Introduction to statistical learning

3.7 Supervised learning: SVM

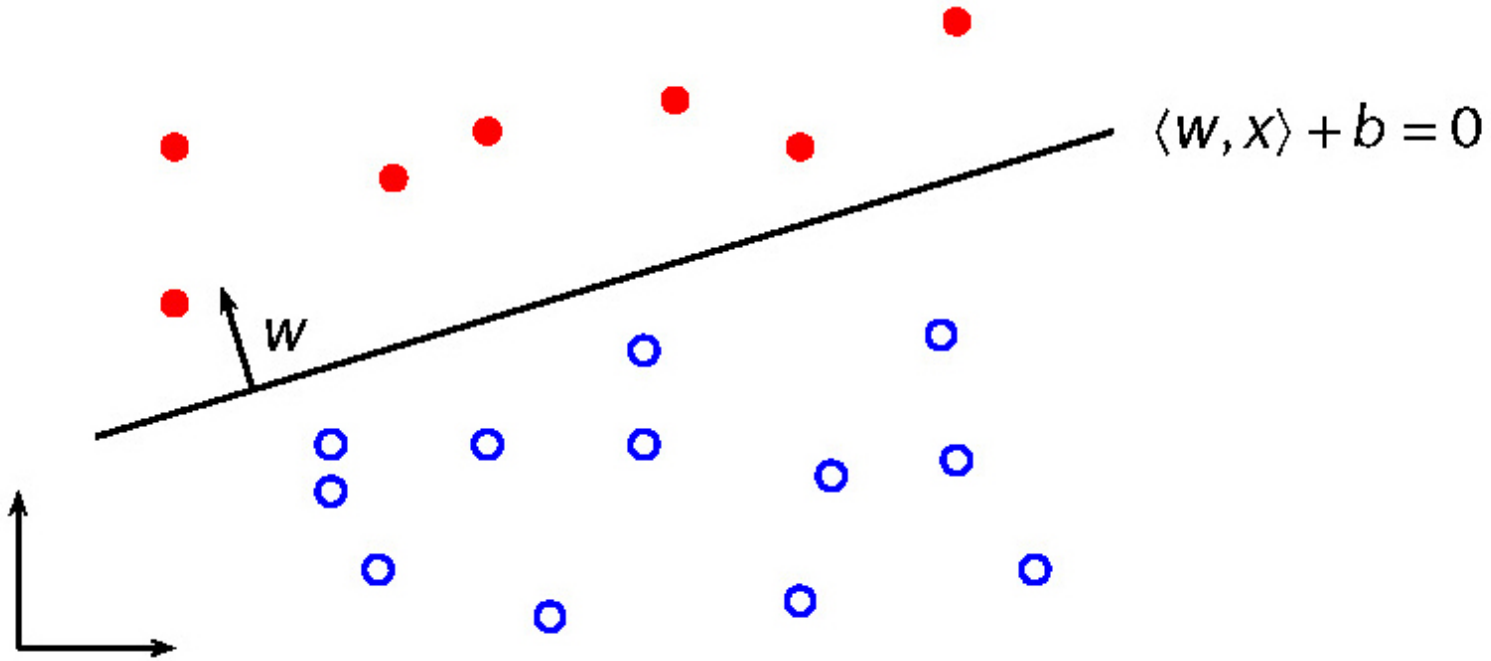
V. Lefieux

June 2018

General points

- SVM (Support Vector Machine) are machine learning algorithms for classification (or regression) problems.
- First idea: find a plane that separates classes in the feature space.

Linearly separable data I



Linearly separable data II

Data are linearly separable if there exists $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that :

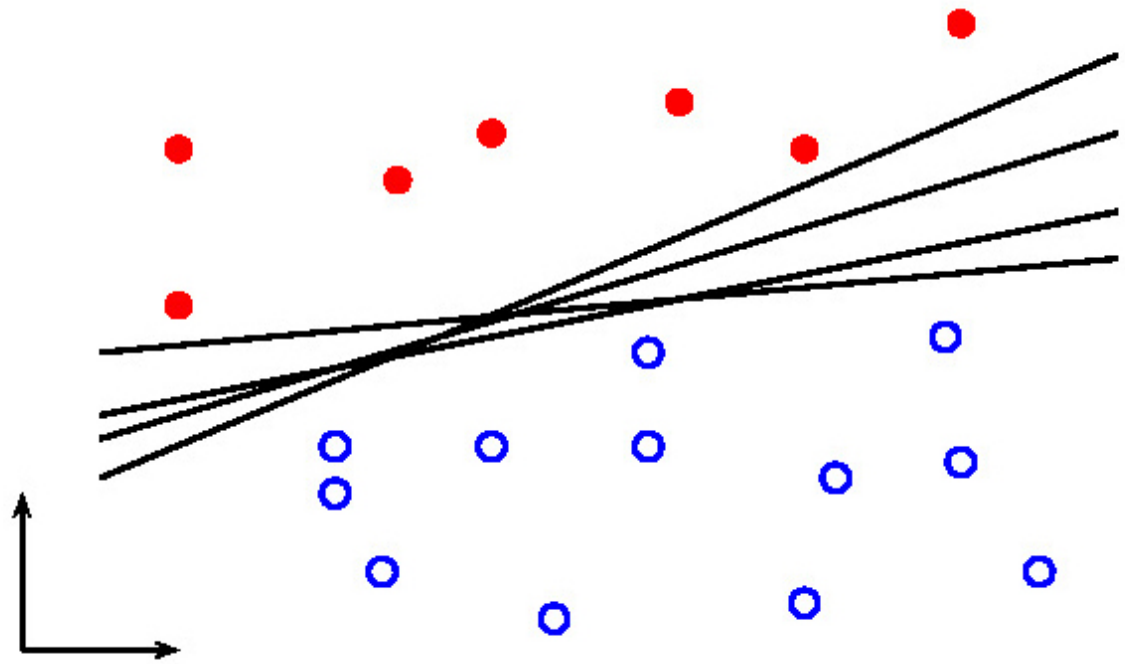
$$\begin{cases} y_i = 1 & \text{if } \langle w_i, x_i \rangle + b > 0 \\ y_i = -1 & \text{if } \langle w_i, x_i \rangle + b < 0 \end{cases}$$

or in an equivalent manner:

$$y_i \left(\langle w_i, x_i \rangle + b \right) > 0 \ .$$

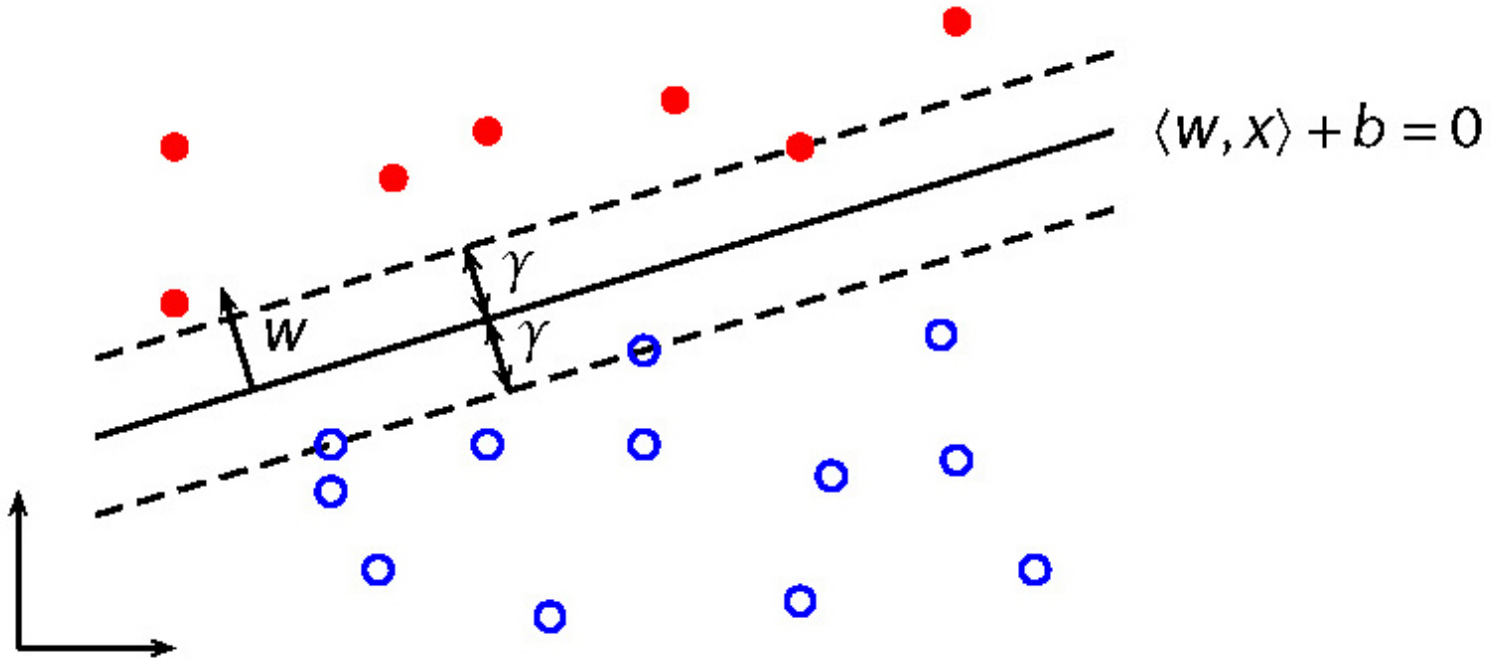
$\langle w_i, x_i \rangle + b = 0$ is the equation of a hyperplane.

A lot of possible separating hyperplanes



Vapnik solution

Among all separating hyperplanes, find the one which maximizes margin γ between classes.



The margin is the distance between the separation boundary and the nearest samples which are called support vectors.

Maximal margin classifier optimization problem I

Maximizing γ is equivalent to:

$$\begin{aligned} &\min \|w\|^2 \\ &\text{subject to } \forall i \in \{1, \ldots, n\} : y_i \left(\langle w_i, x_i \rangle + b \right) \geq 1 \ . \end{aligned}$$

Maximal margin classifier optimization problem II

$$\max \gamma$$
$$\text{subject to } \begin{cases} \sum_{j=1}^{p} w_j^2 = 1 \ , \\ y_i \left(\langle w_i, x_i \rangle + b \right) \geq \gamma \end{cases}$$

where C is a regularization parameter.

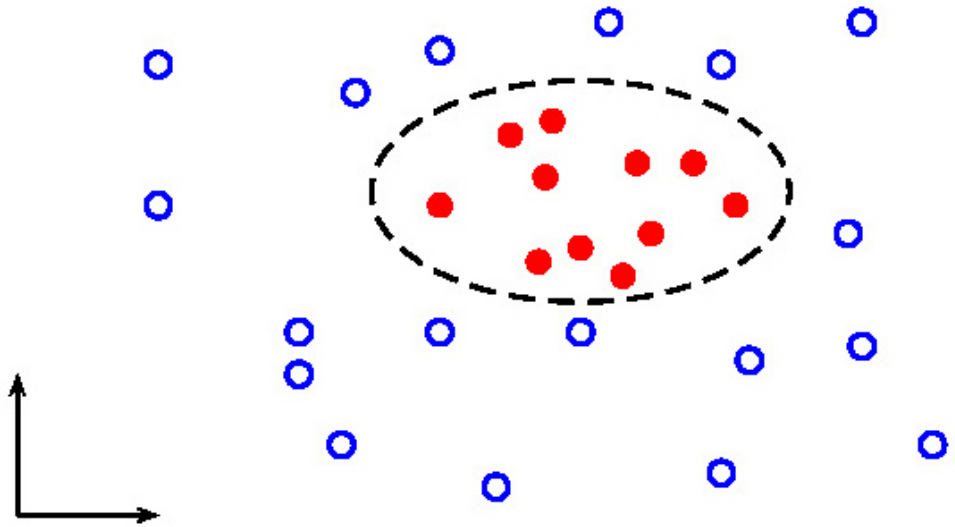
Support vector classifier

The support vector classifier maximizes a soft margin.

$$\max \gamma$$

$$\text{subject to } \begin{cases} \sum_{j=1}^{p} w_j^2 = 1 \ , \\ y_i \left(\langle w_i, x_i \rangle + b \right) \geq \gamma \left(1 - \xi_i \right) \ , \\ \xi_i \geq 0 \ , \\ \sum_{i=1}^{n} \xi_i \leq C \end{cases}$$

In the case of non-separable data I



In the case of non-separable data II

- Sometimes doesn't work, whatever the value of C.
- Enlarge the space of features by including transformations, e.g. X_1, X_2, X_1^2, X_2^2.
- Fit a support-vector classifier in the enlarged space (non-linear decision boundaries in the original space).
- One can use polynomials, more generally one use kernels: polynomial, radial, gaussian,...

References

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An introduction to statistical learning with applications in R*. Springer Texts in Statistics. Springer.