Advanced Institute for Artificial Intelligence

Máquinas de Vetores de Suporte

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

Introdução

As Máquinas de Vetores de Suporte, do inglês *Support Vector Machines* (SVM) são baseadas em conceitos da Teoria do Aprendizado Estatístico (TAE), desenvolvida por Vapnik e colegas. Basicamente, a ideia seria estudar garantias teóricas sobre condições necessárias para o processo de aprendizado.

Como dito anteriormente, temos duas principais limitações durante um processo de aprendizagem:

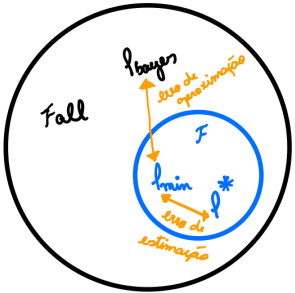
- supertreinamento (*overfitting*): baixa capacidade de generalização no conjunto de teste.
- subtreinamento (*underfitting*): baixa capacidade de aprendizado no conjunto de treinamento.

Qual seria a situação ideal? Um **compromisso** entre as duas situações, ou seja, uma relação custo-benefício entre supertreinamento e subtreinamento.

Definição do problema: dado um espaço de funções $\mathcal{F}$, como escolher uma função $\hat{f} \in \mathcal{F}$ de tal forma que o erro no treinamento seja baixo e a capacidade de generalização seja alta? A TAE nos fornece condições para atingir este objetivo sem assumir uma formulação específica para a distribuição dos dados (abordagem não paramétrica).

Objetivo: Encontrar o melhor classificador $f^* \in \mathcal{F}$ para um conjunto de treinamento fixo com tamanho m de tal forma que se aproxime, ao máximo, do classificador de menor risco, ou seja, f_{bayes}.

Quando o aprendizado é **consistente**, ou seja, quando f^* consegue aprender dos dados? Primeiramente, precisamos definir o espaço de funções $\mathcal{F}$ que o nosso classificador fará parte. Vamos analisar a figura abaixo.



- $\mathcal{F}_{all}$: espaço de todas as funções possíveis
- $\mathcal{F}$: espaço das funções que o classificador pode aprender
- f_{bayes}: classificador de mínimo risco possível
- f_{min}: classificador de mínimo risco em $\mathcal{F}$

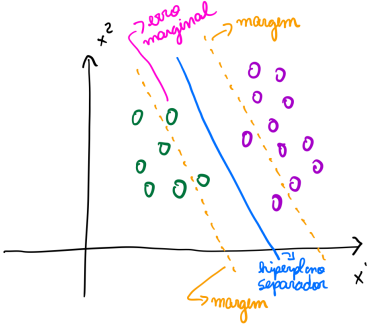
Podemos, então, associar um **risco** R à cada classificador. Desta forma, $R(f^*)$ corresponde ao risco associado ao classificador f^*.

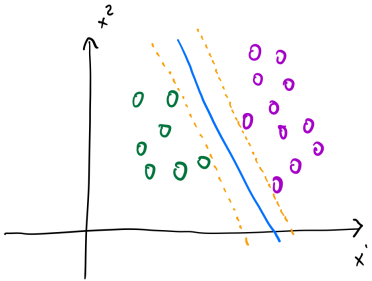
Temos que um classificador é dito ser **consistente** se, e somente se, o seu risco é minimizado quando $m \to \infty$, ou seja, quando o conjunto de treinamento aumenta. Em outras palavras, a consistência nos diz se estamos conseguindo aprender ou não.

O grande problema é que, quando $\mathcal{F} \to \mathcal{F}_{all}$, o nosso aprendizado não é consistente, pois o espaço de funções possíveis aumenta muito. Assim, devemos restringir o tamanho de $\mathcal{F}$. No entanto, o nosso dilema é: **ao restringirmos $\mathcal{F}$, o nosso erro de aproximação fica grande; ao ampliarmos $\mathcal{F}$, o nosso erro de estimação aumenta.**

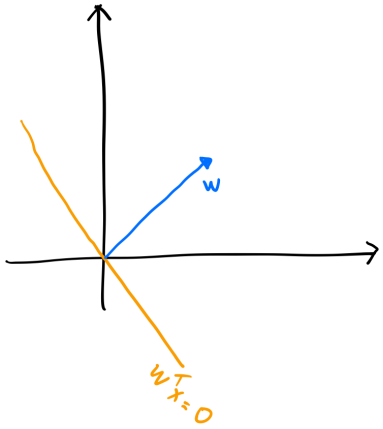
A pergunta principal é: **como escolher** $\mathcal{F} \subset \mathcal{F}_{all}$ **e** $f^* \in \mathcal{F}$**?** A TAE nos ajuda a responder à essa questão propondo as SVMs, que é uma classe de (funções) classificadores ótimas no que diz respeito ao **compromisso** entre os erros de aproximação e estimação.

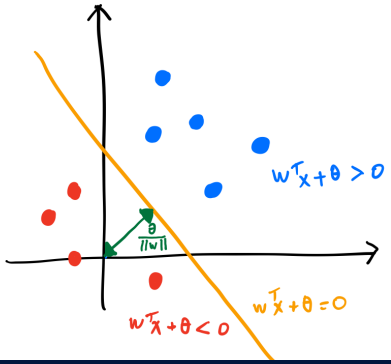
Objetivo: encontrar um hiperplano que maximize a **margem** de segurança e cometa poucos **erros marginais**, isto é, minimize o risco. O que são erros marginais?



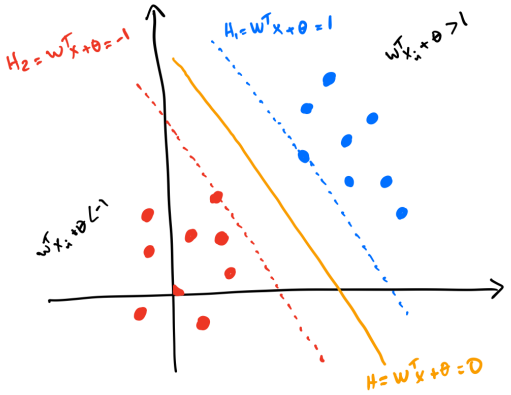


Fazemos uma analogia com o algoritmo do Perceptron no sentido que ambos utilizam uma função de decisão linear, ou seja, um hiperplano. A diferença é que o hiperplano do Perceptron não possui as **propriedades ótimas** que o hiperplano encontrado pelo SVM possui.





Definimos, também, como **hiplerplano canônico** H aquele cujas amostras mais próximas satisfaçam $|\boldsymbol{w}^T\boldsymbol{x} + \theta| = 1$. Como ilustrado abaixo, temos que os hiperplanos H_1 e H_2 definem, então, as margens com relação ao hiperplano canônico.



Definição do problema: dado um conjunto de treinamento rotulado, isto é, classificação supervisionada, $\mathcal{X}^1 = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, as amostras mais próximas do hiperplano canônico H devem satisfazer as seguintes condições:

- $H_1 : \boldsymbol{w}^T \boldsymbol{x}_i + \theta = 1 \implies \boldsymbol{w}^T \boldsymbol{x}_i + (\theta - 1) = 0$
- $H_2 : \boldsymbol{w}^T \boldsymbol{x}_i + \theta = -1 \implies \boldsymbol{w}^T \boldsymbol{x}_i + (\theta + 1) = 0$

Em termos de classificação, temos que:

$$\boldsymbol{w}^T \boldsymbol{x}_i + \theta \leq -1 \text{ se } y_i = -1$$

e

$$\boldsymbol{w}^T \boldsymbol{x}_i + \theta \geq 1 \text{ se } y_i = 1.$$

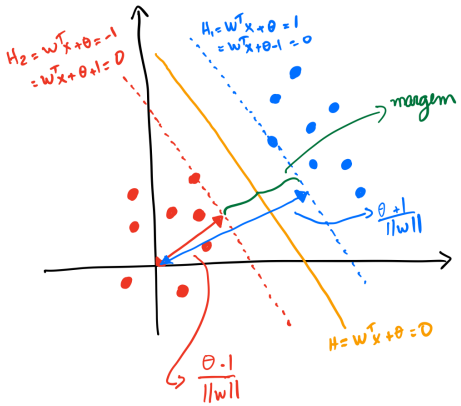
Podemos unificar as equações anteriores da seguinte forma:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1 \geq 0, \tag{1}$$

que é conhecida pela **restrição do problema**.

Note que teremos m restrições, ou seja, uma para cada amostra do conjunto de treinamento, o que é bastante custoso para o classificador. Esse é um dos motivos que torna o SVM uma técnica bastante cara computacionalmente, sendo um dos seus principais pontos negativos.

Como podemos calcular a distância $d(H_1, H_2)$ entre os hiperplanos?



$$
\begin{aligned}
d(H_1, H_2) &= \frac{\theta + 1}{\|\boldsymbol{w}\|} - \frac{\theta - 1}{\|\boldsymbol{w}\|} \\
&= \frac{(\theta + 1) - (\theta - 1)}{\|\boldsymbol{w}\|} \\
&= \frac{\theta + 1 - \theta + 1}{\|\boldsymbol{w}\|} \\
&= \frac{2}{\|\boldsymbol{w}\|}
\end{aligned}
\tag{2}
$$

Desta forma, para **maximizar** a margem de separação, devemos **minimizar** $\|\boldsymbol{w}\|$. Assim sendo, temos o seguinte problema de otimização:

$$\boldsymbol{w}^*, \theta^* = \arg\min_{\boldsymbol{w},\theta} \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{3}$$

sujeito às seguintes restrições:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1 \geq 0, \quad \forall i = 1, 2, \ldots, m.$$

As restrições acima garantem que não existem dados de treinamento entre os hiperplanos de separação.

Quando temos um problema de otimização sujeito à restrições de desigualdade, utilizamos duas principais ferramentas matemáticas: condições **KKT** (Karush–Kuhn–Tucker) e os chamados **Multiplicadores de Lagrange**.

O primeiro passo para a resolução da Equação 3 é criar a função Lagrangiana, a qual vai incorporar as restrições na própria função objetivo:

$$L(\boldsymbol{w}, \theta, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1], \tag{4}$$

também conhecida como forma **primal** do problema de otimização, em que $\boldsymbol{\alpha} \in \mathbb{R}^m$ denota os multiplicadores de Lagrange.

O nosso problema agora passar a ser a otimizar a função Lagrangiana dada pela Equação 4, o que implica em encontrar os valores de $\boldsymbol{\alpha}$, $\boldsymbol{w}$ e θ para os quais o gradiente da função é nulo, ou seja:

$$\nabla L(\boldsymbol{w}, \theta, \boldsymbol{\alpha}) = 0. \tag{5}$$

Isto implica que as derivadas parciais em relação aos parâmetros do hiperplano que desejamos encontrar devem ser nulas, ou seja:

$$\frac{\partial L(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \theta} = 0 \tag{6}$$

e

$$\frac{\partial L(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0. \tag{7}$$

A derivada parcial com relação ao bias, ou seja, parâmetro θ é dada como segue:

$$
\begin{aligned}
\frac{\partial L(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \theta} &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1] = 0 \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} [\alpha_i y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - \alpha_i] = 0 \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) + \sum_{i=1}^{m} \alpha_i = 0 \qquad (8) \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} [\alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i + \alpha_i y_i \theta] + \sum_{i=1}^{m} \alpha_i = 0 \\
&= \sum_{i=1}^{m} \alpha_i y_i = 0.
\end{aligned}
$$

Já a derivada parcial com relação ao vetor $\boldsymbol{w}$ é dada por:

$$
\begin{aligned}
\frac{\partial L(\boldsymbol{w}, \theta, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1] = 0 \\
&= \boldsymbol{w} - \sum_{i=1}^{m} [\alpha_i y_i (\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - \alpha_i] = 0 \\
&= \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i (\boldsymbol{w}^T \boldsymbol{x}_i + \theta) + \sum_{i=1}^{m} \alpha_i = 0 \qquad (9) \\
&= \boldsymbol{w} - \sum_{i=1}^{m} [\alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i + \cancelto{0}{\alpha_i y_i \theta}] + \sum_{i=1}^{m} \alpha_i = 0 \quad \overset{0}{\nearrow} \\
&= \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \implies \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i.
\end{aligned}
$$

Substituindo-se as Equações 8 e 9 na forma primal do problema (Equação 4), encontramos a sua forma **dual**. Relembrando a forma primal, temos que ela é composta por dois termos:

$$L(\boldsymbol{w}, \theta, \boldsymbol{\alpha}) = \overbrace{\frac{1}{2}\|\boldsymbol{w}\|^2}^{\text{Termo 1}} \underbrace{- \sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1]}_{\text{Termo 2}}.$$

Desenvolvendo o primeiro termo e utilizando a Equação 9, temos que:

$$\begin{aligned} \frac{1}{2}\|\boldsymbol{w}\|^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} &= \frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i\right)^T \left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i\right) \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j). \end{aligned} \tag{10}$$

Desenvolvendo o segundo termo, temos que:

$$
\begin{aligned}
-\sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1] &= -\sum_{i=1}^{m} \alpha_i [y_i \boldsymbol{w}^T \boldsymbol{x}_i + y_i \theta - 1] \\
&= -\sum_{i=1}^{m} [\alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i + \alpha_i y_i \theta - \alpha_i] \\
&= -\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i - \sum_{i=1}^{m} \alpha_i y_i \theta + \sum_{i=1}^{m} \alpha_i \\
&= -\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i - \theta \sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i.
\end{aligned}
\tag{11}
$$

Podemos, ainda, desenvolver a Equação 11 um pouco mais.

$$
\begin{aligned}
-\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i - \theta \underbrace{\sum_{i=1}^{m}\alpha_i y_i}_{0 \text{ (Equação 8)}} + \sum_{i=1}^{m}\alpha_i &= -\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \\
&= -\sum_{i=1}^{m}\alpha_i y_i \underbrace{\left(\sum_{j=1}^{m}\alpha_j y_j x_j\right)^T}_{\text{Equação 9}} \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \qquad (12) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=1}^{m}\alpha_i.
\end{aligned}
$$

Substituindo-se, então, as Equações 10 (primeiro termo) e 12 (segundo termo) na Equação 4, temos que:

$$
\begin{aligned}
L(\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^{m}\alpha_i. \qquad (13)
\end{aligned}
$$

Note que, agora, nossa função objetivo depende apenas dos multiplicadores de Lagrange $\boldsymbol{\alpha}$. Lembre-se que, na Equação 4, nossa função dependia de $\boldsymbol{\alpha}$, $\boldsymbol{w}$ e θ.

Como o problema **primal** era de **minimização**, agora o problema **dual** torna-se uma **maximização** (teoria de otimização matemática). Assim sendo, nosso problema de otimização consiste em resolver a seguinte formulação:

$$\begin{aligned}\boldsymbol{\alpha}^* &= \arg\max_{\boldsymbol{\alpha}} \{L(\boldsymbol{\alpha})\} \\ &= \arg\max_{\boldsymbol{\alpha}} \left\{ -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=1}^{m} \alpha_i \right\}, \end{aligned} \tag{14}$$

sujeito às seguintes restrições:

$$\alpha_i \geq 0, \quad \forall i = 1, 2, \ldots, m$$

e

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

Qual a vantagem de termos um problema em sua forma dual? Temos apenas uma variável de interesse, ou seja, $\boldsymbol{\alpha}$. Assim, para resolvermos a Equação 14, precisamos empregar algum método de otimização quadrática (somatório duplo). É por esse motivo que o processo de treinamento do classificador SVM é **bastante custoso computacionalmente**.

Desta forma, obtendo-se $\boldsymbol{\alpha}^*$ por algum método de otimização matemática, basta utilizamos ele na Equação 9 para obtemos $\boldsymbol{w}$. Já para calcularmos o θ, precisamos utilizar as condições de KKT, as quais são condições **necessárias** para que a solução do problema dado pela Equação 14 seja ótima. Essas condições são dadas pelas seguintes restrições:

$$\alpha_i^*(y_i[(\boldsymbol{w}^*)^T\boldsymbol{x}_i + \theta] - 1) = 0,\ \forall i = 1, 2, \ldots, m. \tag{15}$$

Note que $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) - 1 = 0$ ocorre apenas para as amostras que se encontram **sobre** os hiperplanos H_1 e H_2. Assim, α_i pode ser **não nulo** apenas sobre os elementos $\boldsymbol{x}_i$ que se encontram nesses hiperplanos. Para as demais amostras $\boldsymbol{x}_j$, isto é, aquelas estão estão definidas além das margens, temos que $\alpha_j = 0$ para a condição ser satisfeita.

Para esses elementos cujos multiplicadores de Lagrange são não nulos damos o nome de **vetores de suporte**. Eles correspondem às amostras **mais informativas** do conjunto de treinamento, pois estão mais próximas do hiperplano separador. Assim sendo, apenas essas amostras são utilizadas no cálculo de H, conforme apresenta a Equação 9.

Como podemos calcular θ? Sabemos que, para as amostras que estão nas margens, a seguinte igualdade é válida: $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) = 1 \implies y_i\boldsymbol{w}^T\boldsymbol{x}_i + y_i\theta = 1 \implies \theta = \dfrac{1 - y_i\boldsymbol{w}^T\boldsymbol{x}_i}{y_i}$. Temos, então, que θ^* pode ser calculado como segue:

$$\theta^* = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_i \in \mathcal{S}} \left\{ \frac{1}{y_i} - (\boldsymbol{w}^*)^T \boldsymbol{x}_i \right\}, \tag{16}$$

em que $\mathcal{S} \subseteq \mathcal{X}^1$ denota o conjunto dos vetores de suporte. Podemos, ainda, escrever a equação acima substituindo $\boldsymbol{w}^*$ pela Equação 9, isto é, em função dos multiplicadores de Lagrange:

$$\theta^* = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_i \in \mathcal{S}} \left\{ \frac{1}{y_i} - \sum_{\boldsymbol{x}_j \in \mathcal{S}} \alpha_j^* y_j \boldsymbol{x}_j^T \boldsymbol{x}_i \right\}. \tag{17}$$

A função de decisão (classificação) do SVM apenas retorna o sinal da amostra que está sendo avaliada, ou seja:

$$h_{\boldsymbol{w}^*}(x) = \operatorname{sgn}\{(\boldsymbol{w}^*)^T \boldsymbol{x} + \theta^*\} = \operatorname{sgn}\left\{ \underbrace{\sum_{\boldsymbol{x}_i \in \mathcal{S}} y_i \alpha_i^* \boldsymbol{x}_i^T}_{\boldsymbol{w}^*} \boldsymbol{x} + \theta^* \right\}. \tag{18}$$

Neste caso, caso $h_{\boldsymbol{w}^*}(x) < 0$ (sinal negativo), então a amostra é classificada como sendo da classe -1; ou da classe 1 caso contrário. Note que $\boldsymbol{w}^*$ controla a **inclinação** do hiperplano ótimo, θ^* define a sua **posição** em relação às **margens**.

SVMs com Margens Suaves

O objetivo dessa variante é lidar com o problema de sobreposição entre as classes permitindo com que algumas amostras possam violar a restrição e situar-se **entre** as margens.

Lembrando que as restrições originais eram dadas por $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) \geq 1,\ \forall i = 1, 2 \ldots, m.$ Agora, nossas novas restrições são dadas por:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \theta) \geq 1 - \epsilon_i, \tag{19}$$

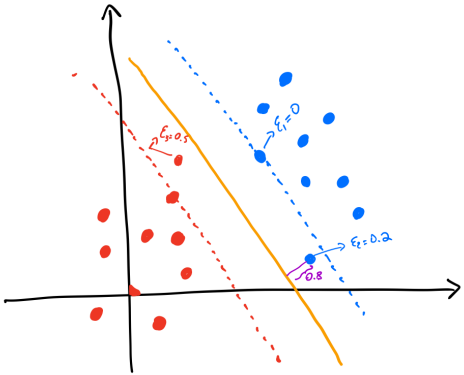
em que $\epsilon_i \geq 0$ representam as variáveis de folga que possuem o papel de "suavizar" as margens.

Agora, nossa função objetivo é dada por:

$$\boldsymbol{w}^*, \theta^*, \boldsymbol{\epsilon}^* = \arg\min \left\{ \frac{1}{2}\|w\|^2 + C \left(\sum_{i=1}^{m} \epsilon_i \right) \right\}, \tag{20}$$

sujeito a $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \theta) \geq 1 - \epsilon_i$, em que $\epsilon_i \in [0, 1]$ indica que os dados estão entre as margens. Note que o parâmetro C controla a relação custo-benefício entre SVMs de margens rígidas e margens suaves.

A ideia, então, é permitir que amostras estejam posicionadas entre as margens, conforme ilustrado abaixo.



A forma dual de nosso problema de otimização é muito similar ao que tínhamos apresentado na formulação das SVMs de margens rígidas (Equação 15), ou seja:

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \left\{ -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=1}^{m} \alpha_i \right\},$$

sujeito às seguintes restrições:

$$0 \leq \alpha_i \leq C, \quad \forall i = 1, 2, \ldots, m$$

e

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

O nosso vetor de pesos $\boldsymbol{w}^*$ continua sendo calculado pela Equação 9. Já as variáveis de folga podem ser calculadas da seguinte maneira:

$$\epsilon_i^* = \max \left\{ 0, 1 - \left[\underbrace{y_i \sum_{j=1}^{m} y_j \alpha_j^* \boldsymbol{x}_j^T}_{\boldsymbol{w}^*} \boldsymbol{x}_i \right] + \theta^* \right\}. \tag{21}$$

Notem que interessante: suponha uma amostra $\boldsymbol{x}_i$ que esteja sobre o hiperplano H_1, ou seja, $y_i = 1$. Neste caso, o termo $y_i \sum_{j=1}^{m} y_j \alpha_j^* \boldsymbol{x}_j^T \boldsymbol{x}_i + \theta^* = (\boldsymbol{w}^*)^T \boldsymbol{x}_i + \theta^* = 1$. A Equação 21 acaba se transformando em $\epsilon_i^* = \max\{0, 1 - 1\} = 0$, o que é verdade, pois $\boldsymbol{x}_i$ é um vetor de suporte.

Suponha, agora, uma outra amostra $\boldsymbol{x}_i$ que pertença à classe 1, ou seja, $y_1 = 1$, mas não está sobre a margem. Neste caso, temos que:

$$y_i \sum_{j=1}^{m} y_j \alpha_j^* \boldsymbol{x}_j^T \boldsymbol{x}_i + \theta^* = (\boldsymbol{w}^*)^T \boldsymbol{x}_i + \theta^* = k > 1.$$

A Equação 21 acaba se transformando em $\epsilon_i^* = \max\{0, 1 - k\} = 0$, o que é também verdade, pois $\boldsymbol{x}_i$ está longe da margem e, portanto, não precisa de variável da folga.

Suponha, agora, que $\boldsymbol{x}_i$ pertença à classe -1, ou seja, $y_i = -1$, mas não está sobre a margem. Neste caso, temos que:

$$y_i \sum_{j=1}^{m} y_j \alpha_j^* \boldsymbol{x}_j^T \boldsymbol{x}_i + \theta^* = -(\boldsymbol{w}^*)^T \boldsymbol{x}_i - \theta^* = -k < -1.$$

Multiplicando-se a equação acima por -1, temos que:

$$-(\boldsymbol{w}^*)^T\boldsymbol{x}_i - \theta^* = -k \implies (\boldsymbol{w}^*)^T\boldsymbol{x}_i + \theta^* = k > 1.$$

A Equação 21 acaba se transformando, novamente, em $\epsilon_i^* = \max\{0, 1-k\} = 0$, o que é também verdade, pois $\boldsymbol{x}_i$ não encontra-se sobre a margem e, portanto, não precisa de variável de folga.

Agora, notem que temos duas condições de KKT:

$$\alpha_i^*(y_i[(\boldsymbol{w}^*)^T\boldsymbol{x}_i + \theta] - 1 + \boldsymbol{\epsilon}^*) = 0,\ \forall i = 1, 2, \ldots, m. \tag{22}$$

e

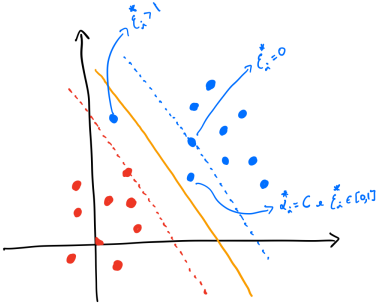
$$(C - \alpha_i^*)\epsilon_i^* = 0. \tag{23}$$

No caso das SVMs com margens rígidas, tínhamos que os vetores de suporte eram as amostras $\boldsymbol{x}_i$ cujos multiplicadores de Lagrange $\alpha_i > 0$. Agora, no caso das SVMs com margens suaves, temos que existem diferentes tipos de vetores de suporte de acordo com os valores de α_i e ϵ_i.

Caso $\alpha_i < C$, temos que $\epsilon_i^* = 0$ para a Equação 23 ser satisfeita. Assim, concluímos que $\boldsymbol{x}_i$ é um vetor de suporte que está sobre a margem H_1 ou H_2 (depende de sua classe) ou é uma amostra que não está posicionada na margem. Este seria o mesmo caso das SVMs com margens rígidas.

Agora, caso $\alpha_i^* = C$, temos três possíveis situações:

1. $\epsilon_i^* > 1$: erros no conjunto de treinamento pois os vetores de suporte cruzam o hiperplano separador.
2. $0 \leq \epsilon_i^* < 1$: amostras situadas entre as margens (não são erros, apenas estão violando as regras).
3. $\epsilon_i^* = 0$: amostras sobre as margens ou distante delas.



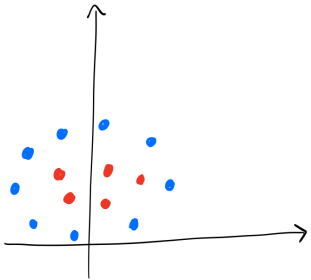
Falta, agora, calcularmos θ^*, que é computado pela média da Equação 17 sobre todos os vetores de suporte com $\alpha_i < C$, ou seja:

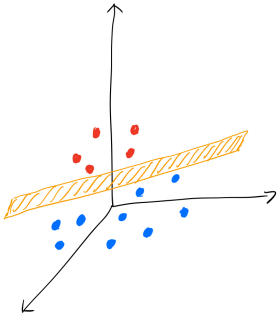
$$\theta^* = \frac{1}{|\mathcal{V}|} \sum_{\boldsymbol{x}_i \in \mathcal{V}} \left\{ \frac{1}{y_i} - \sum_{\boldsymbol{x}_j \in \mathcal{S}} \alpha_j^* y_j \boldsymbol{x}_j^T \boldsymbol{x}_i \right\}. \tag{24}$$

Lembrando que $\mathcal{S}$ denota o conjunto de todos os vetores de suporte (livres e limitados), e $\mathcal{V}$ denota os vetores de suporte limitados apenas, ou seja, aquelas amostras para as quais $\alpha_i^* < C$.

SVMs Não-Lineares

Quando temos dados que são linearmente separáveis, podemos fazer uso das SVMs com margens suaves ou rígidas, como vimos anteriormente. No entanto, como podemos lidar com o caso de dados não lineares, visto que os modelos baseados em SVM aprendidos até agora consegue aprender apenas superfícies de decisão lineares? **A solução é mapear os dados para um outro espaço de maior dimensão** Esta solução é também conhecida por *kernel trick*.





Como é a ideia dessa **função de mapeamento**? Suponha que tenhamos o seguinte exemplo:

- $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ (função de mapeamento)
- $\boldsymbol{x} \in \mathbb{R}^2$ (elemento de entrada)
- $\phi(\boldsymbol{x}) = \phi(x^1, x^2) = ((x^1)^2, \sqrt{2}x^1x^2, (x^2)^2)$ (elemento mapeado)

A ideia é que os dados estarão melhor distribuídos em espaços com maiores dimensões. O **Teorema de Cover** diz que um problema de classificação de padrões mapeado não-linearmente para um espaço de maior dimensão é mais provável de ser linearmente separável do que no espaço original, dado que o espaço não é densamente povoado.

Novamente, o problema dual é dado pela seguinte formulação:

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \left\{ -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_j) + \sum_{i=1}^{m} \alpha_i \right\}. \tag{25}$$

A solução é dada, também, de maneira similar, ou seja:

$$h_{\boldsymbol{w}^*}(x) = \operatorname{sgn}\{(\boldsymbol{w}^*)^T \boldsymbol{x} + \theta^*\} = \operatorname{sgn}\left\{ \underbrace{\sum_{\boldsymbol{x}_i \in \mathcal{S}} y_i \alpha_i^* \phi(\boldsymbol{x}_i)}_{\boldsymbol{w}^*} \phi(\boldsymbol{x}) + \theta^* \right\}. \tag{26}$$

O parâmetro θ^*, calculado na Equação 24, também pode ser obtido de maneira similar:

$$\theta^* = \frac{1}{|\mathcal{V}|} \sum_{\boldsymbol{x}_i \in \mathcal{V}} \left\{ \frac{1}{y_i} - \sum_{\boldsymbol{x}_j \in \mathcal{S}} \alpha_j^* y_j \phi(\boldsymbol{x}_j) \phi(\boldsymbol{x}_i) \right\}. \tag{27}$$

A pergunta é: como projetar o operador de mapeamento ϕ? Uma solução é por meio das funções *kernel*, em que não é necessário conhecer um mapeamento específico para determinado dado, apenas como realizar produtos escalares no novo espaço.

Uma função de kernel K é definida, basicamente, como segue:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j), \tag{28}$$

ou seja, a função kernel recebe dois vetores no espaço de entrada e retorna o valor do produto escalar das amostras no espaço de maior dimensão. Esse é o então chamado *kernel trick*. Vejamos o exemplo abaixo:

- $\phi : \mathbb{R}^2 \to \mathbb{R}^3$, $\phi(\boldsymbol{x}) = \phi(x^1, x^2) = ((x^1)^2, \sqrt{2}x^1x^2, (x^2)^2)$
- $\boldsymbol{x}_i = (x_i^1, x_i^2)$ e $\boldsymbol{x}_j = (x_j^1, x_j^2)$
- $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \overbrace{((x_i^1)^2, \sqrt{2}x_i^1x_i^2, (x_i^2)^2)^T((x_j^1)^2, \sqrt{2}x_j^1x_j^2, (x_j^2)^2)}^{\text{produto interno}} = (x_i^1)^2(x_j^1)^2 + \sqrt{2}x_i^1x_i^2\sqrt{2}x_j^1x_j^2 + (x_i^2)^2(x_j^2)^2 = (x_i^1)^2(x_j^1)^2 + 2x_i^1x_i^2x_j^1x_j^2 + (x_i^2)^2(x_j^2)^2 = (\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \implies$ Lembrando que $(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = (x_i^1x_j^1 + x_i^2x_j^2)^2 = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, o conhecido **kernel polinomial de grau** 2

Desta forma, ao calcularmos $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ para todos os m elementos do conjunto de treinamento, iremos obter a **matriz de kernel** $\boldsymbol{K} \in \mathbb{R}^{m \times m}$. Dizemos que uma matriz de kernel é válida caso ela atenda às **Condições de Mercer**, ou seja, ela é uma matrix positiva e semi-definida (autovalores são todos maiores ou iguais a zero).

Alguns exemplos de função kernel válidas:

- Polinomial: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i, \boldsymbol{x}_j)^d$
- Gaussiano: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma}\right\}$
- Sigmoidal: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\boldsymbol{x}_i, \boldsymbol{x}_j)$

Um outro desafio é: como tornar SVM válido para problemas de classificação com **múltiplas classes**? Basicamente, para um problema com c classes temos duas abordagens:

- OVA (*one-versus-all*): treinamos c SVMs e removemos a função sgn da Equação 26 para que a mesma retorne uma pontuação (*score*). A classe da amostra corresponde àquela SVM que retornou a maior pontuação. **Essa pontuação geralmente é calculada como sendo a distância da amostra a ser classificada até o hiperplano de separação (maior a distância, maior a pontuação).**
- OVO (*one-versus-one*): treinamos $c(c-1)/2$ classificadores SVM, isto é, todos os pares de combinações entre classes. Para cada combinação, o classificador vencedor recebe um voto, e a classe escolhida é a do classificador que recebeu a maior quantidade de votos. **Uma desvantagem dessa abordagem é que, quando o número de classes é alto, ela é muito custosa.**