Intro ML
ML. 6. Support vector machines

DataLab CSIC

Objectives and schedule

Introduce key concepts about SVMs. A case of non-probabilistic ML approach (now made probabilistic). Kernel methods

Contents

- Maximal margin classifier
- Support Vector Classifier
- Support Vector Machines
- Extensions: More than two classes, SVM for reg, SVM vs LR,…

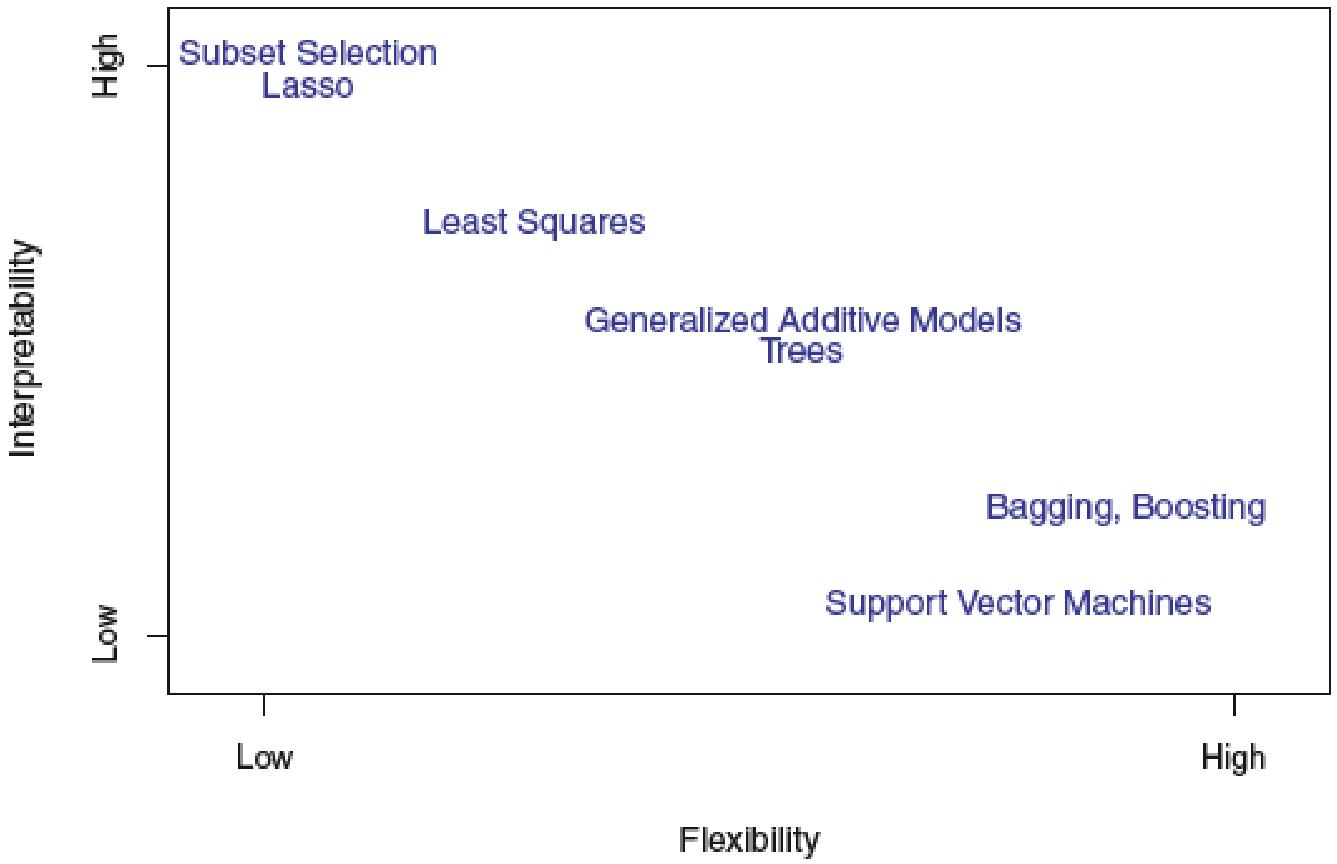
Case by Carlos Oscar Sorzano (CNB) on image processing

ISLR 9, CASI 19, Moguerza, Muñoz (2006), ESL 12, Bishop 7.1

Labs

- Support Vector Classifier
- Support Vector Machines
- SVMs with multiple clases
- Full SVM project
- Comparison of Linear regression, Ridge regression, Lasso, Elastic Net, PLS, SV Regression

A good old friend



SVMs. Motivation

Motivation

Perceptron (Rosenblatt, 1958). Minsky and Papert (1969) ‘blew’ them

Vapnik (Neural nets group at ATT) 1990’s

SVMs lead computer vision competititions until (re)emergence of convolutional neural networks (mid 2000’s)

Still very important

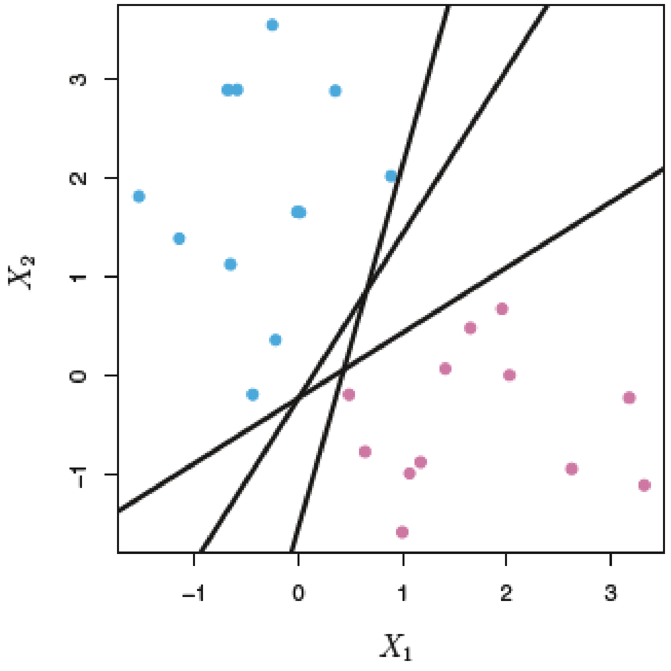
Sparse solution. Computationally efficient

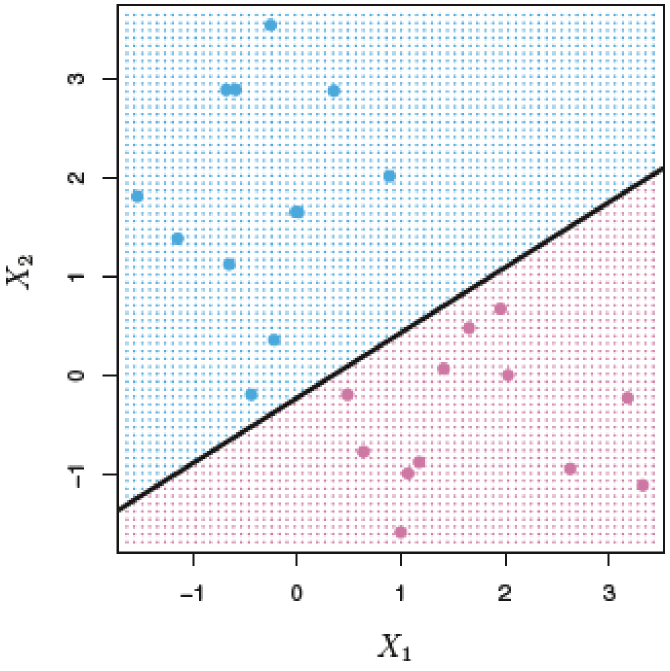
Absence of local minima in SVM optimization

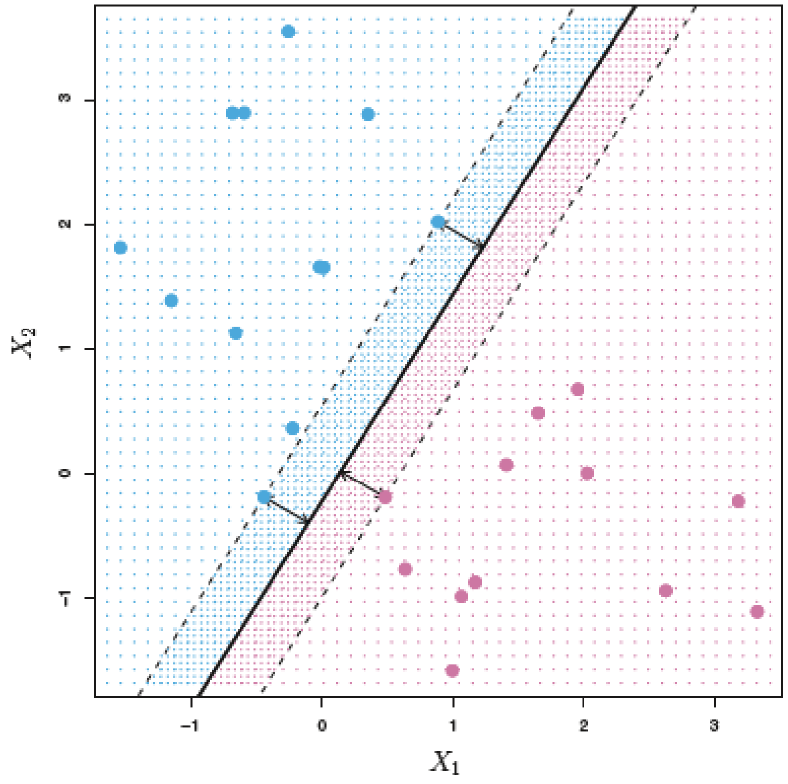
Classification problem reduced to computation of linear decision function

Initially, not a statistical learning theory

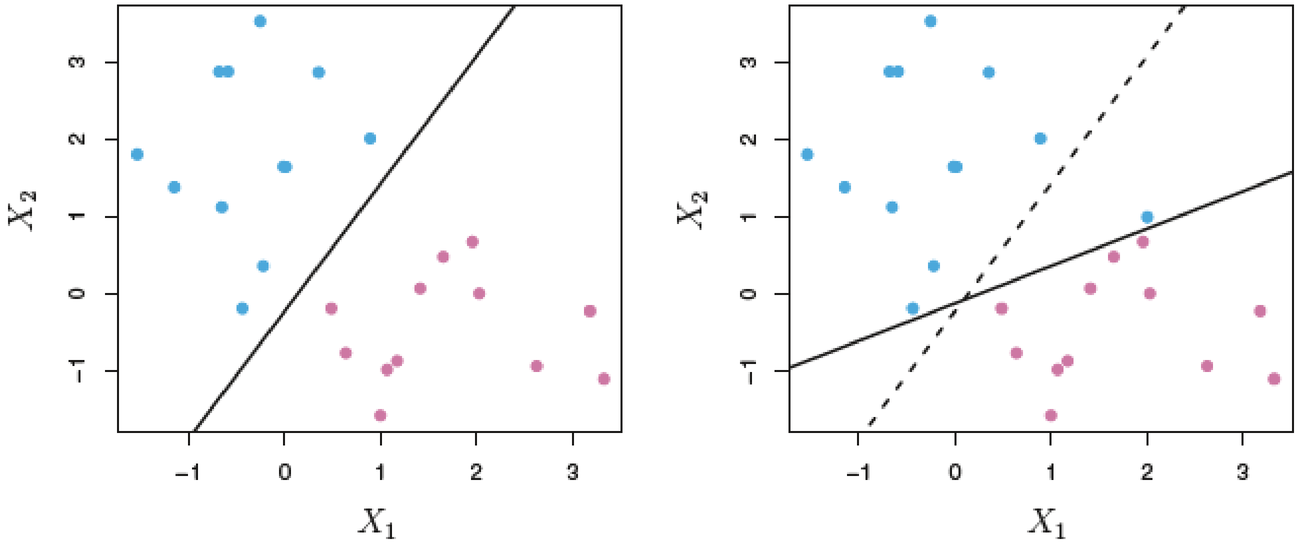
Concept. Linearly separable clases. MMC

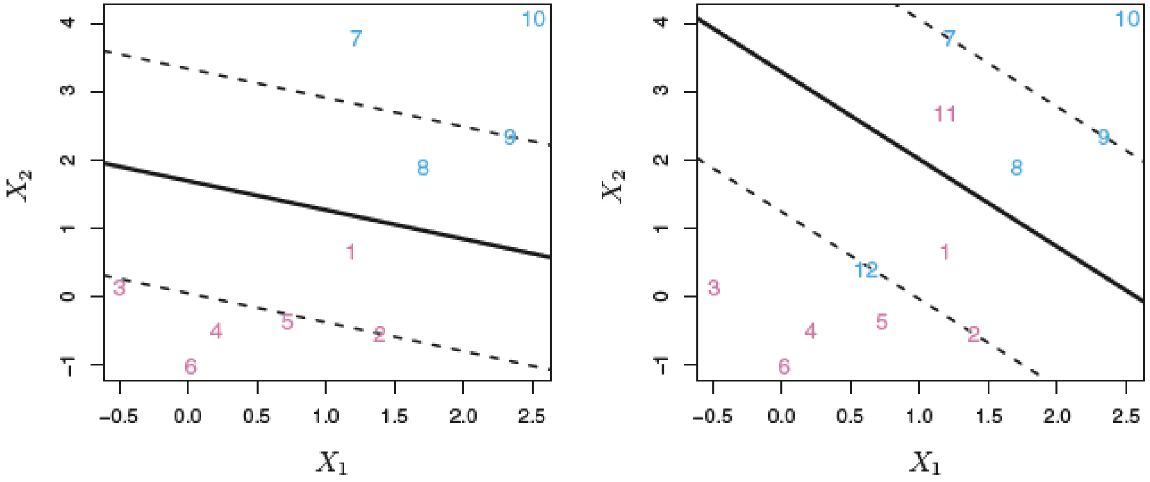




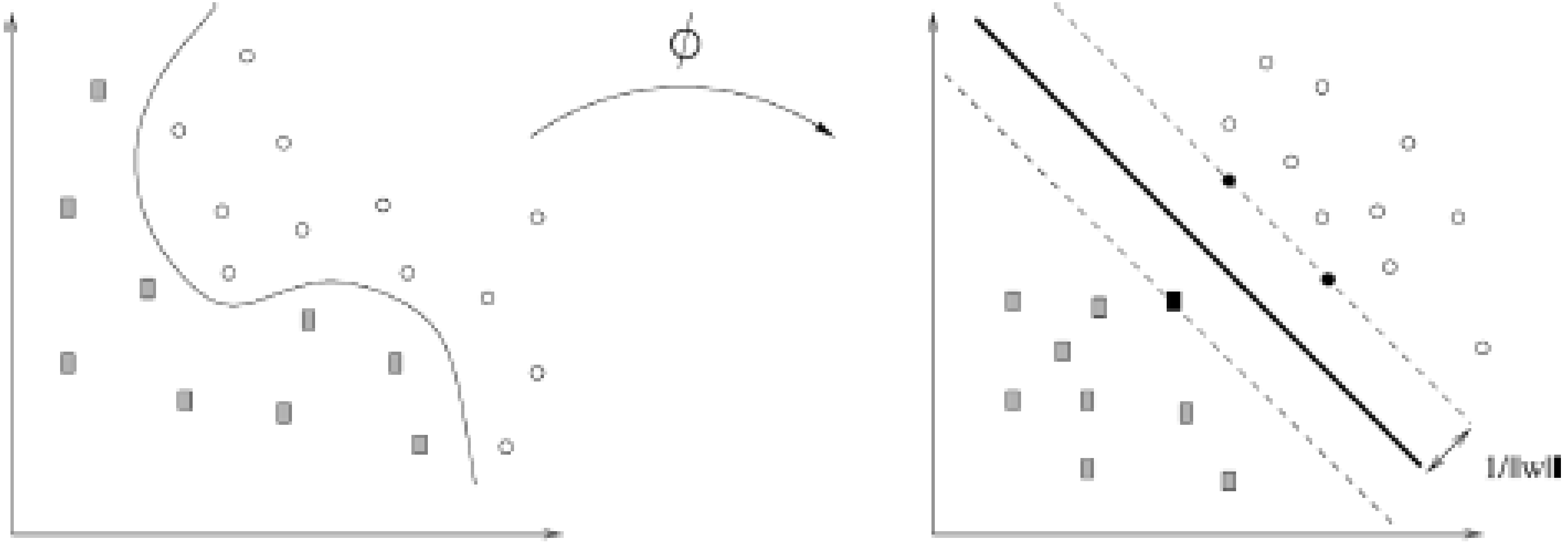


Concept. Soft margins. SVC





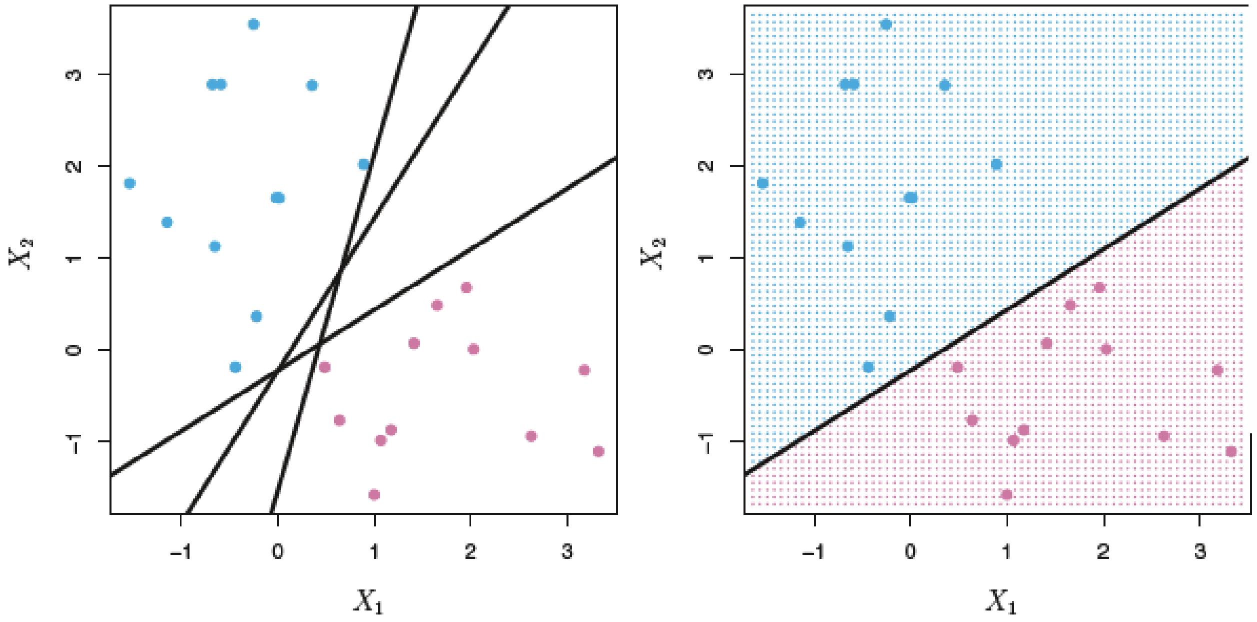
Concept. Kernel trick. SVM



Maximal margin classifier

Maximal margin classifier. Basics

A hyperplane separates two regions



$$\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = 0$$

$$\begin{cases} \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p > 0 \\ \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p < 0 \end{cases}$$

Classifying with a separating hyperplane

A separating hyperplane has the property. Classes +1,-1

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0, \quad y_i = 1$$
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} < 0, \quad y_i = -1$$
$$\Updownarrow$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) > 0, \quad i = 1, \dots, n$$

If a separating hyperplane exists, classify according to

$$x^* \longrightarrow f(x^*) = \beta_0 + \beta_1 x_1^* + \dots + \beta_p x_p^*$$
$$f(x^*) > 0 \longrightarrow 1 \qquad \operatorname{sgn}(f(x^*))$$
$$f(x^*) < 0 \longrightarrow -1$$

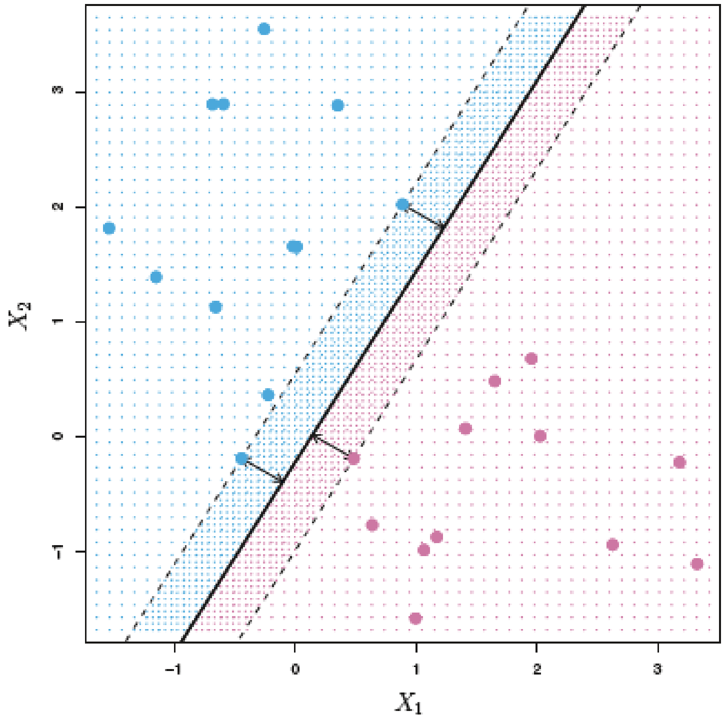
Maximal margin classifier. Concept

Optimal separating classifier→Maximal margin classifier

Separating hyperplane that is farthest from the training observations

Margin, minimal distance from points to hyperplane

Maximise margin



Depends only on the support vectors!!!

Maximal margin classifier. Construction

Through the optimisation problem

$$\max_{\beta_0, \ldots, \beta_p} \quad M$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 = 1$$

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

Maximal margin classifier. Construction

Reformulate

$$\frac{1}{\|\beta\|} y_i (x_i^T \beta + \beta_0) \geq M \qquad y_i (x_i^T \beta + \beta_0) \geq M \|\beta\|$$

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2$$

$$\text{s.t} \quad y_i (x_i^T \beta + \beta_0) \geq 1, \; i = 1, \ldots, n$$

Lagrangian formulation

$$\min \frac{1}{2} \|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i (x_i \beta + \beta_0) - 1 \right]$$

Maximal margin classifier. Construction

Derivatives and equal 0

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad 0 = \sum_{i=1}^{n} \alpha_i y_i$$

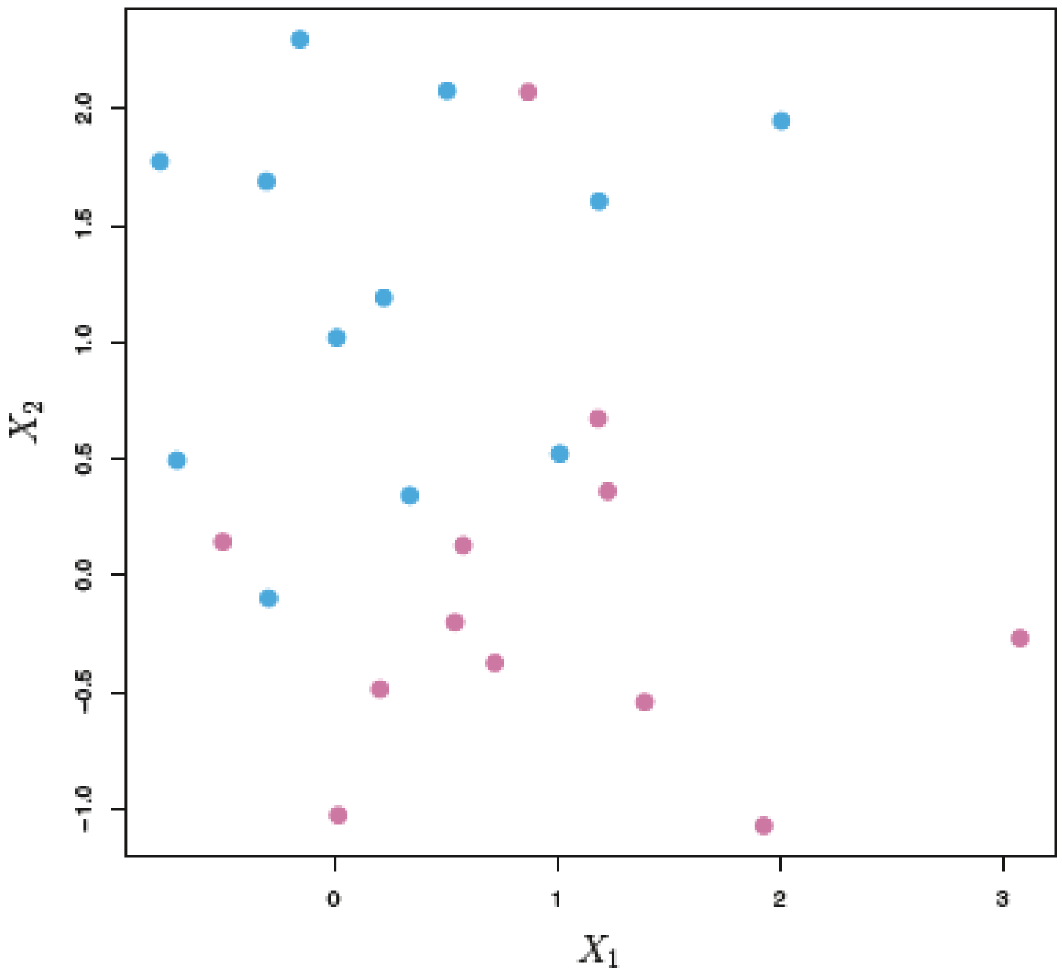
Dual

$$\max \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Maximal margin classifier. The non-separable case

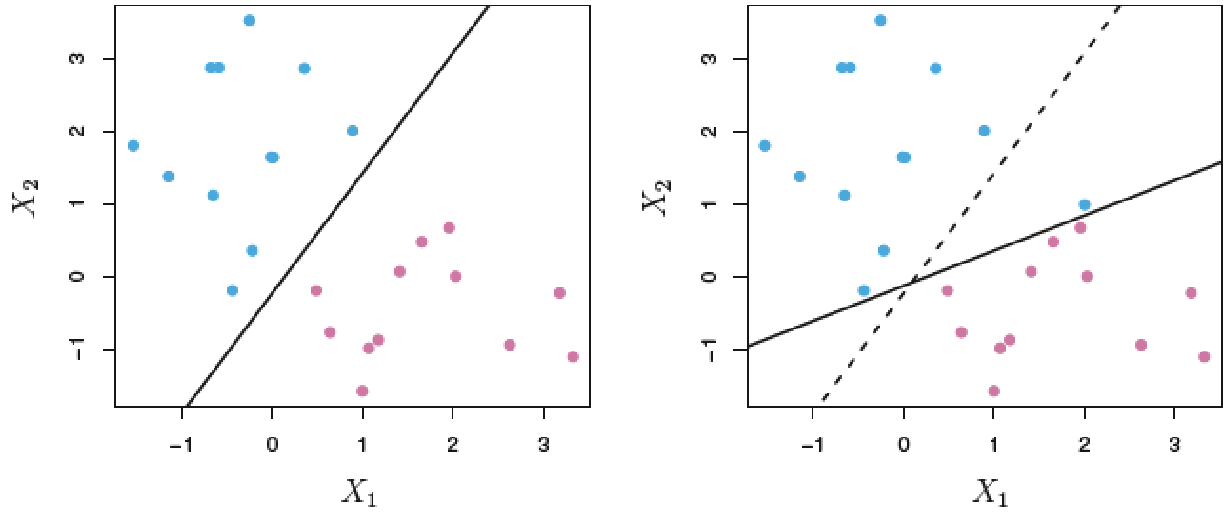
Sometimes no solution exists….



Support vector classifiers

Support vector classifiers. Concept I

MMC very sensitive to change in a single observation. Overfit



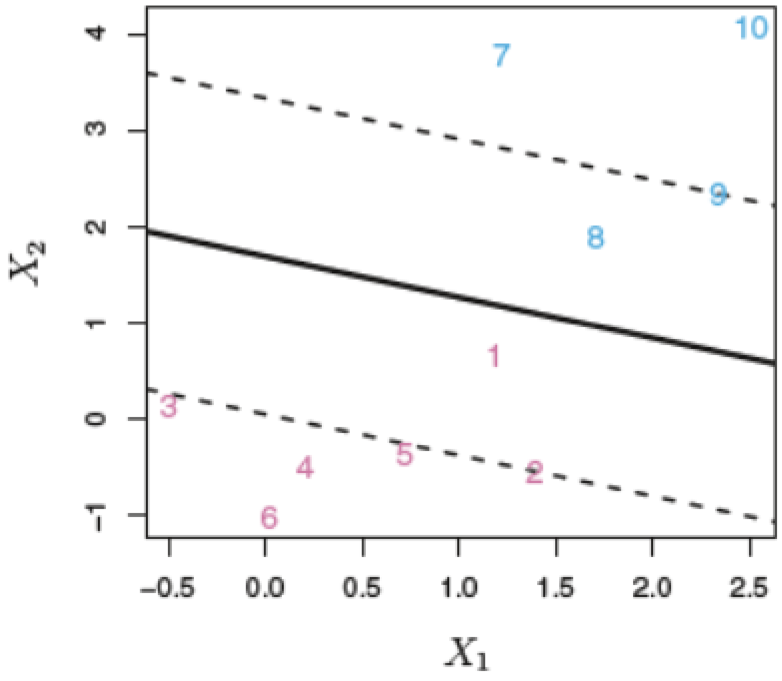
Willing to consider a classifier not perfectly separating to reach?

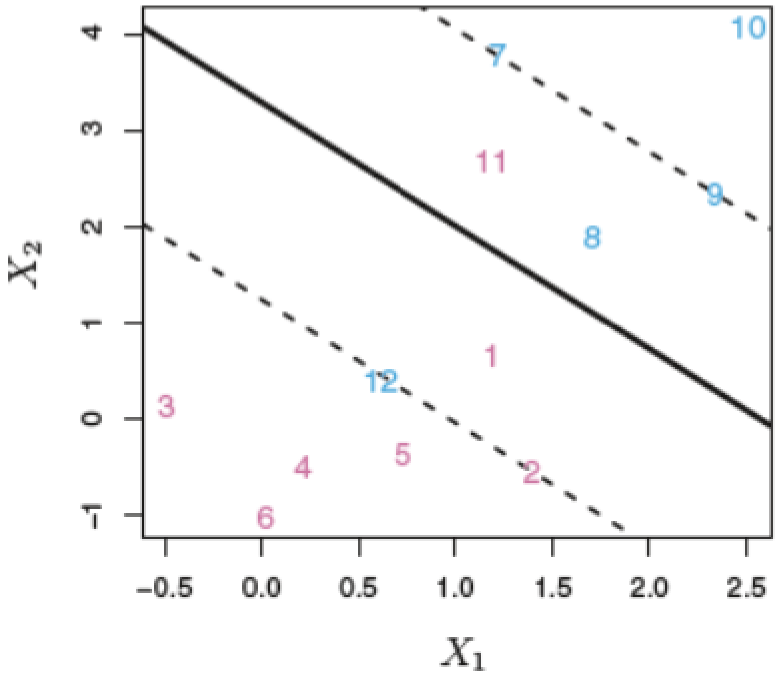
- Greater robustness
- Better classification of most training observations

Support vector classifiers. Concept II

SVC aka soft margin classifier

Allow some observations
In incorrect side of margin
or incorrect side of
hyperplane





Support vector classifiers. Construction I

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

$$\text{s.t.} \quad \sum \beta_i^2$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i)$$

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

Classify through

$$x^* \longrightarrow \operatorname{sgn}(\beta_0 + \beta_1 x_1^* + \ldots + \beta_p x_p^*)$$

Support vector classifiers. Construction II

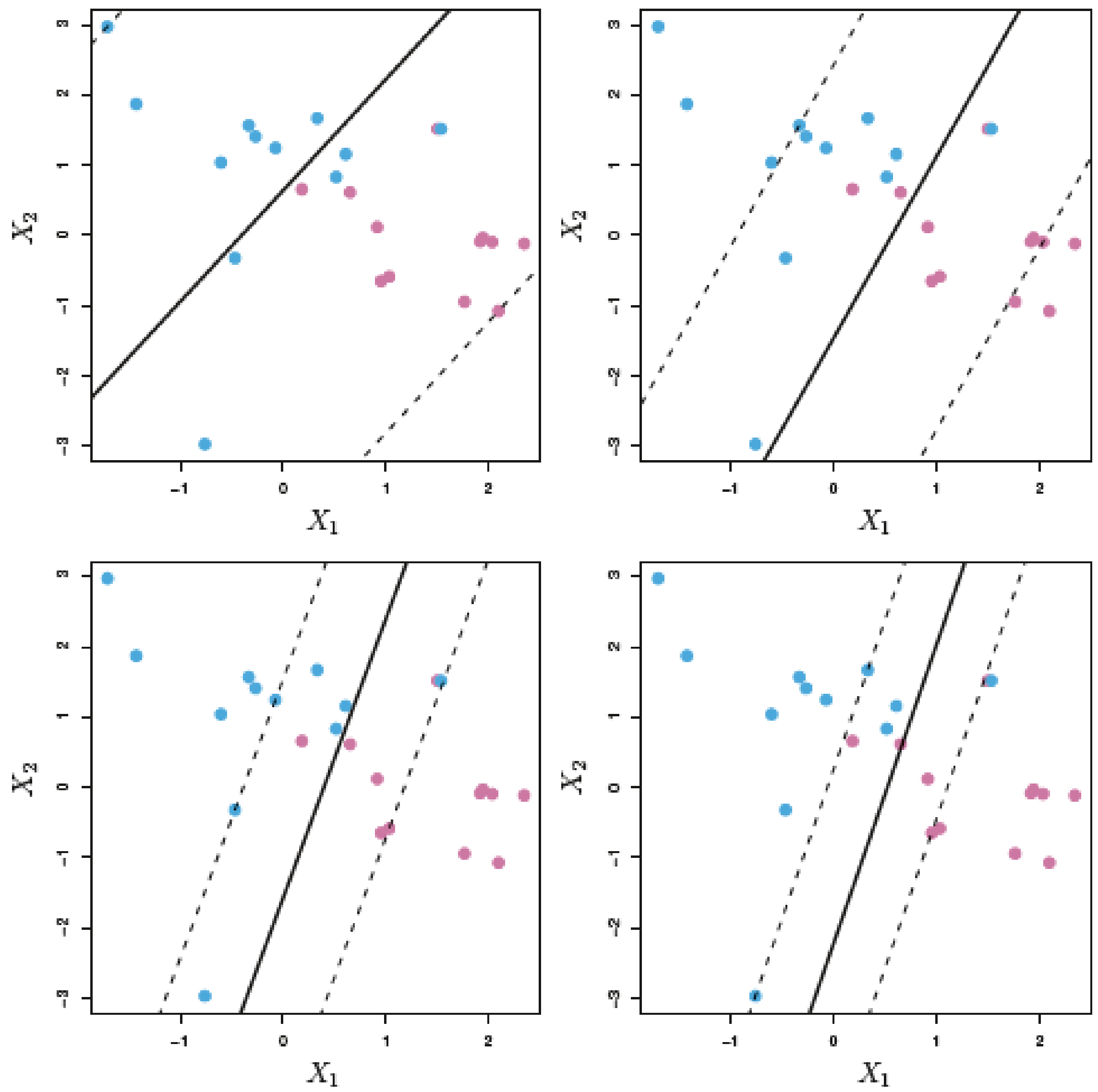
Slack ε.   =0, right side of margin. >0, wrong side of margin. >1, wrong side   of hyperplane

C bounds sum of slacks. C=0, no slacks, MMC. C>0, at most C observations on the wrong side. Bigger C, more tolerant

C tuning parameter chosen by CV

Only observations on the margin, or on the wrong side of the margin for their class matter in the optimisation. Support vectors

Support vector classifiers. Construction III



Support vector classifiers. Construction IV

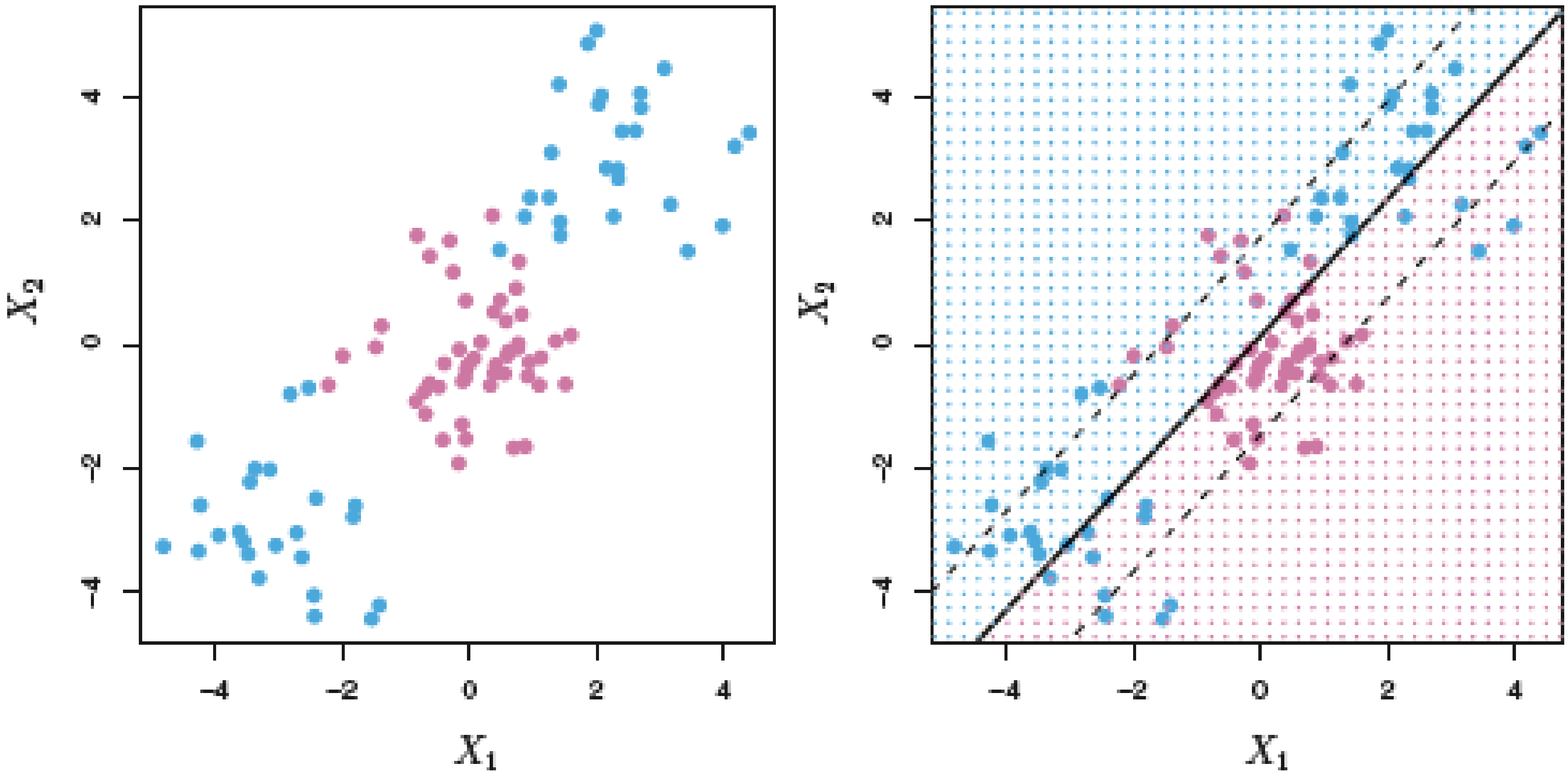
The linear support vector classifier can be represented $f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$

To estimate the parameters just need $\langle x_i, x_{i'} \rangle$

Only non-zero for support vectors $f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle$

We only need inner products!!!!

Support vector classifiers. Failures



Support vector machines

Classifying with non-linear boundaries

From $X_1, X_2, \ldots, X_p$

To $X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$

Solve

$$\max_{\beta, \varepsilon} \; M$$

$$\text{s.t.} \quad y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M (1 - \varepsilon_i)$$

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

Linear in enlarged space, nonlinear in original space. Efficiency?

Kernels

Map data into a higher dimensional space

Cover's theorem. Any data set becomes arbitrarily separable as data dimension grows

Kernels

$$K(x,y) \quad K: X \times X \longrightarrow \mathbb{R}$$

$$\phi: X \longrightarrow Z$$

$$K(x,y) = \phi^T(x)\,\phi(y)$$

Inner product in Z

X, input space. Z, feature space

Kernel. Example

$$K(x_1, x_2) = (1 + x_1^T x_2)^2$$

$$= (1 + x_{11} x_{21} + x_{12} x_{22})^2 = \phi^T(x_1)\,\phi(x_2)$$

$$\phi(x_i) = (1, \sqrt{2} x_{i1}, \sqrt{2} x_{i2}, x_{i1}^2, x_{i2}^2, \sqrt{2} x_{i1} x_{i2})$$

$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^6$$

Kernel

Positive definite function admiting an expansion

Mercer's theorem provides conditions

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) \qquad \lambda_i \in \mathbb{R}^+$$

Kernel

Linear kernel $K(x,y) = x^T y$

Polynomial kernel degree d $K(x,y) = (c + x^T y)^d$

Radial kernel $K_c(x,y) = e^{-\|x-y\|^2/c}$

Kernel

Given K

$$f(x) = \sum \alpha_j \; K(x_j, x)$$

Reproducing Kernel Hilbert space (RKHS)

$$f(x) = \sum_j \alpha_j \, K(x_j, x) + b$$

Adopt the form

$$f(x) = \sum_s \alpha_j \cdot K(x_i, x) + b$$

Support vector machines

SVC enlarging feature space with (nonlinear) kernel!!!!

SVMs. Other topics

SVMs. More than two classes

One vs one classification

Consider all pairs of clases

Classify to the most frequently assigned class

One vs all classification

Fit SVM for each class against all others

Assign to that maximising

SVMs vs Logistic Regression

Rewrite SVC as

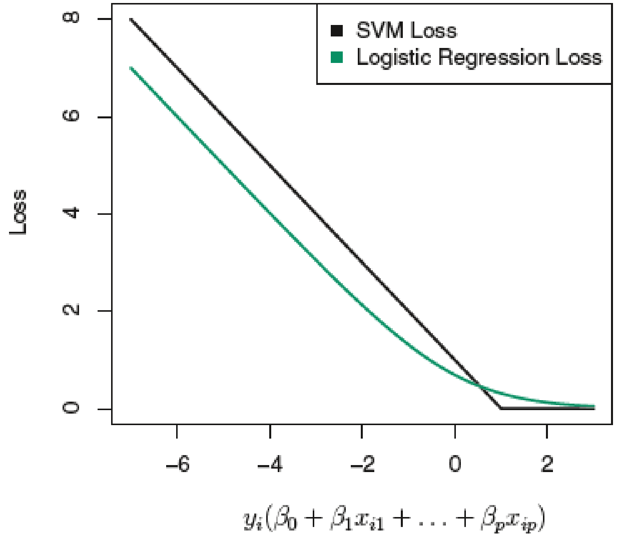
$$\min_{\beta} \underbrace{\sum_{i=1}^{n} \max\left(0, 1 - y_i f(x_i)\right)}_{\text{LOSS}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{PENALTY}}$$

Loss+penalty

Hinge loss

$$L(X, y, \beta) = \sum_{i=1}^{n} \max\left[0, 1 - y_i\left(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}\right)\right]$$

$$\max\left[0, 1 - y_i f(x_i)\right]$$



Further topics

Support vector regression

Computational learning theory

Platt trick to make it probabilistic

Final comments

Perceptron

CNNs

Function fitting using kernels

Kernel smoothing and local regression