{17}{25}.
\end{aligned}
$$

Agora, a partir de (24), podemos escrever

$$
\begin{aligned}
h(\mathbf{x}) &= \mathbf{w}^\top \mathbf{x} + w_0 = \sum_{i=1}^{m} \alpha_i y^i \langle \mathbf{x}^i, \mathbf{x} \rangle + w_0 \\
&= \frac{2}{25}\langle \mathbf{x}^1, \mathbf{x} \rangle + \frac{2}{25}(-1)\langle \mathbf{x}^2, \mathbf{x} \rangle + \frac{17}{25} \\
&= \frac{2}{25}\begin{pmatrix} 2 \\ 4 \end{pmatrix}^\top \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{2}{25}\begin{pmatrix} 6 \\ 1 \end{pmatrix}^\top \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{17}{25} \\
&= \frac{4}{25}x_1 + \frac{8}{25}x_2 - \frac{12}{25}x_1 - \frac{2}{25}x_2 + \frac{17}{25} \\
&= -\frac{8}{25}x_1 + \frac{6}{25}x_2 + \frac{17}{25}
\end{aligned}
$$

O hiperplano separador $h(\mathbf{x}) = 0$ estabelece a relação $-8x_1 + 6x_2 = -17$, que é a reta mostrada na figura 1. Os pontos $\mathbf{x}^1$ e $\mathbf{x}^2$ são mostrados em azul e vermelho, respectivamente.

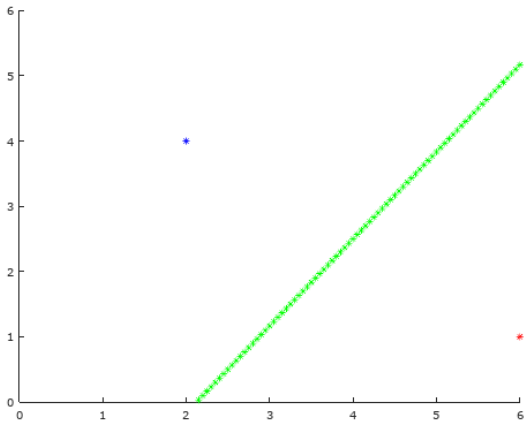

Figura 1: Hiperplano separador para exemplo.

8 Exemplo: Superfícies de Separação

Esta seção contém representações aproximadas das superfícies de separação obtidas com o classificador SVM (`functions.SMO` do pacote WEKA). O conjunto de dados utilizado foi o `Iris2D`, mas com apenas duas classes: os exemplos `Iris-versicolor` compõem a classe 1, enquanto os demais exemplos compõem a classe 2. O conjunto de treinamento é mostrado na figura 2. É fácil ver que o conjunto não é linearmente separável. Portanto, um *kernel* não-linear será necessário para se obter boa precisão.

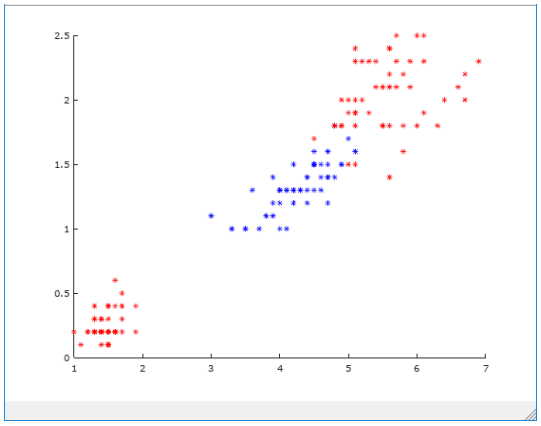

Figura 2: Conjunto `Iris2D` com classes `Iris-versicolor` (classe 1) e não-`Iris-versicolor` (classe 2).

Três classificadores SVM foram treinados, com variações nos parâmetros `c` (chamada de *parâmetro de complexidade*) e `E` (expoente do *kernel* polinomial). A tabela 1 mostra as configurações.

Nome do classificador	C	E
SVM1	1	3
SVM2	20	5
SVM3	50	10

Tabela 1: Configurações dos classificadores SVM.

As figuras 3, 4, e 5 mostram representações simplificadas das superfícies de separação dos três classificadores. Para isso, uma grade de pontos foi construída sobre o domínio do conjunto `Iris2D` e cada ponto da grade foi classificado por cada um dos três classificadores. Os círculos indicam a classificação feita pelo classificador, com círculos azuis indicando a classe 1 e círculos vermelhos indicando a classe 2. Nas figuras, ainda é possível ver os pontos originais, mostrados como asteriscos azuis e vermelhos.

Com base nas figuras e na tabela 1, é possível perceber que, à medida em que elevamos a dimensão do espaço aumentado onde o classificador é construído, e à medida em que o parâmetro de complexidade `c` aumenta (penalizando mais e mais os erros cometidos pelo classificador), a região prevista como pertencente à classe 1 (círculos azuis) se torna progressivamente mais ajustada aos exemplos do conjunto de treinamento que pertencem à classe 1.

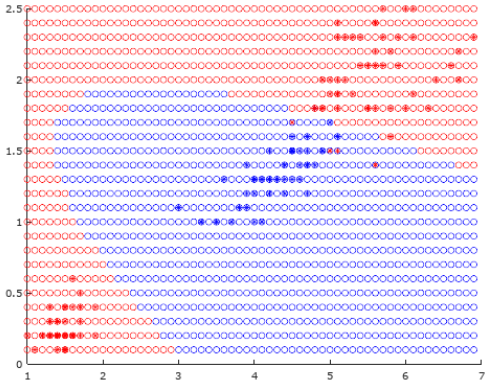

Figura 3: Classificador SVM1.

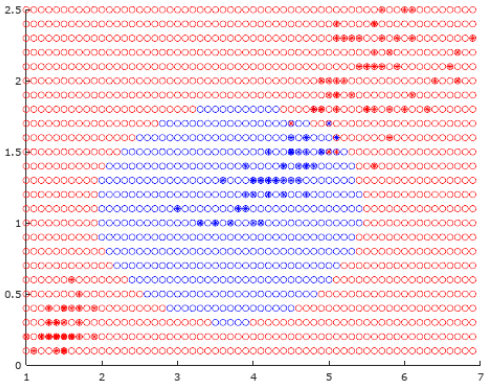

Figura 4: Classificador SVM2.

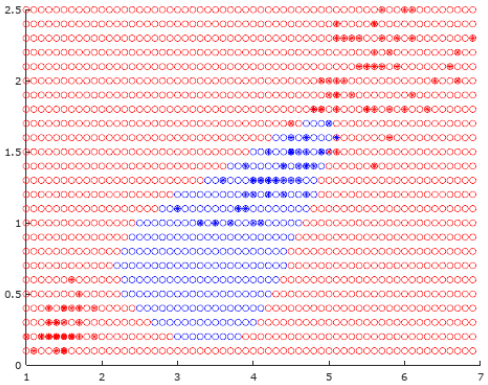

Figura 5: Classificador SVM3.