Máquinas de Vetores-Suporte (SVMs)
Parte I

1.Introdução às Máquinas de Vetores-Suporte

- Conforme vimos anteriormente no curso, o problema de classificação corresponde à tarefa de categorizar padrões pertencentes a um espaço de características n-dimensional em m classes (tipicamente disjuntas).
- No caso supervisionado (de que trataremos aqui), o problema é abordado por meio do projeto de uma *máquina* (conhecida como *classificador*) a partir de exemplos devidamente rotulados.

- Naturalmente, o interesse não jaz na obtenção de erros de treinamento reduzidos, mas sim num treinamento que leve a uma adequada generalização.
- Se considerarmos que a estrutura de classificação é linear, ou seja, que o classificador dá origem a um *hiperplano*[1], é possível mostrar que a qualidade da generalização tem a ver com a ideia de **margem**. Uma compreensão mais aprofundada dessa conexão requer um estudo sistemático da teoria de aprendizado estatístico (VAPNIK, 1998; BISHOP, 2006), estudo este que transcende o escopo deste curso. Ficaremos, portanto, com a ideia geral: **maximização da margem se relaciona com uma melhor generalização**.
- Talvez estejamos nos apressando, pois nem chegamos a definir **margem**. Felizmente, o conceito é intuitivo: é uma espécie de "folga" que o hiperplano

---

[1] Em duas dimensões, o hiperplano se reduz a uma reta; em três, a um plano.

tem com respeito à classificação dos dados disponíveis. A Fig. 1 traz uma ilustração.

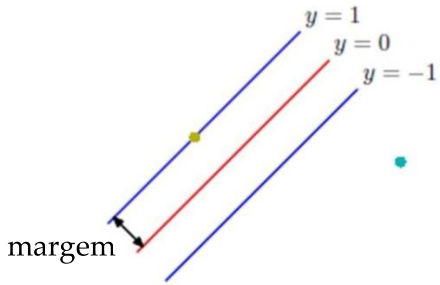

Fig. 1 – Ilustração do Conceito de Margem.

- Em linguagem mais formal, a margem pode ser definida como a *distância perpendicular entre a fronteira de decisão e o(s) dado(s) mais próximo(s) a ela*.
- É natural, a esta altura, que indaguemos como seria o projeto de um classificador linear de máxima margem.

1.1. Classificador Linear de Máxima Margem

- Se os dados forem linearmente separáveis, haverá infinitos hiperplanos capazes de separá-los. À guisa de exemplo, consideremos os dois hiperplanos (no caso, retas) da Fig. 2.

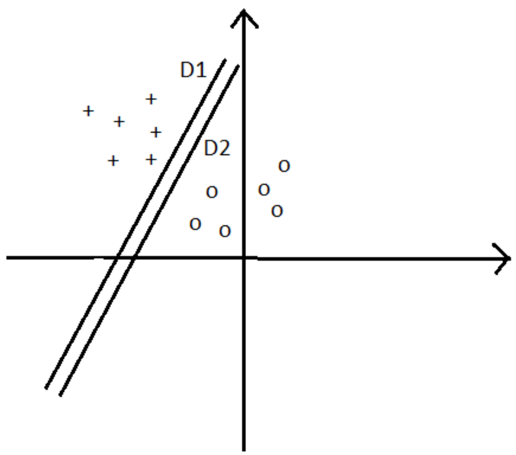

Figura 2 – Duas Fronteiras Lineares de Separação.

- Há dados de duas classes, os dados D1 e D2. Perceba como uma das retas está bem distante dos dados D2, mas "perigosamente" próxima aos dados D1. Já a outra reta paira de maneira mais segura entre as duas classes[2]. Situações do segundo tipo são mais interessantes se buscamos uma melhor generalização[3].

- Nesse espírito, consideraremos o problema de projetar um classificador linear (i.e., um hiperplano) de máxima margem. Seguiremos, via de regra, a linha de raciocínio do clássico trabalho (CORTES & VAPNIK, 1995).
- Tomemos um conjunto de dados $\{\mathbf{x}_i, d_i\}, i = 1, \ldots, N$, com $\mathbf{x}_i \in \mathbb{R}^n$ e $d_i \in \{+1, -1\}$. Trocando em miúdos, temos um problema com N amostras, cada uma caracterizada por n atributos. Há duas classes, às quais, por simetria, associamos os números "-1" e "$+1$" (a pertinência disso ficará clara adiante).
- Se os dados forem linearmente separáveis, valerá a seguinte condição (para algum $\mathbf{w}$, algum b e $i = 1, \ldots, N$):

[2] O que traz à mente o célebre trecho de Ovídio: "*medio tutissimus ibis*".

[3] No treinamento, ambos os classificadores são perfeitos, pois separam os dados sem erros.

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq 1 \text{ se } d_i = 1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1 \text{ se } d_i = -1 \end{cases}$$

- Note que ser maior ou igual a 1 ou menor ou igual a -1 é arbitrário: outros valores poderiam ter sido usados (2 e -2, ou 0,5 e $-0{,}5$, por exemplo). O que importa é que haverá um valor χ e um valor $-\chi$ que serão, de certo modo, limiares de classificação para o conjunto de dados.
- A condição acima pode ser reescrita de maneira mais sucinta:

$$d_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1, \dots, N$$

- Dentre os hiperplanos que separam os dados, ou seja, que atendem à condição expressa acima, desejamos o de *máxima margem*. Essa condição é simples de formalizar no âmbito das projeções engendradas pelo classificador.
- Para obter essas projeções, esqueçamos por ora o termo de *bias* (b). O classificador já gera sua saída por meio de uma espécie de projeção, o produto escalar $\mathbf{w}^T\mathbf{x}$. Entretanto, esse termo é sensível a fatores de escala: se

multiplicarmos o vetor de pesos por, digamos, dez, as projeções serão também multiplicadas por esse valor.

- Para evitar esse efeito artificial, consideremos a projeção realizada pelo vetor de pesos normalizado, ou seja, $\frac{\mathbf{w}}{\|\mathbf{w}\|}$. Vamos tomar os dados projetados que são da classe $+1$ e os dados projetados que são da classe -1. Consideraremos então os menores valores projetados da classe $+1$ e os maiores valores projetados da classe -1: eles formarão os *dados limítrofes*, por assim dizer. Maximizar a distância entre esses casos limítrofes significará, destarte, *maximizar a margem de separação entre classes*. Passemos a uma análise mais rigorosa.
- A distância mencionada entre projeções é:

$$\rho(\mathbf{w}, b) = min_{\{\mathbf{x}:d=+1\}} \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} - max_{\{\mathbf{x}:d=-1\}} \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|}$$

Repitamos agora a expressão de separabilidade linear:

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, \dots, N$$

- Percebe-se que, nessa expressão, o caso de mínimo valor (em ρ) para a classe $d = +1$ será a condição limítrofe $\mathbf{w}^T\mathbf{x} + b = 1$, ou seja, $\mathbf{w}^T\mathbf{x} = 1 - b$. Já o caso de máximo valor para a classe $d = -1$ será a condição $\mathbf{w}^T\mathbf{x} + b = -1$, ou seja, $\mathbf{w}^T\mathbf{x} = -1 - b$. Usando esses valores em ρ, vemos que, no ponto ótimo $(\mathbf{w}_o, \mathrm{b}_o)$:

$$\rho(\mathbf{w}_o, b_o) = \frac{2}{\|\mathbf{w}_o\|}$$

  A Fig. 3 ilustra tudo isso.

- Maximizar a distância ou margem, desse modo, significa minimizar o denominador, ou seja, a norma do vetor de pesos $\|\mathbf{w}\|$. Podemos, por uma questão de tratabilidade, minimizar $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$. Essa minimização deve, evidentemente, obedecer às restrições de separabilidade linear.

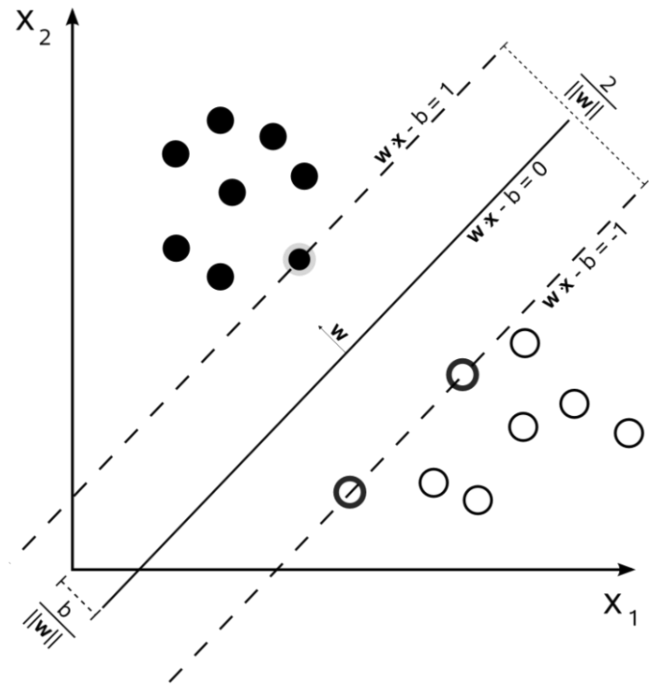

Fig. 3 – Ilustração dos Conceitos relacionados à ideia de Máxima Margem.

- Matematicamente, o problema de obtenção do classificador linear de máxima margem tem a seguinte forma:

$$
\begin{aligned}
&min_{\mathbf{w},b}\ \ \Phi = \|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} \\
&s.a.\ \ (\mathbf{w}^T\mathbf{x}_i + b)d_i \geq 1, i = 1, \dots, N
\end{aligned}
$$

- O primeiro passo para resolver esse problema de otimização com restrições é construir o seguinte lagrangiano[4]:

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i [d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

onde $\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_N]^T$ representa um vetor contendo os N multiplicadores de Lagrange.

- O próximo passo é minimizar o lagrangiano $L(\cdot)$ com respeito aos parâmetros $\mathbf{w}$ e b. Há, não obstante, uma ressalva. O lagrangiano foi construído como se as restrições fossem de igualdade, o que não é o caso. É preciso efetuar ainda a maximização de $L(.)$ com respeito aos multiplicadores de Lagrange, o que caracteriza um *problema dual* (CRISTIANINI E SHAWE-TAYLOR, 2000).

---

[4] A introdução do fator $\frac{1}{2}$ junto a $\mathbf{w}^T\mathbf{w}$ serve apenas para que a potência de dois da derivada "desapareça".

- Comecemos pelas derivadas de $L(.)$ com respeito aos pesos e *bias* do classificador:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \lambda_i d_i \mathbf{x}_i = \mathbf{0}$$

e

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \lambda_i d_i = 0$$

- A primeira equação nos diz que o vetor ótimo de pesos $\mathbf{w}_{\mathrm{o}}$ terá a seguinte forma:

$$\mathbf{w}_{\mathrm{o}} = \sum_{i=1}^{N} \lambda_i d_i \mathbf{x}_i$$

Um ponto muito interessante é que o vetor de pesos ótimo é uma combinação linear de padrões do conjunto de dados. Em outras palavras, *o vetor de pesos é* *"composto diretamente" pelos estímulos de entrada para os quais* $\lambda_i \neq 0$. Mais sobre isso em breve...

- Substituindo as derivadas nulas mostradas há pouco no lagrangiano, chega-se a uma função que depende apenas dos multiplicadores:

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\mathbf{w_o}^T \mathbf{w_o}$$

Usando a expressão obtida para $\mathbf{w_o}$, tem-se, finalmente, a seguinte função quadrática:

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

- Em notação mais compacta:

$$L(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \mathbf{1}_N - \frac{1}{2}\boldsymbol{\lambda}^T \mathbf{D} \boldsymbol{\lambda}$$

onde $\mathbf{1}_N$ é o vetor formado por N "1s" e $\mathbf{D}$ é uma matriz simétrica com elementos da seguinte forma:

$$D_{ij} = d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

- É preciso então maximizar $L(.)$ com respeito a $\boldsymbol{\lambda}$, sob duas restrições:

$$\sum_{i=1}^{N} \lambda_i d_i = \boldsymbol{\lambda}^T \mathbf{d} = 0$$

sendo $\mathbf{d}$ um vetor com os rótulos associados a todos os N dados. Essa restrição surgiu da derivada com respeito ao *bias* feita anteriormente. Também deve valer a restrição:

$$\lambda_i \geq 0, i = 1, \ldots, N$$

- Não trataremos das filigranas matemáticas, mas a solução desse problema de otimização é obtida por meio das *condições de Karush-Kuhn-Tucker (KKT)* (BISHOP, 2006). Uma dessas condições é que a seguinte igualdade vale para todos os multiplicadores:

$$\lambda_i [d_i ({\mathbf{w}_o}^T \mathbf{x}_i + b_o) - 1] = 0, i = 1, \ldots, N$$

- Cada uma dessas equações pode ser satisfeita de duas formas. Pode valer a restrição de igualdade / separação linear, ou seja:

$$d_i(\mathbf{w}_o{}^T\mathbf{x}_i + b_o) - 1 = 0$$

Nesse caso, pode-se ter $\lambda_i \neq 0$. A outra forma é mais trivial: não vale a restrição de igualdade e, portanto, $\lambda_i = 0$.

- Os pontos do conjunto de dados que satisfazem a primeira condição, que é não trivial, são chamados de **vetores-suporte** (*support vectors*). Eles são **os únicos pontos** que desempenham algum papel na definição dos pesos **w** classificador. Para que essa afirmação fique clara, vejamos novamente a expressão de $\mathbf{w}_o$:

$$\mathbf{w}_o = \sum_{i=1}^{N} \lambda_i d_i \mathbf{x}_i$$

Percebe-se que, se $\lambda_i = 0$, o dado $\mathbf{x}_i$ correspondente não entra na composição do vetor de pesos.

- Isso significa que apenas os vetores-suporte influenciam a determinação de $\mathbf{w_o}$. Desse modo, chegamos à nossa ideia inicial: encontrar um hiperplano de máxima margem…ora, apenas os pontos "limítrofes" definem a margem. A Fig. 4 revisita a ideia.

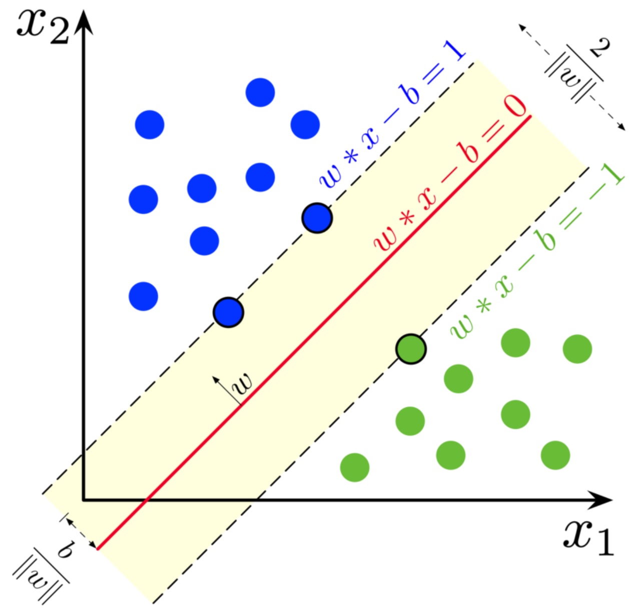

Fig. 4 – Margem e Vetores-Suporte.

- Recapitulando: de posse do conjunto de dados, é possível construir a matriz **D** e o vetor **d**. Deve-se resolver, então, o problema quadrático com restrições de otimização de $L(\boldsymbol{\lambda})$. De posse do vetor $\boldsymbol{\lambda}$, são obtidos os parâmetros do hiperplano. Não discutiremos aqui os métodos matemáticos empregados para realizar esse processo de otimização. Mais detalhes podem ser obtidos em referências como (CRISTIANINI E SHAWE-TAYLOR, 2000).

- O parâmetro de *bias* (b) poderia, em tese, ser escolhido a partir de qualquer uma das expressões seguidas pelos vetores-suporte:

$$d_i(\mathbf{w_o}^T\mathbf{x}_i + b_o) - 1 = 0$$

- Mais robusto, no entanto, é fazer uma média entre os vetores-suporte (BISHOP, 2006). Multiplicando a equação acima por d_i e fazendo a média nos vetores-suporte[5], temos:

---

[5] Note que $d_i^2 = 1$.

$$b = \frac{1}{N_S} \sum_{j \in SV} \left(d_j - \mathbf{w}_o^T \mathbf{x}_i \right)$$

onde SV representa o conjunto de índices dos vetores-suporte e N_S é a quantidade total desses vetores.

- O problema que acabamos de ver é a base da teoria de máquinas de vetores-suporte (SVMs, do inglês *support vector machines*) lineares. Essa base, entretanto, será estendida em duas direções: 1) a de abranger problemas em que a condição de separabilidade linear não vigora de maneira estrita e 2) a de incluir o caso de classificação não-linear.

2. Referências bibliográficas

BISHOP, C., *Pattern Recognition and Machine Learning*, Springer, 2006.

CORTES, C., VAPNIK, V., “Support Vector Networks”, *Machine Learning*, Vol. 20, pp. 273 – 297, 1995.

CRISTIANINI, N., SHAWE-TAYLOR, J., *Support Vector Machines and other Kernel-Based Methods*, Cambridge University Press, 2000.

VAPNIK, V., *Statistical Learning Theory*, Wiley, 1998.