SVM (Support Vector Machines)

Máquinas* de Vetores-*Suporte

Evelyn Perez Cervantes

Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

Antiga União Soviética

CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

Editor: Lorenza Saitta

Abstract. The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data.

High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

Keywords: pattern recognition, efficient learning algorithms, neural networks, radial basis function classifiers, polynomial classifiers.

Support-vector networks

C Cortes, V Vapnik - Machine learning, 1995 - Springer

The support-vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed ...

A note on a class of perceptrons

Vapnik–Chervonenkis (VC) dimension

Aprendizado com estatística

↓ reconhecimento de padrões

↓ Inteligência artificial

Como resolver um problema de classificação de duas classes de forma directa?

Tentamos encontrar um plano que separa as classes no espaço de características.

E se um plano não consegue separar? o que podemos fazer?

- Nós suavizamos o que queremos dizer com "separar", e
- Enriquecemos e ampliamos o espaço de características para que a separação seja possível.

O que é um hiperplano?

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

$$\beta = (\beta_1, \beta_2, \cdots, \beta_p)$$

Em particular num espaço de duas dimensões (***p*=2**) o hiperplano é uma linha.

No espaço tridimensional o hiperplano é um plano bidimensional

Um hiperplano num espaço de ***p*** dimensões é um sub-espaço de ***p*-1** dimensão.

Como separar com um hiperplano?

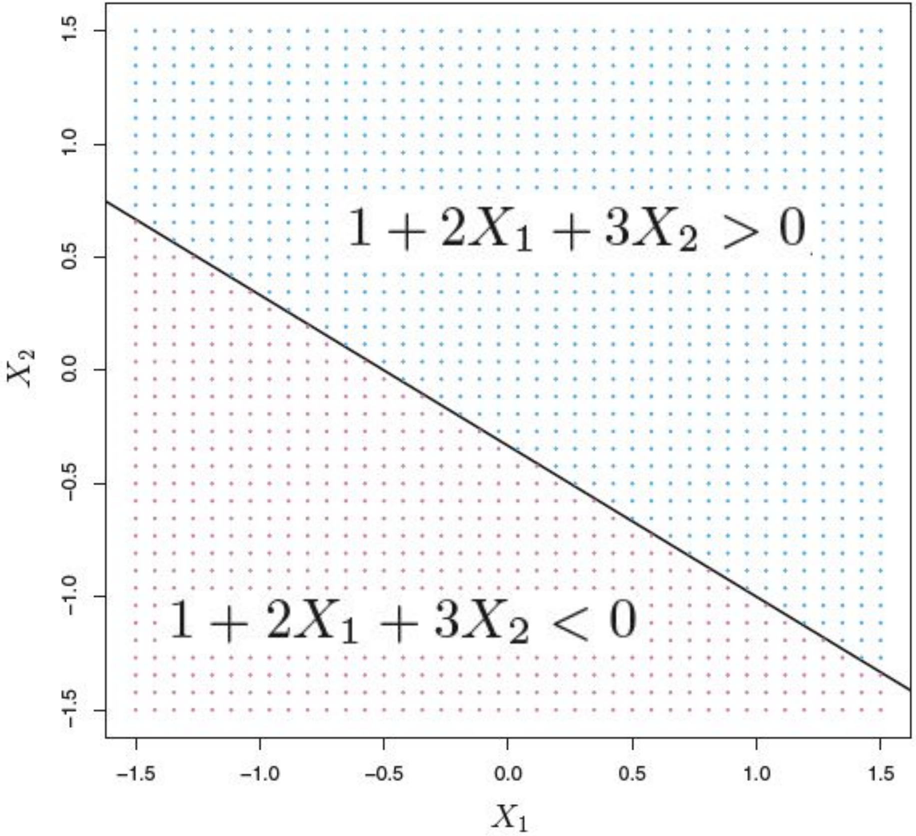
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0,$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0.$$

Se $X = (X_1, X_2, \ldots, X_p)^T$ cumpre com a equação, então pertence ao hiperplano, caso contrário ele não pertence ao hiperplano.

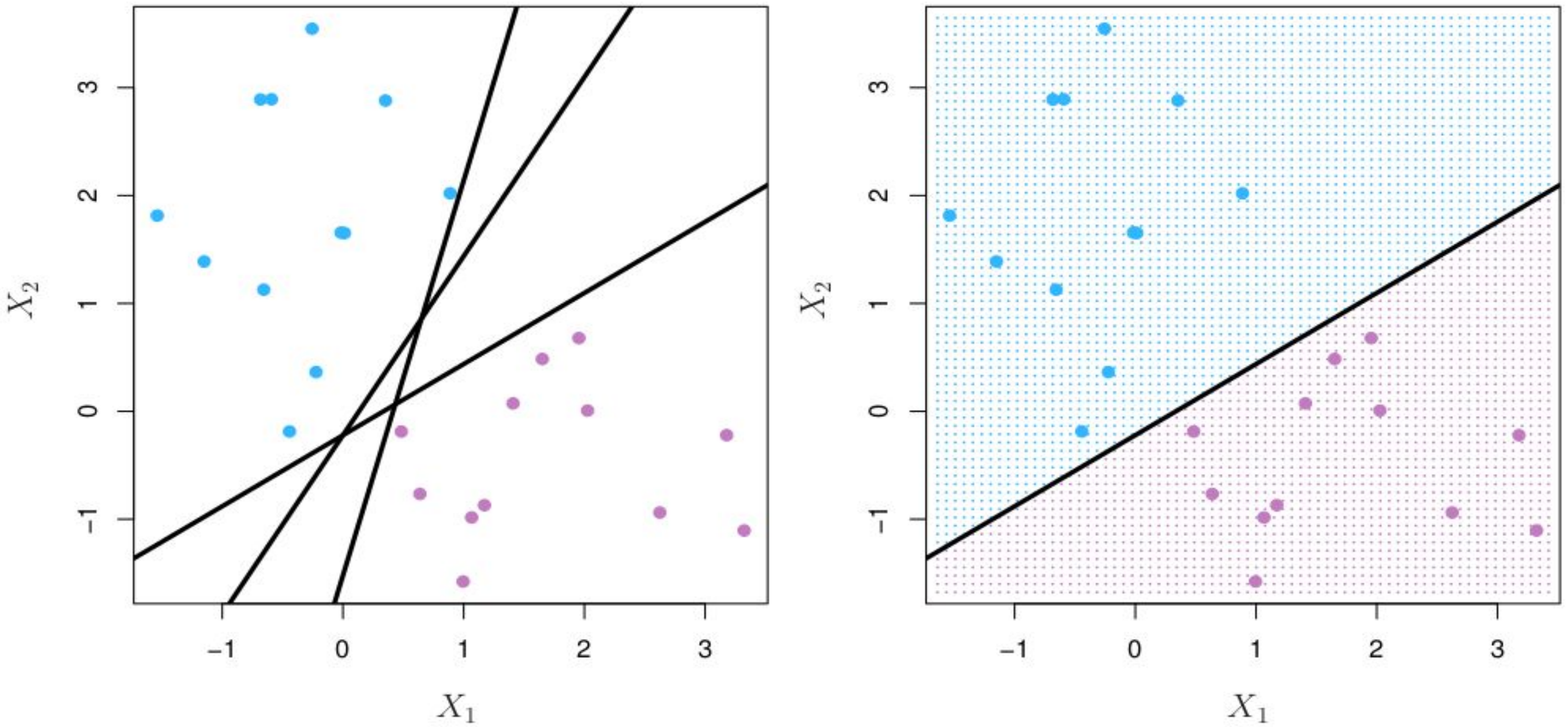
Como separar com um hiperplano?



$1 + 2X_1 + 3X_2 = 0$

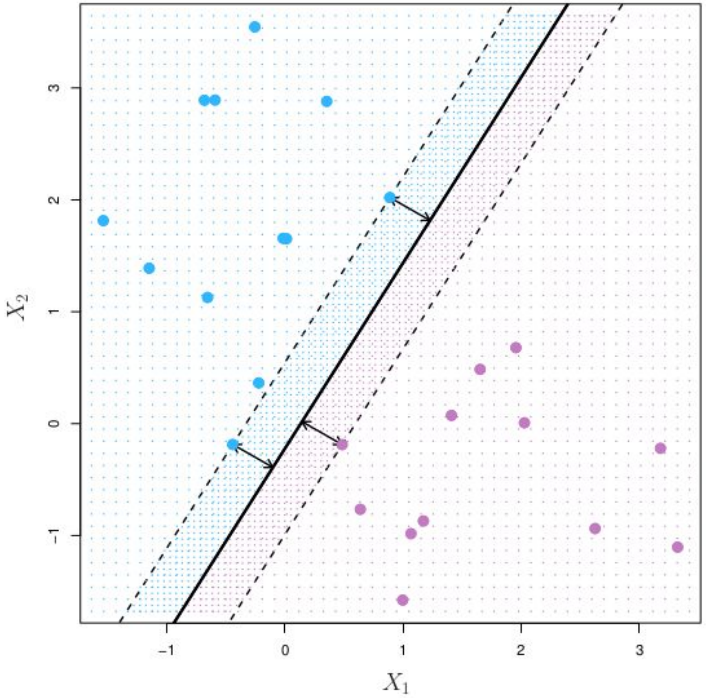
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Como separar com um hiperplano?



Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Classificador de Margem Máxima



Constrained optimization problem

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\; M$$

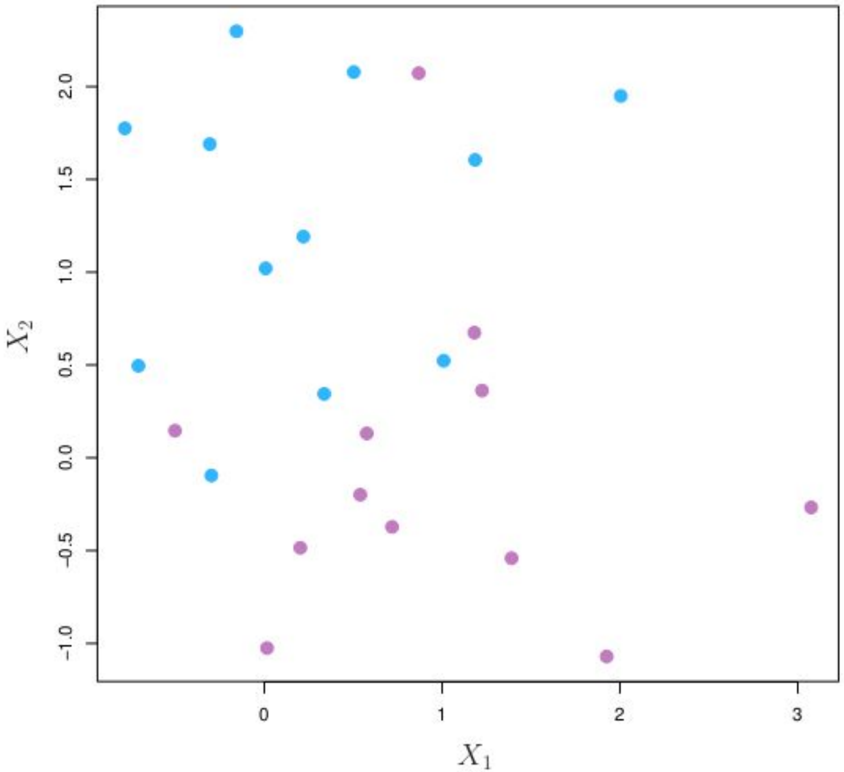
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

$$\text{for all } \; i = 1, \ldots, N.$$

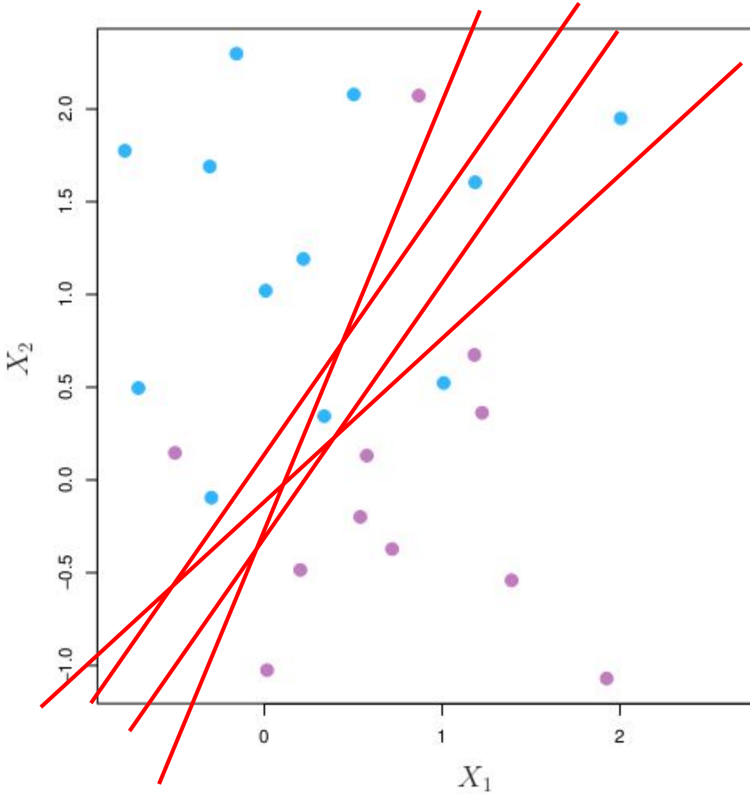
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Datos no separables, o que fazer?



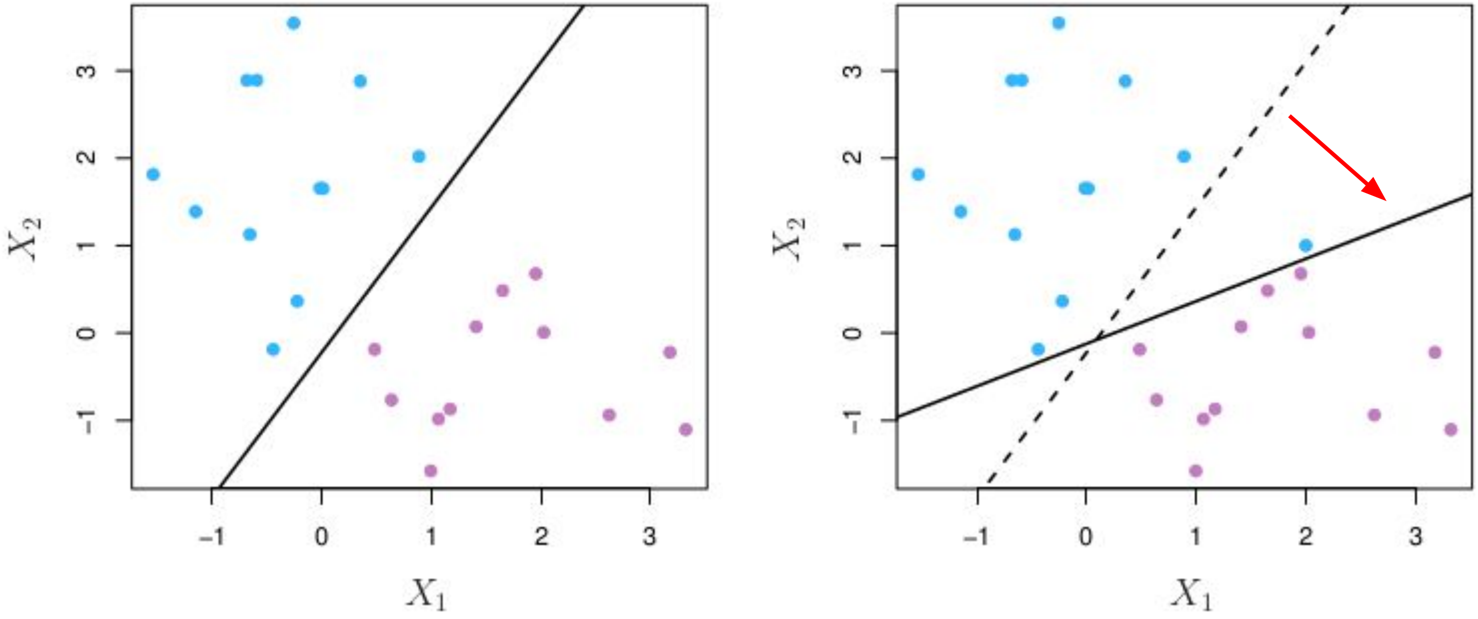
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Datos no separables, o que fazer?



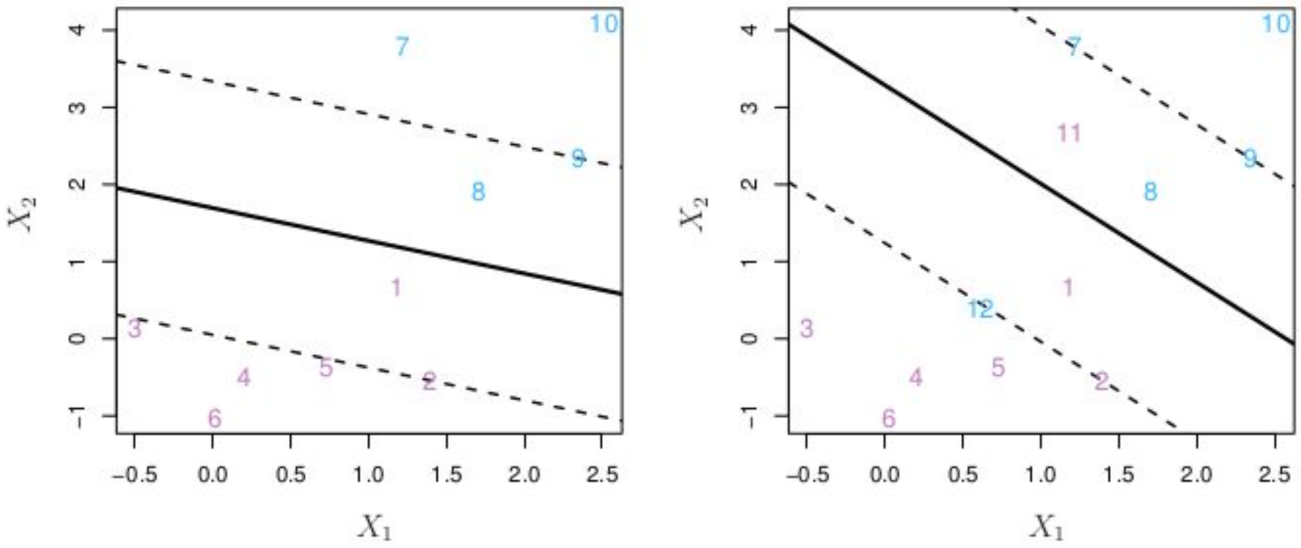
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Datos separables ... mas com ruído



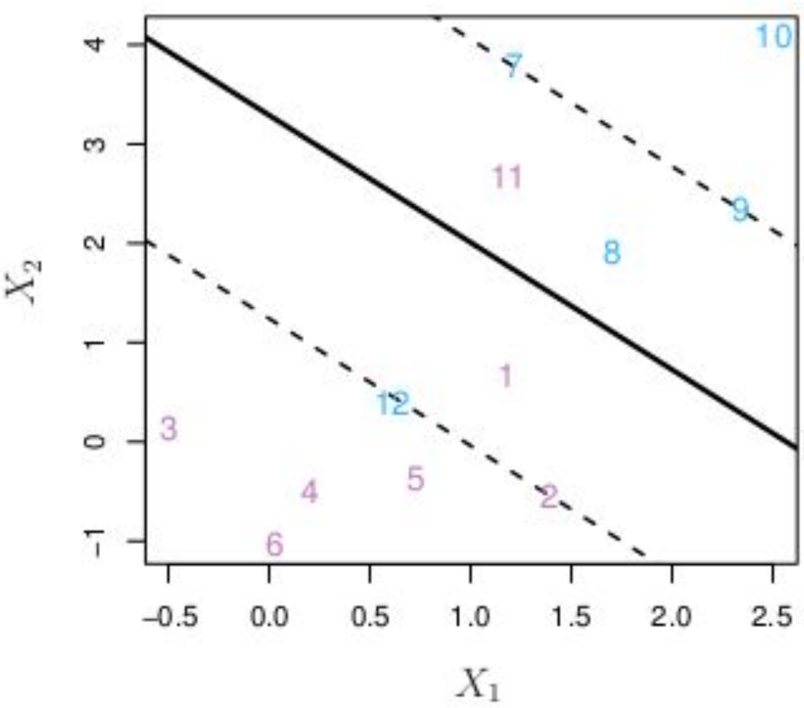
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Classificador de vetor de suporte



Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Classificador de vetor de suporte



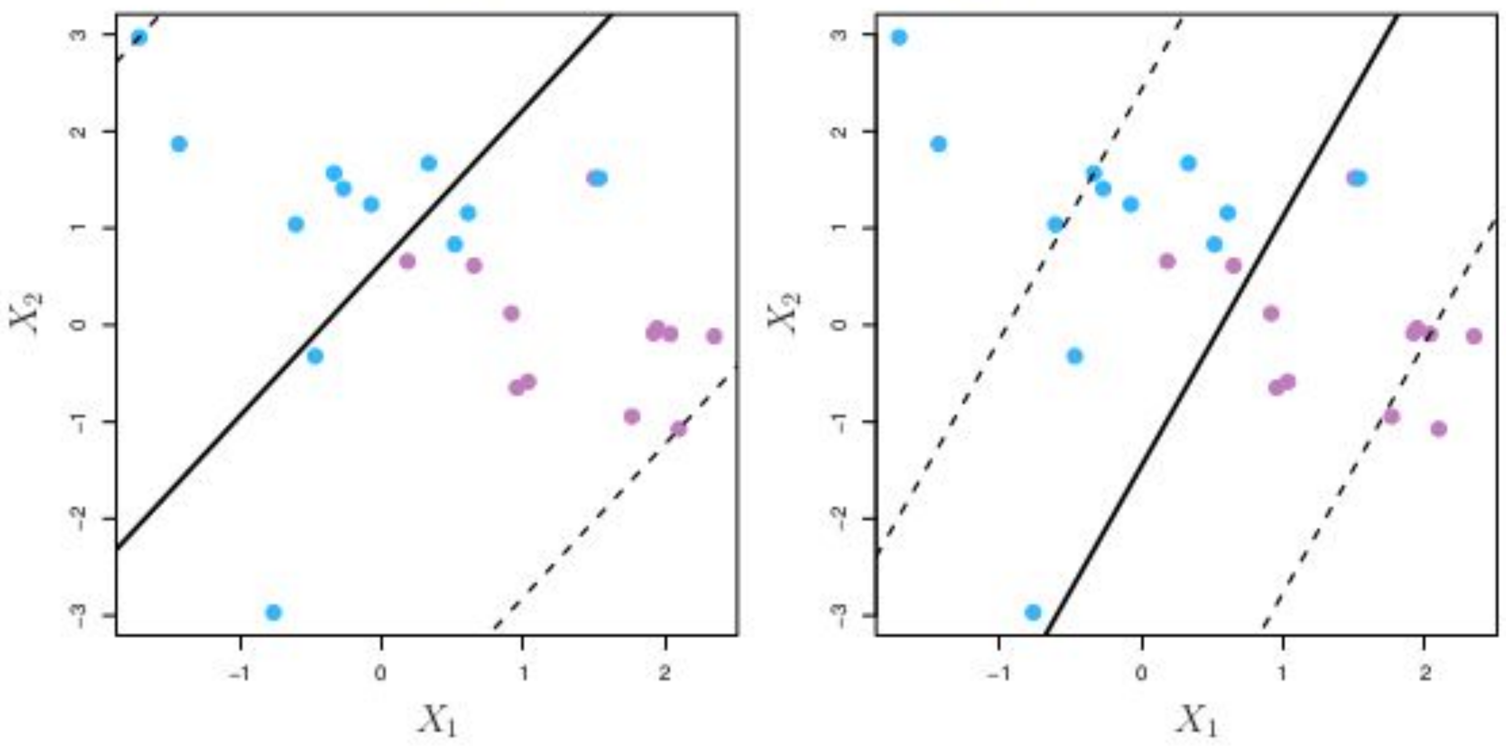
$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \ \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i \leq C,$$

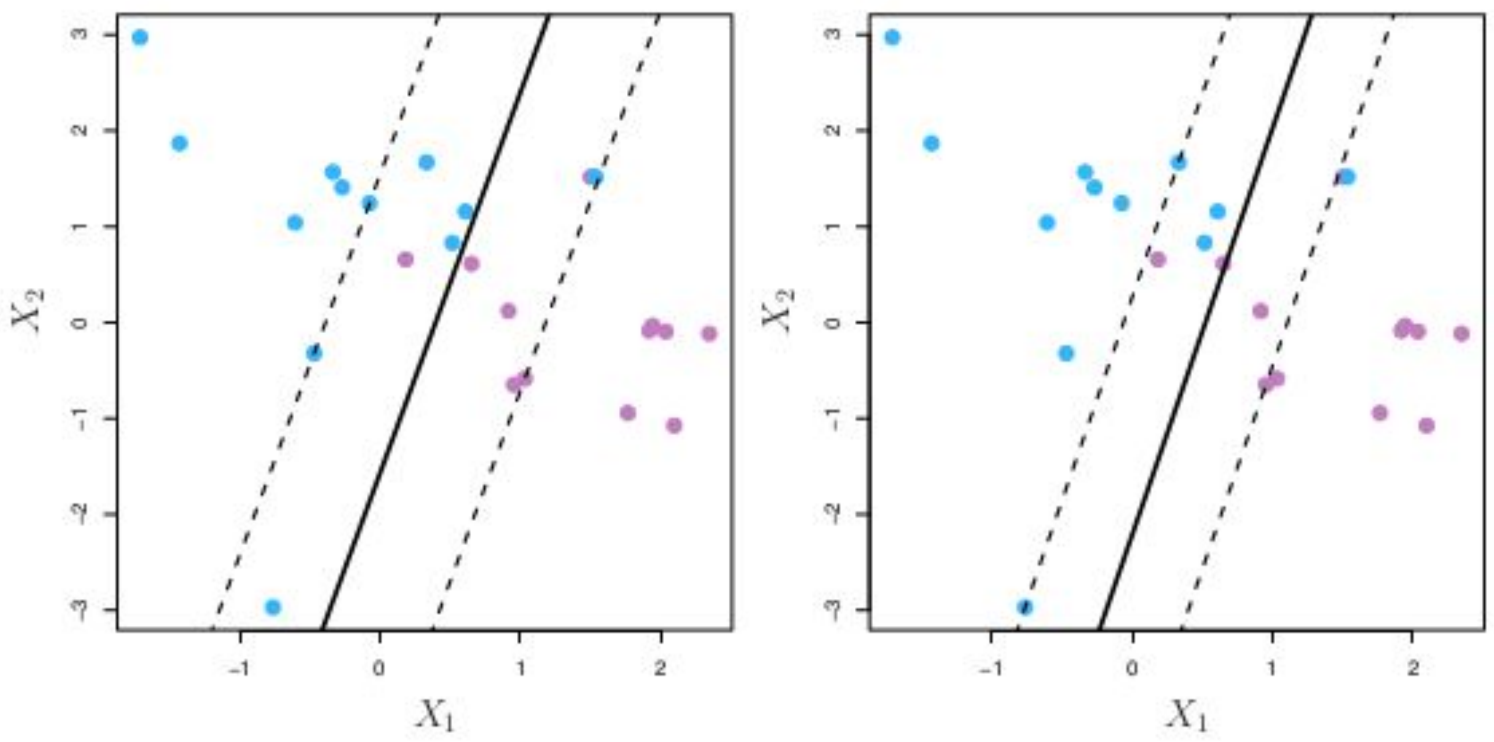
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Parâmetro C



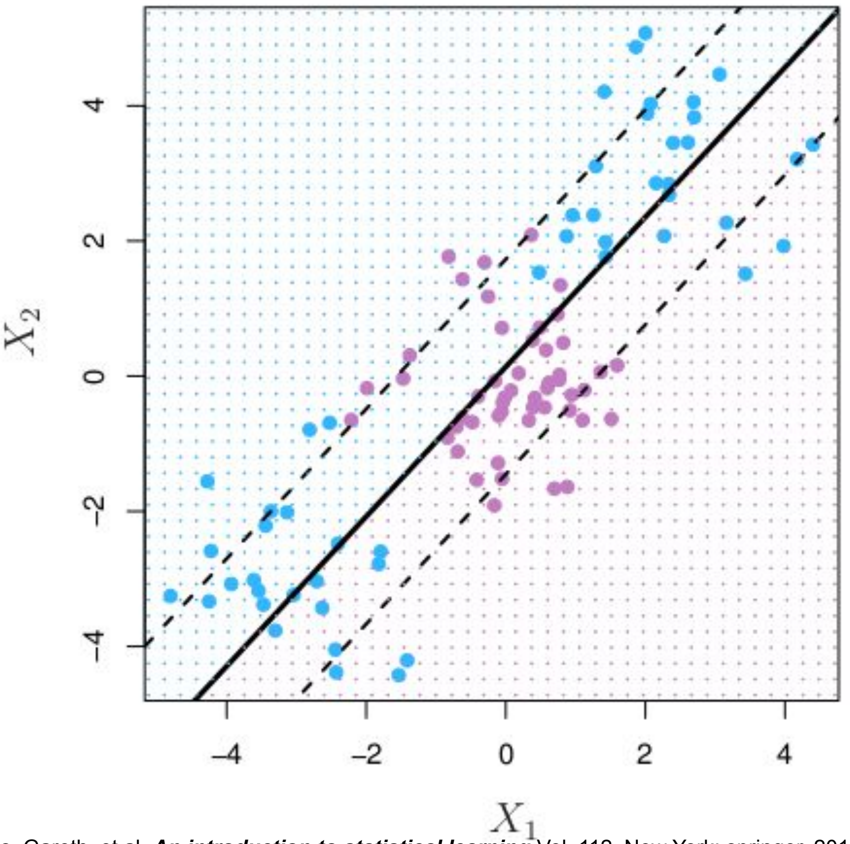
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Parâmetro C



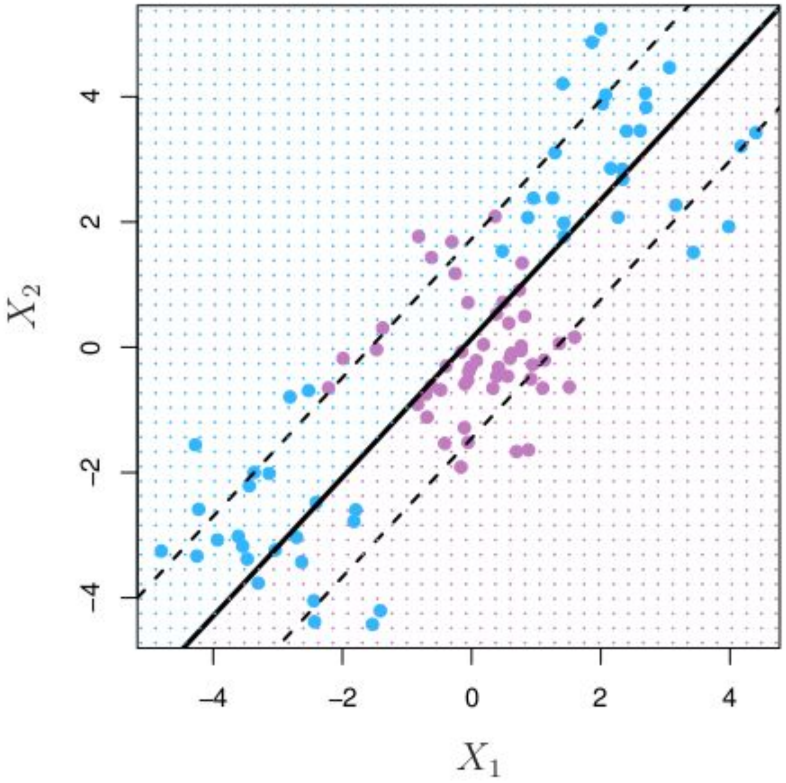
Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

e quando o parâmetro C não dá conta?

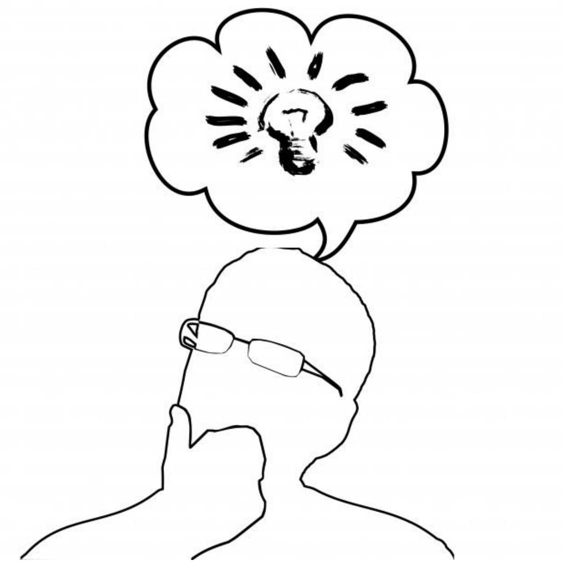


Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

e quando o parâmetro C não dá conta?

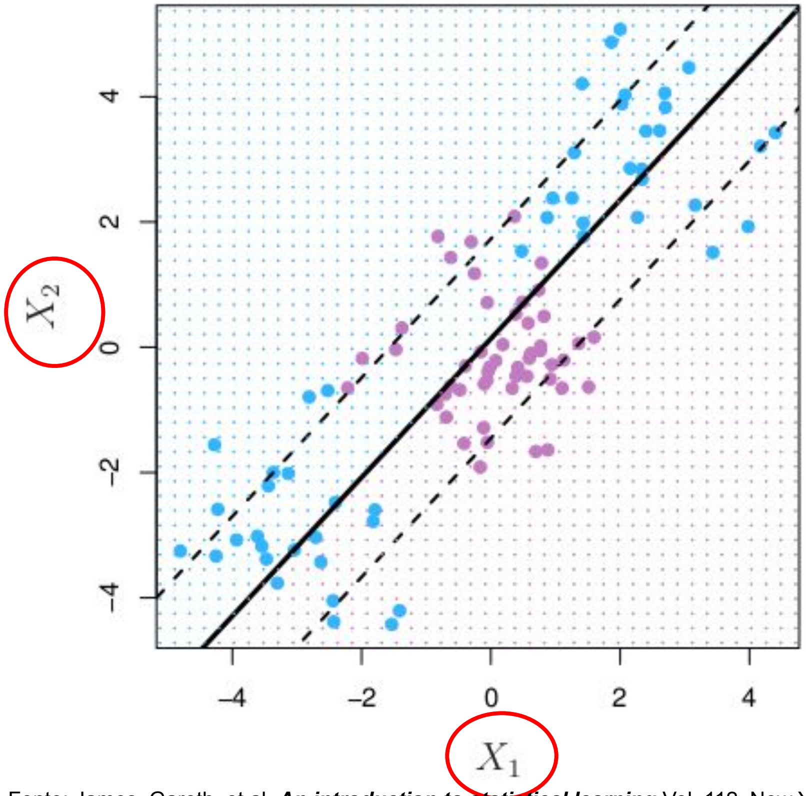


Expansão de características



Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Expansão de características



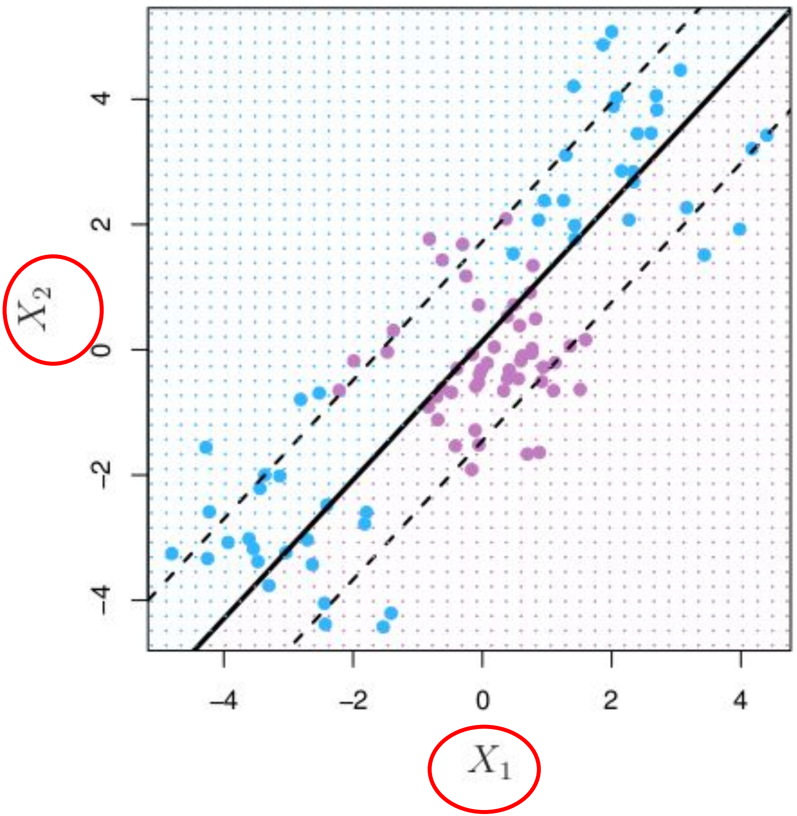
podemos combinar de alguma maneira as características que temos.

$$X_1^2,\ X_1^3,\ X_1X_2,\ X_1X_2^2, \ldots$$

De maneira tal que vamos nos mover de um espaço de duas dimensões ***(p=2)*** para um outro de ***q*** en que ***q > p***

Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Expansão de características



$$(X_1,\ X_2,\ X_1^2,\ X_2^2,\ X_1X_2)$$

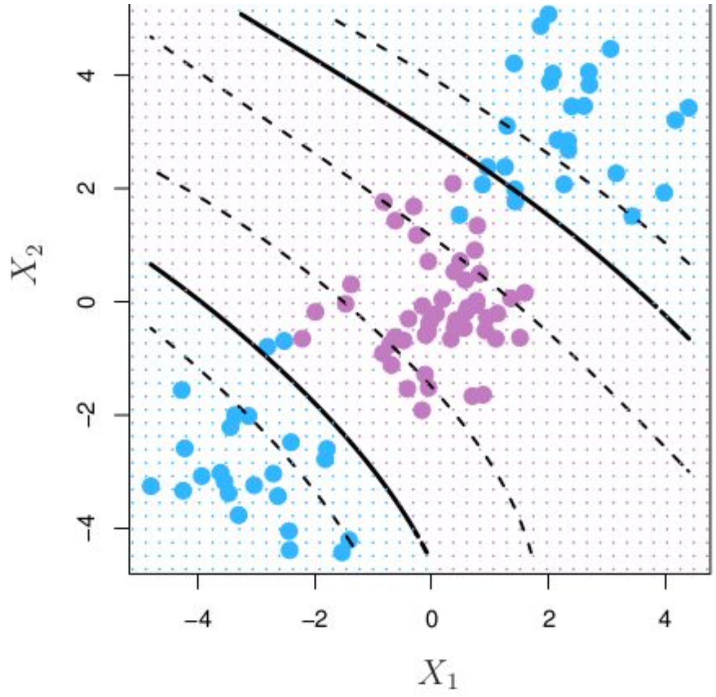
Então trabalhando no novo espaço de 5 dimensões o novo hiperplano de separação estaria definido da seguinte maneira:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0$$

Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

Expansão de características

$$\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1^2+\beta_4X_2^2+\beta_5X_1X_2+\beta_6X_1^3+\beta_7X_2^3+\beta_8X_1X_2^2+\beta_9X_1^2X_2=0$$



O problema do espaço de duas dimensões é resolvido no espaço expandido, e se projetamos a resposta no espaço de duas dimensões resulta num limite de separação não linear, neste caso dois segmentos

Fonte: James, Gareth, et al. ***An introduction to statistical learning*** Vol. 112. New York: springer, 2013

SVM e produto interno

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \quad \textit{— inner product between vectors}$$

Um classificador linear de vetores de suporte pode ser representado utilizando o produto interno

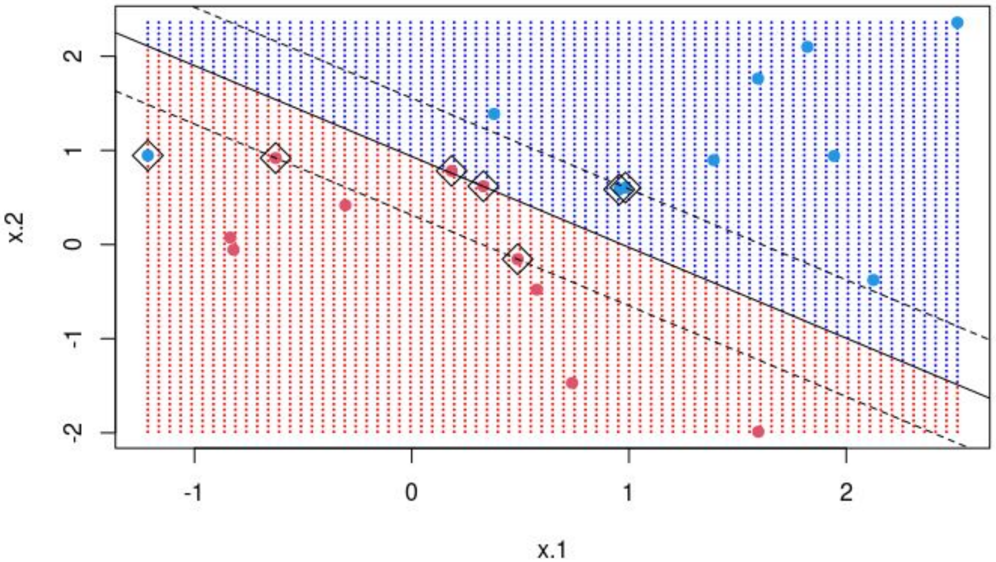
$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \quad \textit{— n parameters} \qquad \binom{n}{2}$$

A maioria dos $\hat{\alpha}_i$ é 0 pelo qual ficamos com a seguinte fórmula em q S é o conjunto de índices em que $\hat{\alpha}_i > 0$

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

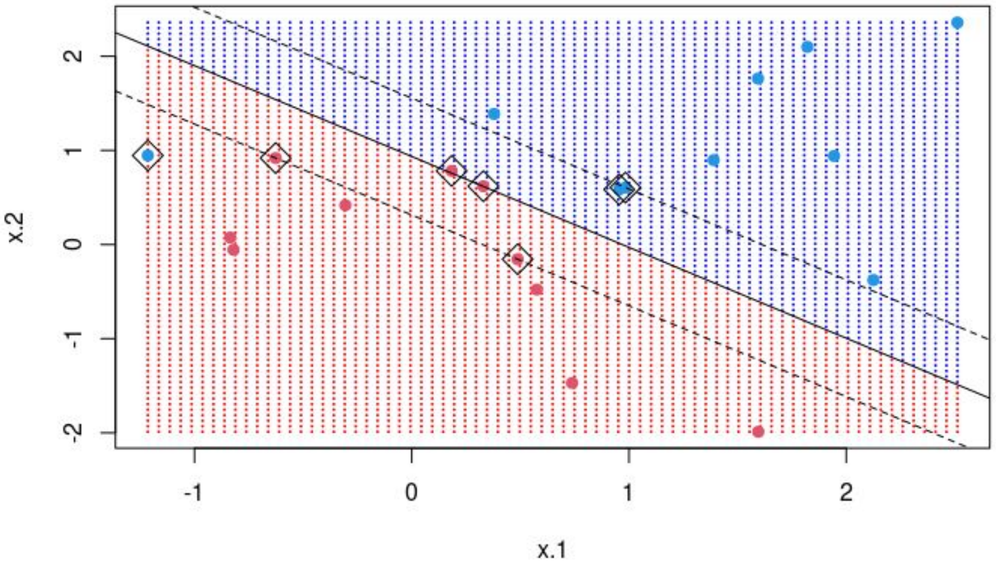
SVM e produto interno

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



SVM e produto interno

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



SVM e Kernels

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

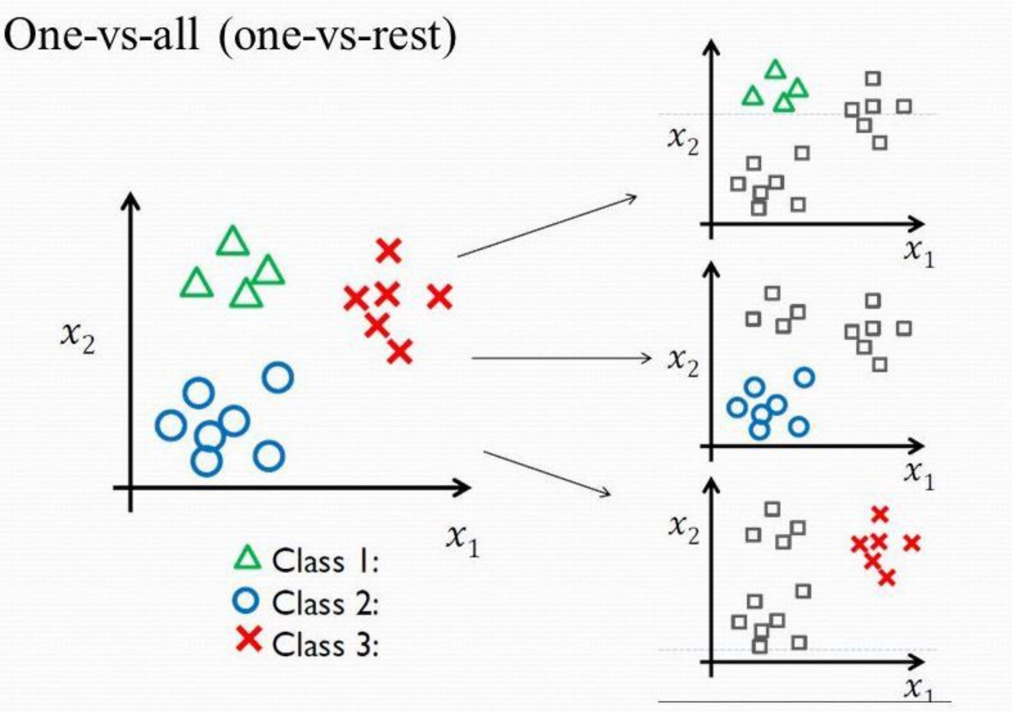
Em resumo se conseguimos calcular o produto interno entre nossas observações*, podemos treinar um classificador de vetores de suporte.
Mas existem outras funções que podem fazer isso e são chamadas de funções kernel.

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

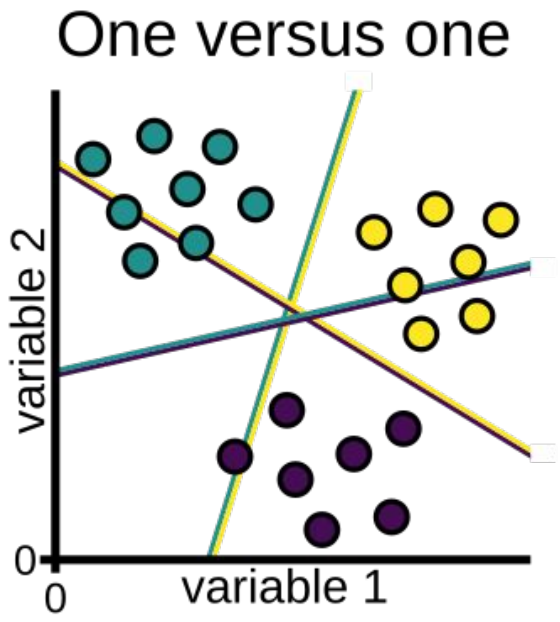
$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i K(x, x_i).$$

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

SVM com mais de duas clases



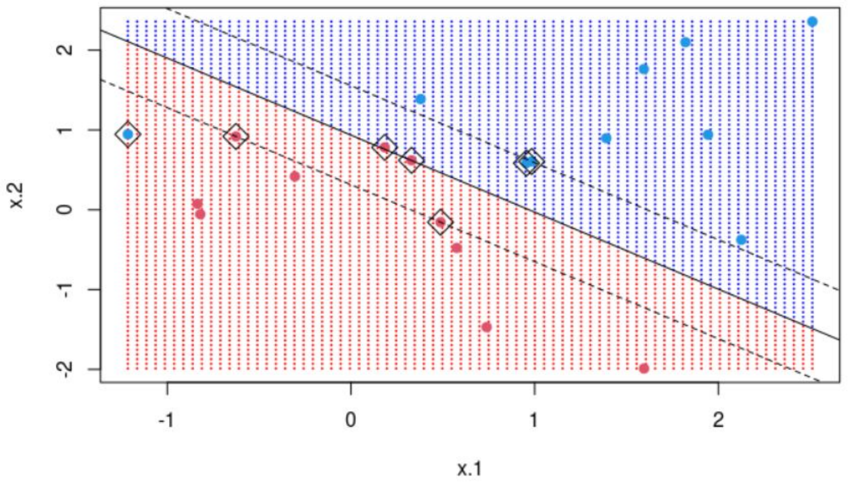
Fonte: https://medium.com/@b.terryjack/tips-and-tricks-for-multi-class-classification-c184ae1c8ffc



Fonte: https://www.r-bloggers.com/2019/10/support-vector-machines-with-the-mlr-package/

LAB

```{r}
beta=drop(t(svmfit$coefs)%*%x[svmfit$index,])
beta0=svmfit$rho #negative intercept
plot(xgrid,col=c("red","blue")[as.numeric(ygrid)],pch=20,cex=.2)
points(x,col=y+3,pch=19)
points(x[svmfit$index,],pch=5,cex=2)
abline(beta0/beta[2],-beta[1]/beta[2])
abline((beta0-1)/beta[2],-beta[1]/beta[2],lty=2)
abline((beta0+1)/beta[2],-beta[1]/beta[2],lty=2)
```



Obrigado!

Contato: Evelyn P Cervantes

epcervantes7@gmail.com

https://www.linkedin.com/in/epcervantes/

Fonte: **Diego Delso** Creative Commons Atribuição-CompartilhaIgual 4.0 Internacional.

Principais referências

- James, Gareth, et al. ***An introduction to statistical learning***. Vol. 112. New York: springer, 2013.
- Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.
- **Statistical Learning Course (Trevor Hastie** and Rob Tibshirani) https://web.stanford.edu/~**hastie**/MOOC-Slides/svm.pdf
- http://faculty.marshall.usc.edu/gareth-james/ISL/code.html
- https://altaf-ali.github.io/ISLR/chapter9/lab.html