SVM

Hugo Carvalho

Oferecido por:

Departamento de Métodos Estatísticos

Instituto de Matemática

Universidade Federal do Rio de Janeiro

SVM – *support-vector machines*

- Ideias iniciais: 1963 – V. Vapnik e A. Chervonenkis
- ...
- Formulação atual: 1995 – C. Cortes e V. Vapnik
- Generalização esperta de ideias simples:
  - Classificador de margem máxima
  - Classificador de vetor suporte
- Relação interessante com regressão logística e outros classificadores

Classificador de margem máxima – motivação

- Classificadores com fronteira de decisão linear:
  - LDA – modelo probabilístico em $\mathbf{X}|Y = d$
  - Regressão logística – modelo probabilístico com $\ln\left(\frac{p}{1-p}\right)$ linear em $\mathbf{X}$, sendo $p = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$

$\Longrightarrow$ Tentar aproveitar a geometria e esquecer os modelos probabilísticos!

Interlúdio – Hiperplanos em $\mathbb{R}^p$

Definição: Em $\mathbb{R}^p$, um hiperplano é um sub-espaço afim de dimensão $p-1$. Equivalentemente, é o conjunto de pontos $\mathbf{x} = (x_1, \ldots, x_p)^T$ satisfazendo

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0,$$

para parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ escolhidos

- Se $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p > 0$, então $\mathbf{x}$ está "de um lado" do hiperplano...
- ...e se $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p < 0$, então $\mathbf{x}$ está "do outro lado" do hiperplano

$\Longrightarrow$ Tem cara que dá para ser usado como um classificador!

Interlúdio – Hiperplanos em $\mathbb{R}^p$

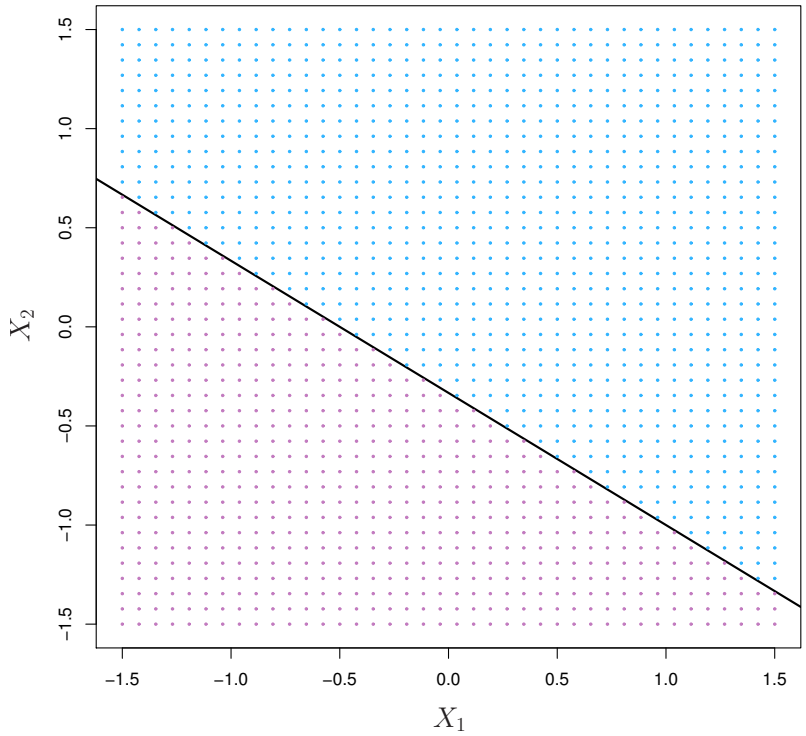

Figura 1: (Figura 9.1 de [ITSL]) Hiperplano $1 + 2x_1 + 3x_2 = 0$ em $\mathbb{R}^2$ (linha sólida). A região azul são os pontos $\mathbf{x}$ tais que $1 + 2x_1 + 3x_2 > 0$ e região roxa são os pontos $\mathbf{x}$ tais que $1 + 2x_1 + 3x_2 < 0$.

Classificador de margem máxima – construção

- Matriz $\mathbf{X}_{n\times p}$ de n observações em $\mathbb{R}^p$:

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

- Observações em duas classes: $y_1, \dots, y_n \in \{-1, 1\}$
- Nova observação $\mathbf{x}^* = (x_1^*, \dots, x_p^*)^T$ que desejamos classificar
- Assuma que as classes sejam linearmente separáveis

$\Longrightarrow$ Como escolher o "melhor" hiperplano classificador?

Classificador de margem máxima – construção

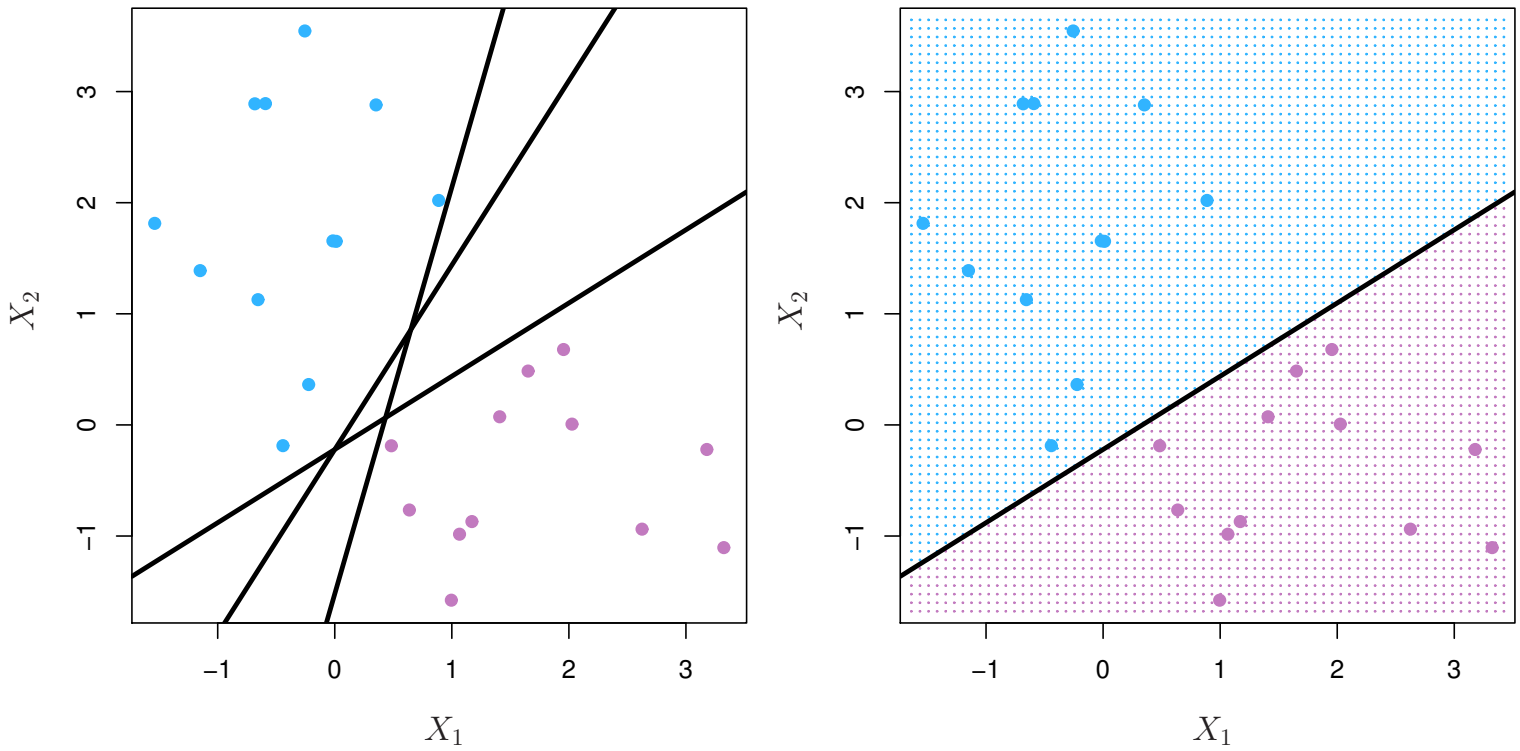

Figura 2: (Figura 9.2 de [ITSL]) Esquerda: Duas classes e três hiperplanos separadores. Direita: Um hiperplano separador e sua respectiva regra de decisão.

Classificador de margem máxima – construção

- Azul $\iff y_i = 1$ e roxo $\iff y_i = -1$
- Um hiperplano separador satisfaz

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0 \text{ se } y_i = 1$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0 \text{ se } y_i = -1$$

- Equivalentemente,

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0, \forall i = 1, \ldots, n$$

- $\implies$ Podemos construir um classificador com base no sinal de $f(\mathbf{x}^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$...
- ...e a magnitude de $f(\mathbf{x}^*)$ nos dá a "distância" ao hiperplano (confiança na classificação)

Classificador de margem máxima – construção

- Dados linearmente separáveis $\implies$ falta de unicidade no hiperplano separador!
- Solução: escolher aquele que esteja mais longe das observações de treinamento!
  - Distância de cada observação de treinamento a um dado hiperplano
  - Menor dessas distâncias: margem do hiperplano
  - Escolher o hiperplano de maior margem – aquele que tem maior distância mínima às observações de treinamento
  - Daí o nome do classificador!
- No classificador de margem máxima, as observações que atingem a margem são chamados de vetores suporte

Classificador de margem máxima – construção

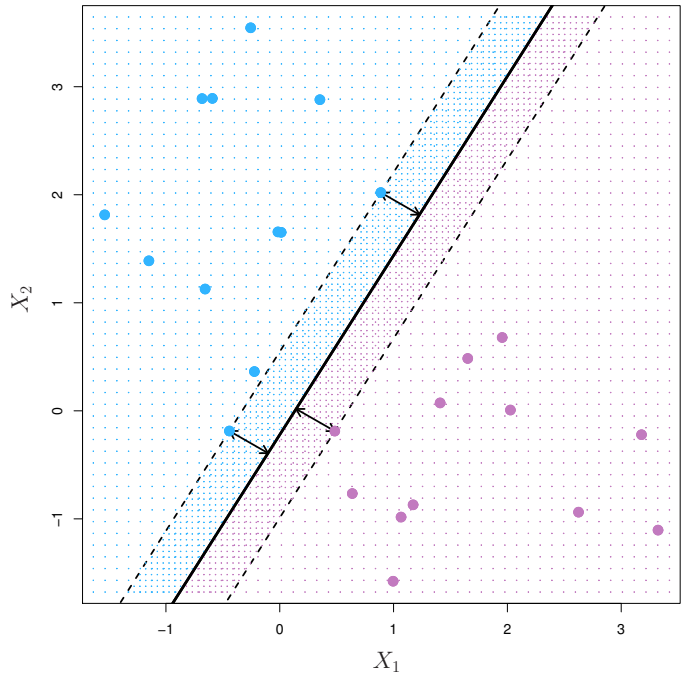

Figura 3: (Figura 9.2 de [ITSL]) Classificador de margem máxima (linha sólida), margens (linhas tracejadas) e os respectivos vetores suporte.

$\Longrightarrow$ O classificador de margem máxima depende somente dos vetores suporte, e não das outras observações!

Classificador de margem máxima – construção

Definição: O classificador de margem máxima, sob a hipótese de classes linearmente separáveis, é a solução do seguinte problema de otimização:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,M} M \quad (\star)$$

$$\text{sujeito a } \sum_{j=1}^{p} \beta_j^2 = 1 \quad (\star\star)$$

$$\text{e } y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M, \forall i = 1, \ldots, n \quad (\star\star\star)$$

- $(\star\star\star)$: Todas as observações estão do lado "certo" do hiperplano, com uma "gordurinha" $M > 0$
- $(\star\star)$: Unicidade na fórmula do hiperplano de margem máxima

$\Longrightarrow$ Se as classes não são linearmente separáveis, o problema de otimização $(\star, \star\star, \star\star\star)$ não tem solução com $M > 0$!

Classificador de margem máxima – limitação

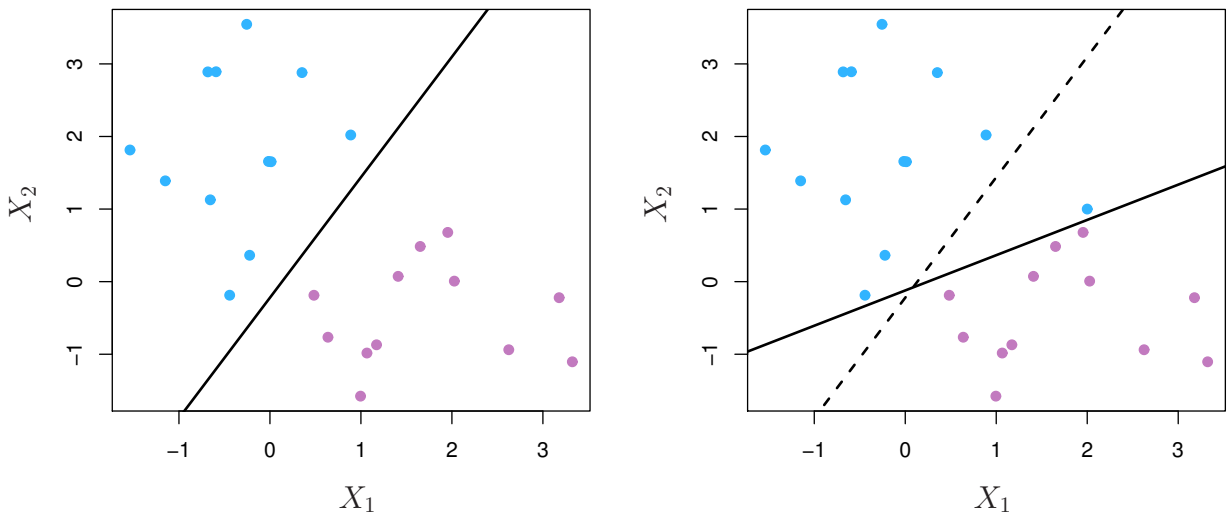

Figura 4: (Figura 9.5 de [ITSL]) Esquerda: Classificador de margem máxima. Direita: Classificadores de margem máxima antigo (linha pontilhada) e novo (linha sólida) após acrescentar uma nova observação no conjunto de treinamento.

$\Longrightarrow$ Alta sensibilidade a novas observações, possível sobreajuste!

Classificador de vetor suporte – motivação

- Pode valer a pena classificar errado algumas observações em prol de:
  - Maior robustez a observações individuais
  - Classificação mais confiável (i.e., mais distante do hiperplano separador) da *maioria* das observações

$\Longrightarrow$ Em vez de exigir que todas as observações estejam do lado "certo" da margem, permitir que algumas estejam do lado "errado" da margem, talvez até do lado "errado" do hiperplano!

Observação: Além de resolver os dois problemas acima, resolve o problema das classes não serem linearmente separáveis

Classificador de vetor suporte – motivação

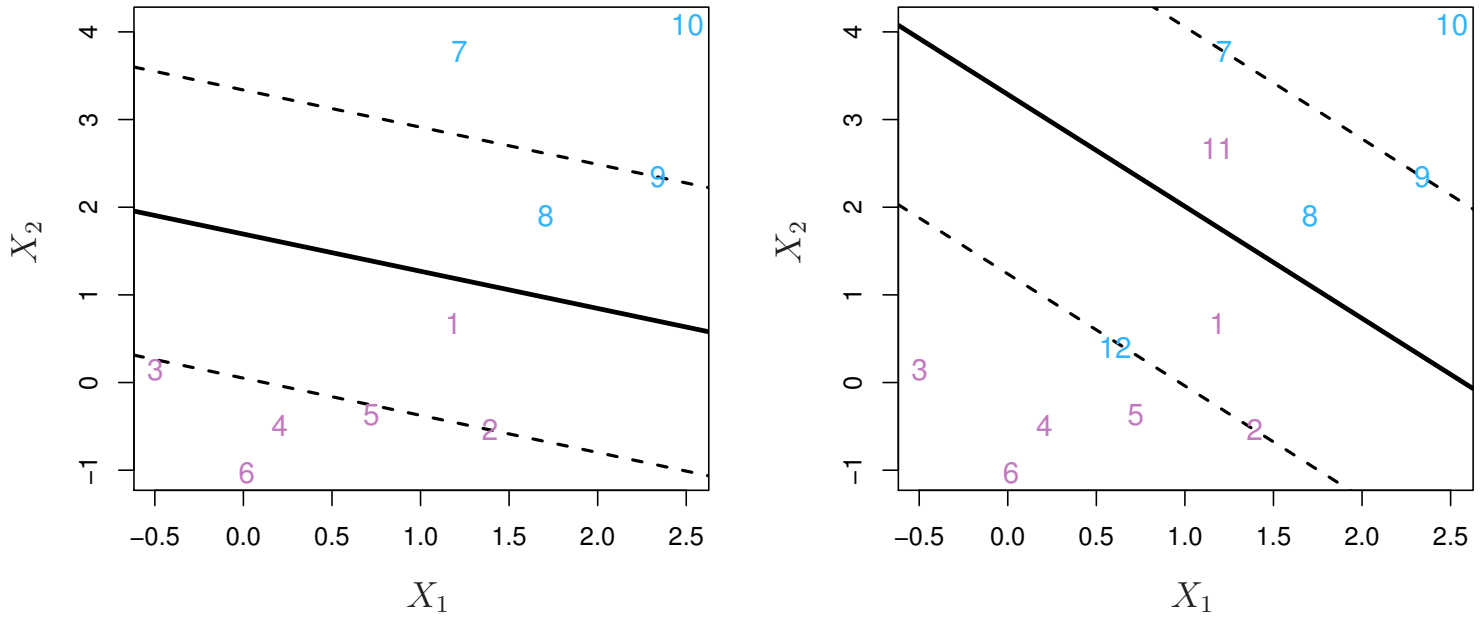

Figura 5: (Figura 9.6 de [ITSL]) Hiperplanos de margem máxima (linha sólida) e margens (linhas tracejadas). Esquerda: Algumas observações do lado "errado" da margem, porém nenhuma do lado "errado" do hiperplano. Direita: Observações do lado "errado" da margem e do hiperplano, após acréscimo de duas novas observações, 11 e 12.

Classificador de vetor suporte – construção

Definição: O classificador de vetor suporte é a solução do seguinte problema de otimização:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M} M \quad (*)$$

$$\text{sujeito a } \sum_{j=1}^{p} \beta_j^2 = 1 \quad (**)$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i), \forall i = 1, \ldots, n \quad (***)$$

$$\varepsilon_i \geq 0 \text{ e } \sum_{i=1}^{n} \varepsilon_i \leq C \quad (****)$$

- $\varepsilon_i = 0 \iff$ observação i do lado "certo" do hiperplano
- $\varepsilon_i > 1 \iff$ observação i do lado "errado" do hiperplano
- $0 < \varepsilon_i < 1 \iff$ observação i do lado "certo" do hiperplano porém do lado "errado" da margem

Classificador de vetor suporte – análise

- $(****)$: $\sum_{i=1}^{n} \varepsilon_i \leq C \iff C$ representa um orçamento para as violações do hiperplano:
  - $C = 0$ reduz ao caso do classificador de margem máxima: nenhuma violação é permitida ($\varepsilon_1 = \cdots = \varepsilon_n = 0$)
  - $C > 0 \iff$ não mais que C observações podem estar classificadas erroneamente, no pior dos casos
- $\uparrow C \implies$ mais tolerantes com classificações errôneas $\implies$ maior margem $\implies \uparrow$ viés e $\downarrow$ variância
- $\downarrow C \implies$ menos tolerantes com classificações errôneas $\implies$ menor margem $\implies \downarrow$ viés e $\uparrow$ variância
- Escolher C através de validação cruzada, por exemplo

Classificador de vetor suporte – análise

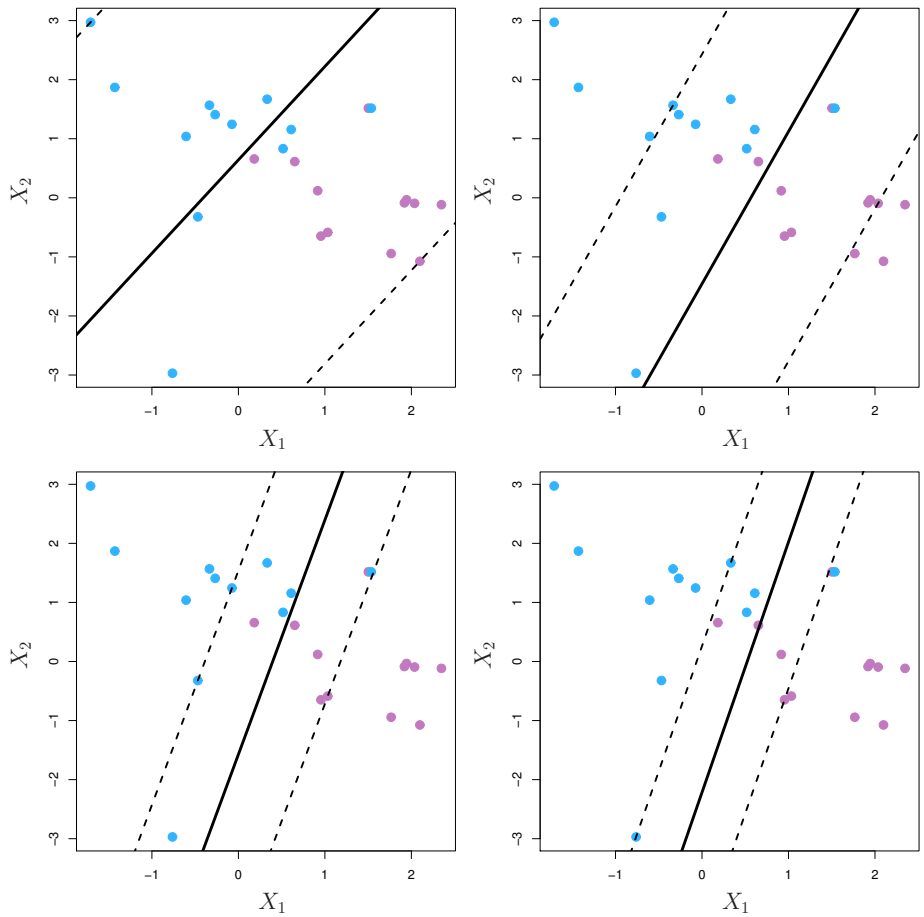

Figura 6: (Figura 9.7 de [ITSL]) Classificadores de vetor suporte ajustados com quatro valores de C distintos. Hiperplanos separadores em linha sólida e margens em linha pontilhada.

Classificador de vetor suporte – análise

- O hiperplano depende somente das observações que estão do lado "errado" da margem e do hiperplano $\implies$ vetores suporte
- De acordo com o balanço entre viés e variância:
  - $\uparrow C \implies$ margem larga $\implies$ mais vetores suporte $\implies \downarrow$ variância
  - $\downarrow C \implies$ margem estreita $\implies$ menos vetores suporte $\implies$ $\uparrow$ variância

$\implies$ Insensibilidade a observações "longe" do hiperplano separador!

Observação: Distinto de outros métodos com fronteira de decisão lineares!

- LDA depende (e é sensível) a todas as observações
- Regressão logística é pouco sensível a observações "longe" da fronteira de decisão

Classificador de vetor suporte – observação importante

Observação: O que é resolvido numericamente não são as equações $(*, **, ***, ****)$, mas sim:

$$\min_{\substack{\beta_0,\beta_1,\ldots,\beta_p,\\ \varepsilon_1,\ldots,\varepsilon_n}} \|(\beta_1,\ldots,\beta_p)\|_2 \text{ t.q. } \begin{cases} y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}) \geq 1-\varepsilon_i \\ \varepsilon_i \geq 0, \sum_{i=1}^n \varepsilon_i \leq C, \forall i=1,\ldots,n \end{cases}$$

$\implies$ Problema quadrático com restrições lineares $\implies$ problema de otimização convexa!

Equivalente a resolver o problema (multiplicadores de Lagrange):

$$\min_{\substack{\beta_0,\beta_1,\ldots,\beta_p,\\ \varepsilon_1,\ldots,\varepsilon_n}} \frac{1}{2}\|(\beta_1,\ldots,\beta_p)\|_2^2 + C\sum_{i=1}^n \varepsilon_i, \text{ t.q. } \varepsilon_i \geq 0$$

$$\text{e } y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}) \geq 1-\varepsilon_i, \forall i=1,\ldots,n$$

Classificador de vetor suporte – limitação

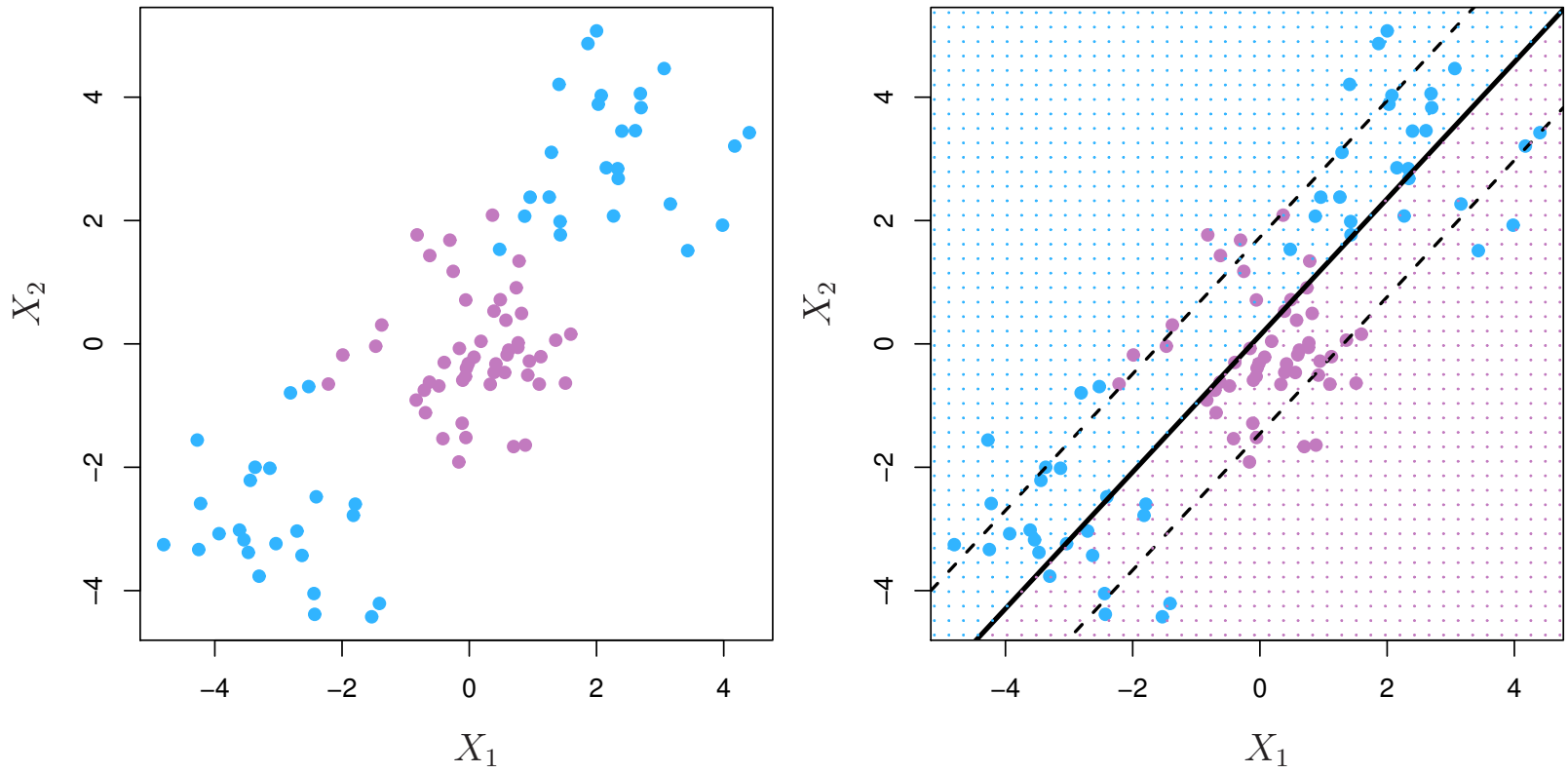

Figura 7: (Figura 9.8 de [ITSL]) Esquerda: Observações em duas classes. Direita: Tentativa de ajustar um classificador de vetor suporte a tal conjunto de dados.

Máquina de vetores suporte (SVM) – primeira ideia

Lembremos do classificador de vetor suporte:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M} M$$

$$\text{sujeito a } \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i), \forall i = 1, \ldots, n$$

$$\varepsilon_i \geq 0 \text{ e } \sum_{i=1}^{n} \varepsilon_i \leq C$$

$\Longrightarrow$ Em vez de ajustar o classificador com $X_1, \ldots, X_p$, acrescentar também $X_1^2, \ldots, X_p^2$!

Máquina de vetores suporte (SVM) – primeira ideia

$$\max_{\beta_0,\beta_{11},\ldots,\beta_{p1},\beta_{12},\ldots,\beta_{p2},\varepsilon_1,\ldots,\varepsilon_n,M} M$$

$$\text{sujeito a } \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2 = 1, \varepsilon_i \geq 0 \text{ e } \sum_{i=1}^{n}\varepsilon_i \leq C$$

$$y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right) \geq M(1-\varepsilon_i), \forall i = 1,\ldots,n$$

- No espaço aumentado $X_1,\ldots,X_p,X_1^2,\ldots,X_p^2 \implies$ fronteira de decisão linear
- No espaço original $X_1,\ldots,X_p \implies$ fronteira de decisão quadrática
- Quantidade de coeficientes a estimar aumenta muito! $\implies$ baixa eficiência computacional

Podemos tornar os dados separáveis em um espaço de dimensão infinita sem aumentar a quantidade de parâmetros a estimar!

Máquina de vetores suporte (SVM) – construção

Podemos mostrar que:

- A fronteira de decisão do classificador de vetor suporte pode ser escrita como

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle,$$

onde $\mathbf{x}_1, \ldots, \mathbf{x}_n$, são as observações de treinamento

- Para estimar $\beta_0, \alpha_1, \ldots, \alpha_n$ precisamos somente dos $\binom{n}{2}$ produtos internos $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, para $1 \leq i < j \leq n$
- Os únicos $\alpha_i \neq 0$ são os referentes aos vetores suporte:

$$f(\mathbf{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle,$$

onde $\mathcal{S}$ é o conjunto dos índices dos vetores suporte

$\Longrightarrow$ Calcular e estimar f só precisa de produtos internos!

Digressão de Álgebra Linear

- O produto interno $\langle \mathbf{x}, \mathbf{y} \rangle$ mede similaridade entre os vetores $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$
- Como podemos medir similaridade de outras formas?
  - $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R} \implies k(\mathbf{x}, \mathbf{y})$ generalização de $\langle \cdot, \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, dito um *kernel*
  - $k(\mathbf{x}, \mathbf{y}) = \left(1 + \sum_{j=1}^{p} x_j y_j\right)^d$, *kernel* polinomial de grau d
  - $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_j - y_j)^2\right)$, *kernel* radial ou exponencial quadrático
  - Obviamente $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$! *Kernel* linear
  - ... muitos outros!

Digressão de Álgebra Linear

Mas quais funções $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ são "boas" para medir similaridade entre dois vetores?

Resposta: Devem existir $\mathcal{H}$ espaço de Hilbert e $\varphi : \mathbb{R}^p \to \mathcal{H}$ tais que

$$k(\mathbf{x}, \mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_{\mathcal{H}}$$

$\Longrightarrow$ Intuitivamente, os dados devem ser linearmente separáveis em algum espaço, potencialmente de dimensão infinita.

Teorema: A função $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ é um *kernel* (i.e., existem $\mathcal{H}$ e φ como acima) se e somente se ela é positiva semidefinida:

$$\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} g(\mathbf{x}) k(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) \, d\mathbf{x} d\mathbf{y} \geq 0,$$

para toda $g : \mathbb{R}^p \to \mathbb{R}$ tal que $\int_{\mathbb{R}^p} \|g(\mathbf{x})\|^2 \, d\mathbf{x} < \infty$.

Digressão de Álgebra Linear

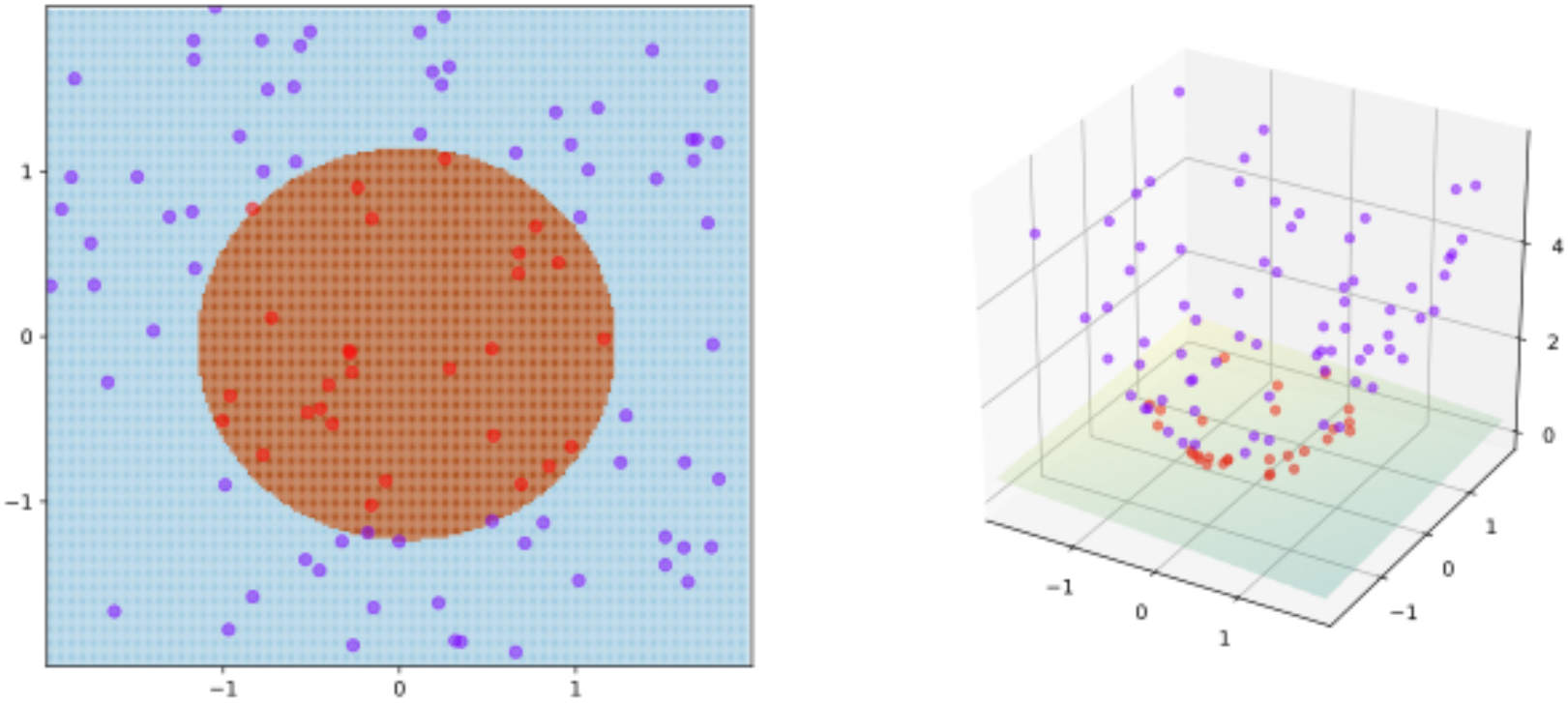

Figura 8: [Wikipedia] $\varphi : \mathbb{R}^2 \to \mathbb{R}^3$ dada por $\varphi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$ e portanto $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$. Dados não-separáveis linearmente em $\mathbb{R}^2$ mas sim em $\mathbb{R}^3$.

Máquina de vetores suporte (SVM) – construção

- Escolha o seu *kernel* preferido k
- Encontre a fronteira de decisão da forma $f(\mathbf{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$

$$\Longrightarrow \text{ Eis o SVM!}$$

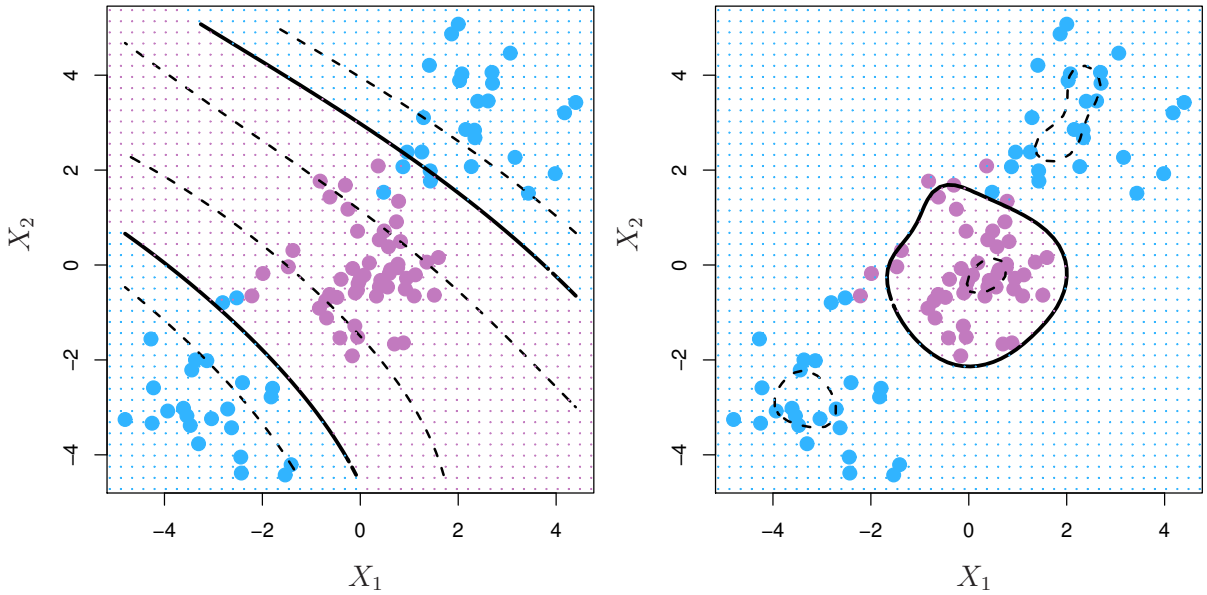

Figura 9: (Figura 9.9 de [ITSL]) Esquerda: Ajuste com *kernel* polinomial de grau 3. Direita: Ajuste com *kernel* radial.

Relação com regressão logística

Lembremos do classificador de vetor suporte (caso particular do SVM com *kernel* linear:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M} M$$

$$\text{sujeito a } \sum_{j=1}^{p} \beta_j^2 = 1, \varepsilon_i \geq 0 \text{ e } \sum_{i=1}^{n} \varepsilon_i \leq C$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i), \forall i = 1, \ldots, n$$

Esse problema de otimização pode ser escrito como

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i f(\mathbf{x}_i)] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

onde $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Relação com regressão logística

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i f(\mathbf{x}_i)] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\Downarrow$$

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \{L(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\},$$

onde L é uma função custo e P é uma penalidade

- A função custo mede o ajuste do modelo aos dados
- A penalidade impõe uma restrição desejada *a priori* no modelo
- O modelo de regressão logística também pode ser escrito nessa forma

Relação com regressão logística

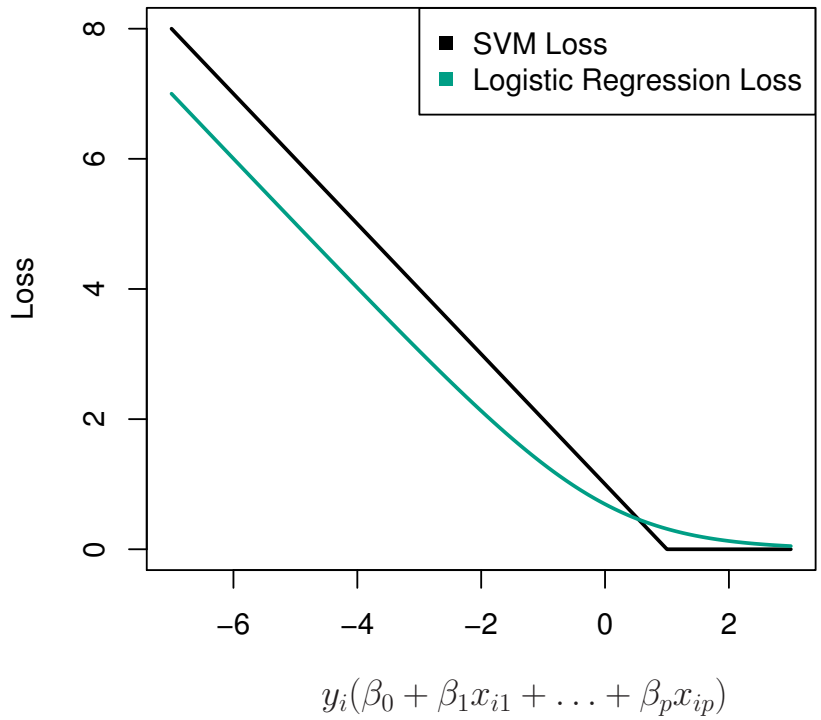

Figura 10: (Figura 9.12 de [ITSL]) Funções perda associadas à regressão logística (verde) e classificador de vetor suporte (preto), como função de $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$.

Relação com regressão logística

- Na formulação de custo + penalidade, a margem do classificador de vetor suporte corresponde ao valor 1, e o seu tamanho é dado por $\sum_{j=1}^{p} \beta_j^2$
- Observação do lado "certo" da margem $\iff y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq 1 \iff$ perda do classificador de vetor suporte é zero
- A perda da regressão logística nunca é zero! Porém, é próxima de zero quando a perda do classificador de vetor suporte é zero!
- Além disso, ambas as perdas são próximas para observações do lado "errado" da margem

$\implies$ Ambas as técnicas dão resultados semelhantes!