Support Vector Machines

University of Santiago de Compostela
Manuel Mucientes

Introduction

- Support Vector Machines (SVMs) are one of the best classifiers
- SVMs are a generalization of the maximal margin classifier

esquema que vamos a seguir

- Maximal margin classifiers require that the classes are separable by a linear boundary
- Support vector classifiers are an extension of maximal margin classifiers
- SVMs extend support vector classifiers to accommodate non-linear boundaries

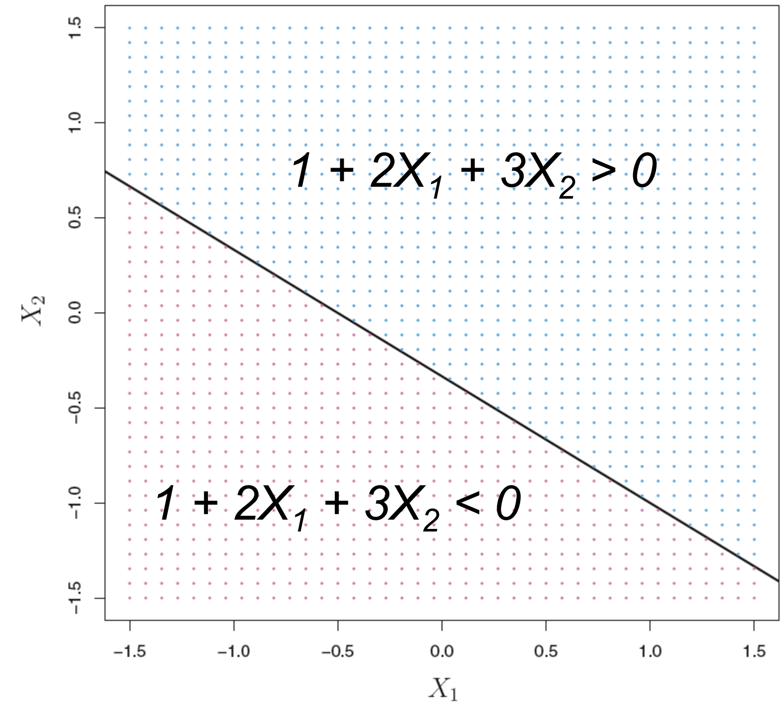
Hyperplanes

- In a p-dimensional space, a hyperplane is a flat affine (needs not to pass through the origin) subspace of dimension p-1

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

- A hyperplane divides a p-dimensional space into two halves
  - $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0$
  - $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$

normalizar dividiendo entre el modulo de beta (sin incluir beta 0!!!)

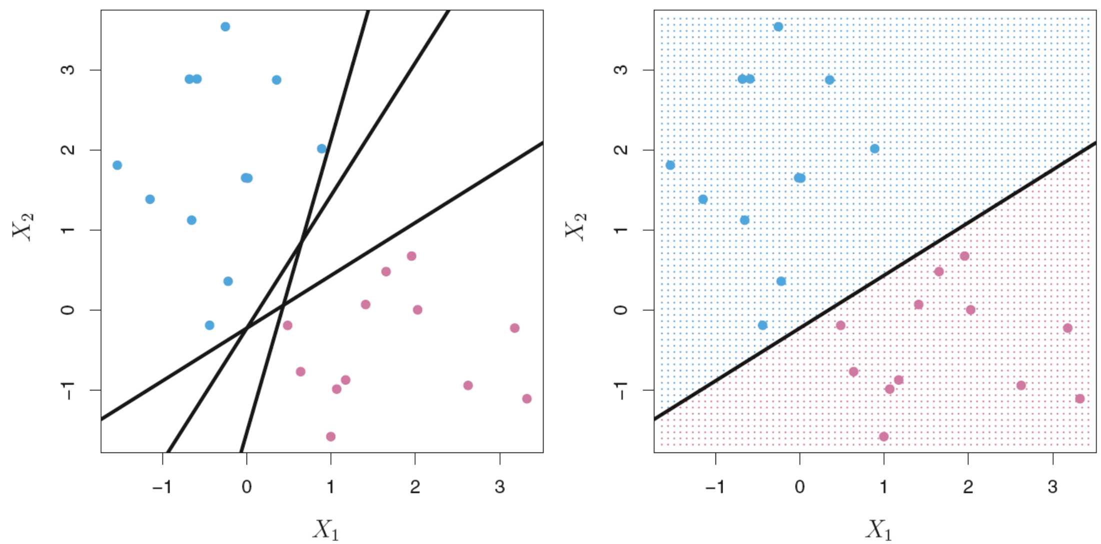


Classification using a Separating Hyperplane

- Separating hyperplane:
  - $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \text{ if } y_i = 1$
  - $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \text{ if } y_i = -1$
  - $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$
- Classify a test observation based on the sign of:

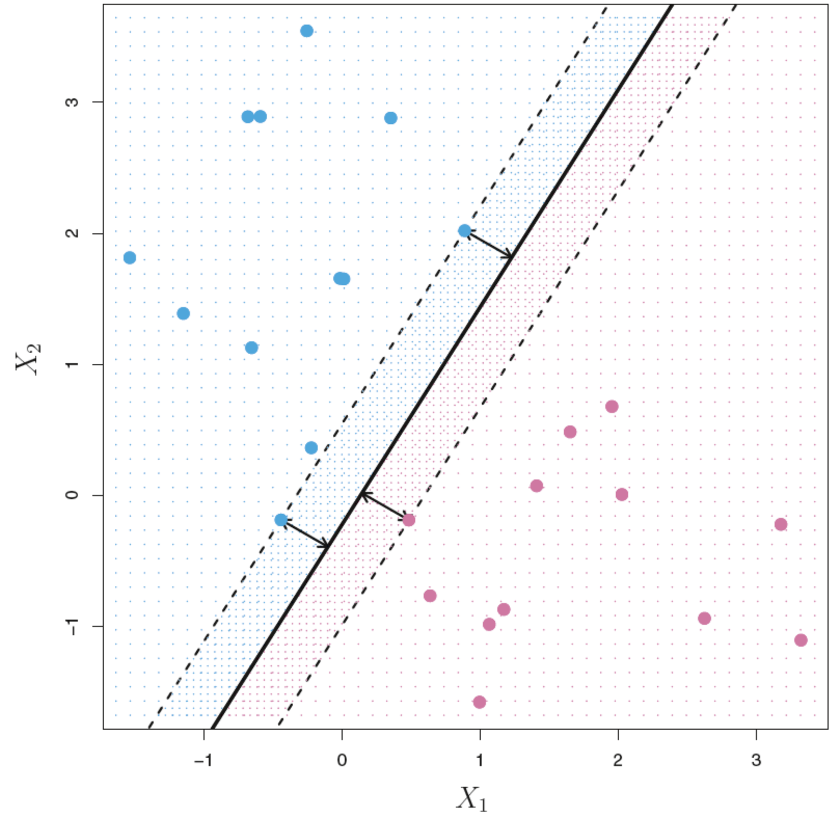
$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$$

- The magnitude of $f(x^*)$ gives the confidence
- This classifier leads to a linear decision boundary



Maximal Margin Classifier

- If data is separable by a hyperplane, there exist an infinite number of such hyperplanes
- Maximal margin hyperplane
  - Maximal Margin Classifier (MMC)
- Support vectors: observations in p-dimensional space that “support” the hyperplane
  - If they were moved the maximal margin hyperplane would move as well



Maximal Margin Classifier

- Solution to the optimization problem:

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

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

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \ \ \forall\, i = 1, \ldots, n$$

  - Second condition: each observation in the correct side, at least at a distance M
  - First condition: adds meaning to the second constraint; distance to the hyperplane

- Classification rule: $G(x) = \text{sign}[x^T\beta + \beta_0]$

Maximal Margin Classifier

- Get rid of the ||β||=1 constraint

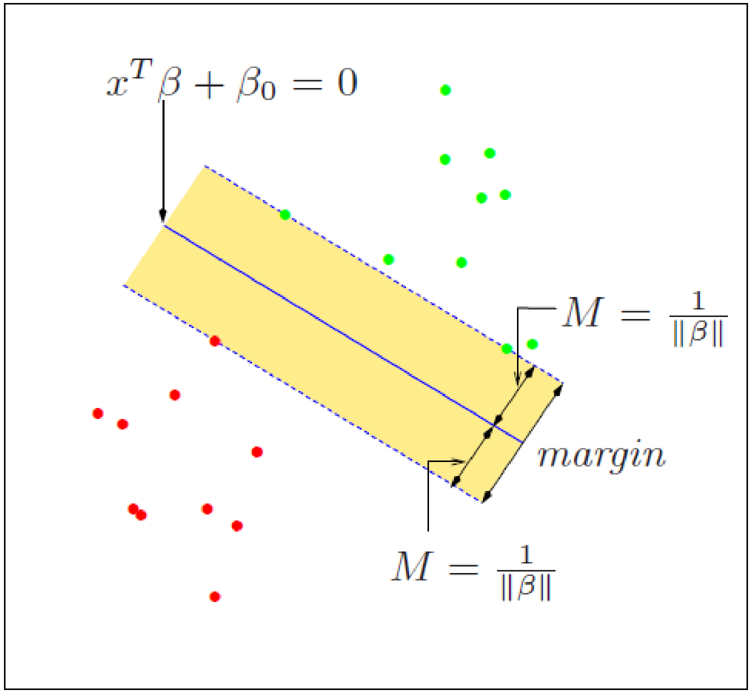
$$\frac{1}{||\beta||} y_i(x_i^T \beta + \beta_0) \geq M$$

$$y_i(x_i^T \beta + \beta_0) \geq M||\beta||$$

- We can arbitrarily set ||β||=1/M

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

$$\text{subject to } y_i(x_i^T \beta + \beta_0) \geq 1, \; i = 1, \dots, N$$



- Convex optimization problem:
  - Quadratic criterion
  - Linear inequality constraints

Maximal Margin Classifier

- Lagrange multipliers method:

$$\text{Maximize } f(x)$$
$$\text{subject to } g_j(x) = 0 \text{ for } j = 1, \ldots, J,$$
$$\text{and } h_k(x) \geq 0 \text{ for } k = 1, \ldots, K.$$

- Lagrangian function:

$$L(x, \{\lambda_j\}, \{\mu_k\}) = f(x) + \sum_{j=1}^{J} \lambda_j g_j(x) + \sum_{k=1}^{K} \mu_k h_k(x)$$
$$\text{subject to } \mu_k \geq 0 \text{ and } \mu_k h_k(x) = 0 \text{ for } k = 1, \ldots, K.$$

  - Karush-Kuhn-Tucker (KKT) conditions

- In our optimization problem, the Lagrange (primal) function to be **minimized** w.r.t. β and β_0 is:

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1]$$

  - Minimization: inverted sign
  - α_i: Lagrange multipliers (μ_k)

Maximal Margin Classifier

- $$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1] \qquad (1)$$
- Deriving w.r.t. β and β_0 and setting derivatives to zero:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \quad (2) \qquad 0 = \sum_{i=1}^{N} \alpha_i y_i, \quad (3)$$

- Substituting Eqs. 2-3 in Eq. 1: Lagrangian (Wolfe) dual func.

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0.$$

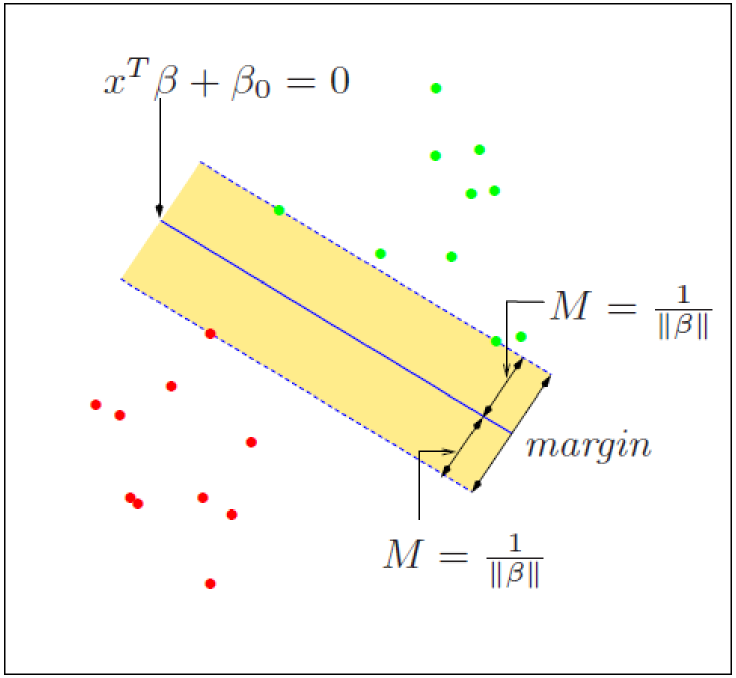
(KKT conditions)

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - 1] = 0 \; \forall i. \qquad (4)$$

  - Maximize L_D: simpler convex optimization problem
    - Obtains α_i

Maximal Margin Classifier

- $\alpha_i[y_i(x_i^T\beta + \beta_0) - 1] = 0 \; \forall i.$ (4)
- if $\alpha_i > 0$, then $y_i(x_i^T\beta + \beta_0) = 1$:
  - x_i is in the edge of the margin
- if $y_i(x_i^T\beta + \beta_0) > 1$: $\alpha_i = 0$
  - x_i is outside the margin
- Support vectors: x_i with $\alpha_i > 0$



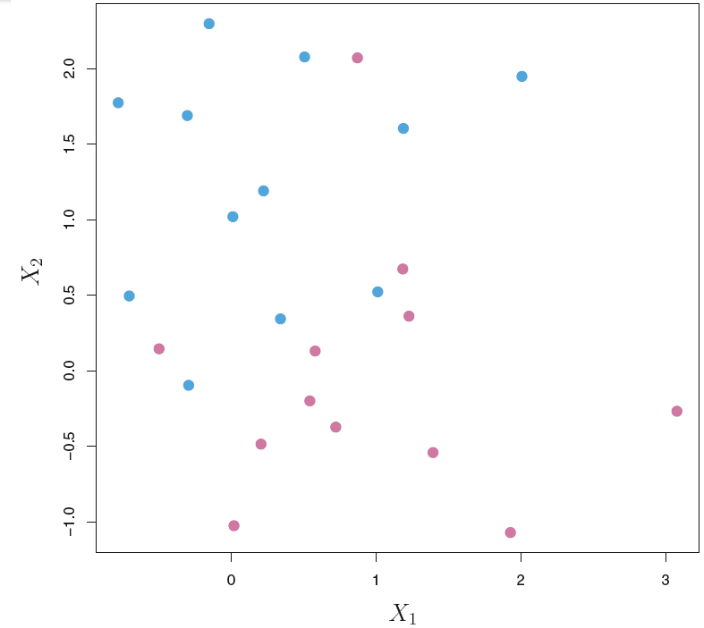
- β: linear combination of the support vectors (eq. 2)

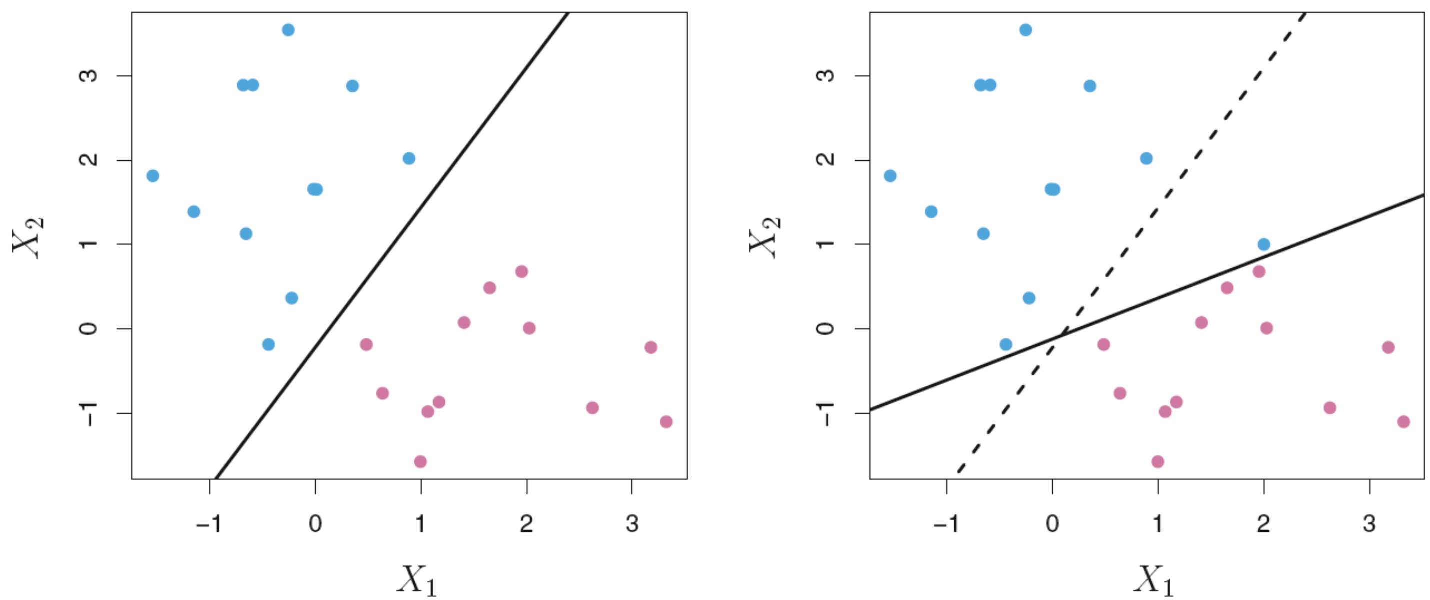
$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i,$$

- β_0 obtained solving eq. 4 for any support vector
  - Average of all the solutions for numerical stability

Support Vector Classifiers

- No separating hyperplane exists
- Sometimes, a classifier based on a separating hyperplane is not desirable
  - Extremely sensitive to one observation: overfitting

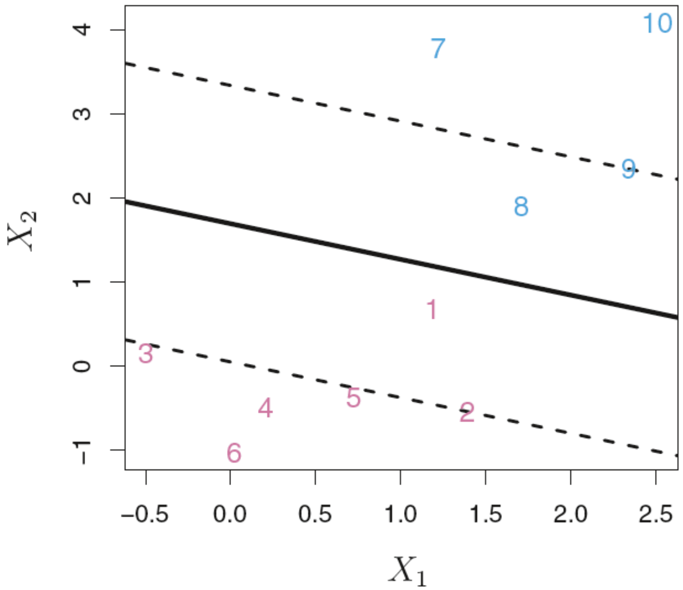




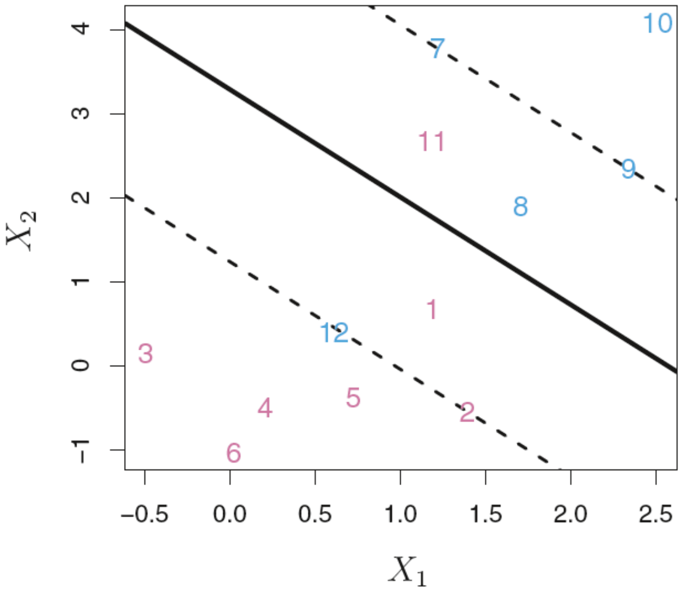
Support Vector Classifiers

- A classifier that does not perfectly separate the two classes
  - Greater robustness to individual observations
  - Better classification of most training observations

On the margin: 2, 9
Wrong side of the margin: 1, 8

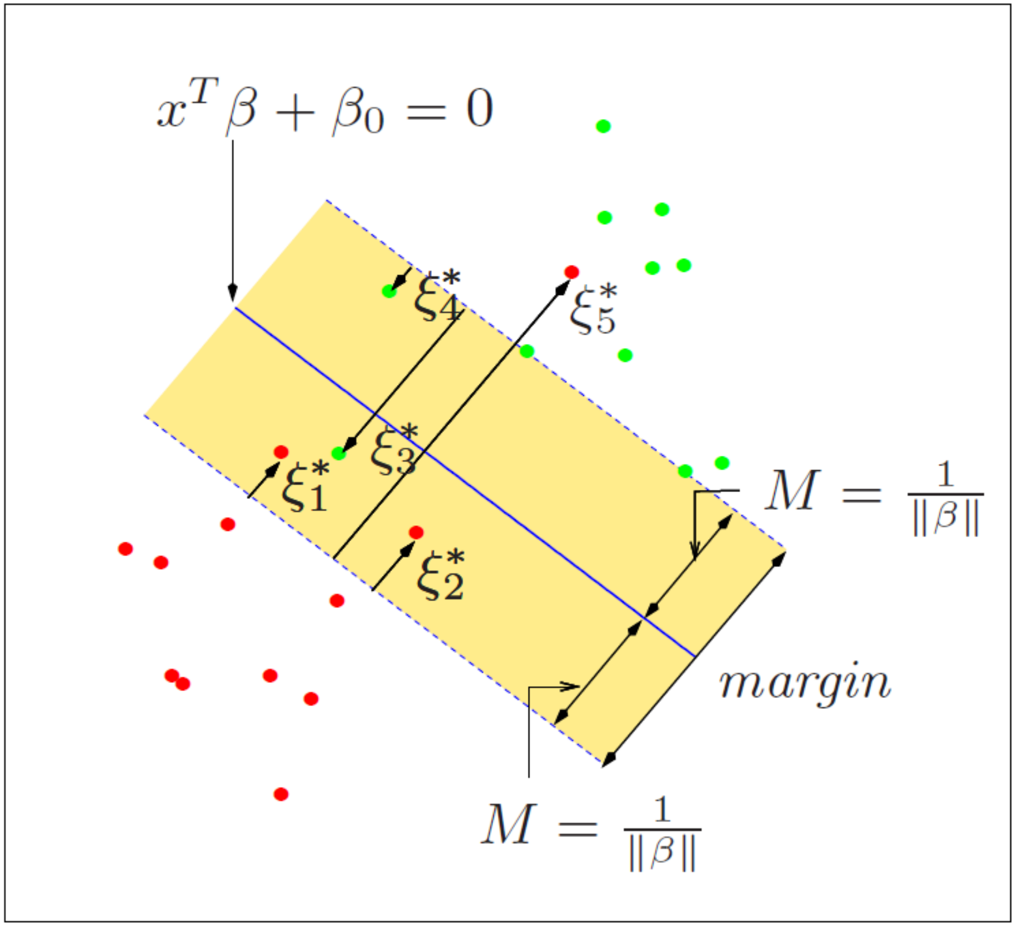


On the margin: 2, 7, 9
Wrong side of the margin: 1, 8
Wrong side of the hyperplane: 11, 12



Support Vector Classifiers

- ε_i tells where the i-th observation is located: percentage of M
  - $\varepsilon_i=0$: observation in the correct side of the margin
  - $\varepsilon_i>0$: observation in the wrong side of the margin
  - $\varepsilon_i>1$: observation in the wrong side of the hyperplane (misclassification)



Support Vector Classifiers

- Optimization problem:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

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

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

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq \textit{constant}$$

Support Vector Classifiers

- Rephrasing the problem:

$$\min \|\beta\| \quad \text{subject to} \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i, \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{constant}. \end{cases}$$

- Computationally convenient to re-express as:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{subject to} \quad \xi_i \geq 0, \ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i$$

  - *C* proportional to the inverse of the constant
    - Inverse of a regularization parameter
    - Separable case: $C=\infty$

Support Vector Classifiers

- Lagrange (primal) function: minimize w.r.t. β, β_0, ε_i

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i \quad (1)$$

- Setting the derivatives to zero:

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \quad (2) \qquad 0 = \sum_{i=1}^{N}\alpha_i y_i \quad (3)$$

$$\alpha_i = C - \mu_i,\ \forall i \quad (4) \qquad \alpha_i,\ \mu_i,\ \xi_i \geq 0\ \forall i \quad (5)$$

- Substituting eqs. 2-4 in eq. 5: Lagrangian (Wolfe) dual func.

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'}$$

Support Vector Classifiers

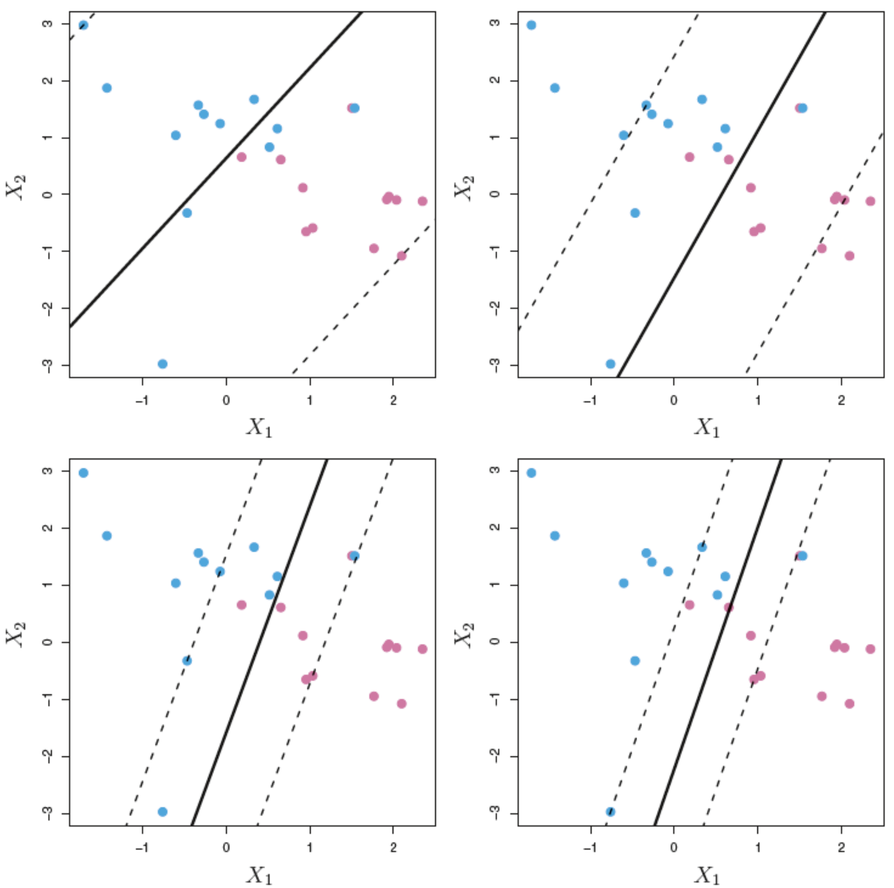
- Maximize L_D: $\alpha_i,\ \mu_i,\ \xi_i \geq 0\ \forall i$
  - $\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] \ = \ 0,$ (1)
  - $\mu_i\xi_i \ = \ 0,$ (2)
  - $y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) \ \geq \ 0,$ (3)
  - $\beta \ = \ \sum_{i=1}^{N} \alpha_i y_i x_i$ (4) $\alpha_i \ = \ C - \mu_i,\ \forall i$ (5)
- Support vectors: $\alpha_i>0$ (eq. 1)
  - Support vectors in the edge: $\varepsilon_i=0$, $0<\alpha_i<C$ (eqs. 2, 5)
    - From eq. 1 use any of these margin points to solve for β_0
      - Average all the solutions for numerical stability
  - The remainder support vectors: $\varepsilon_i>0$, $\alpha_i=C$ (eqs. 2, 5)
- Decision function: $G(x) = \text{sign}[x^T\beta + \beta_0]$

Support Vector Classifiers

- *C* is the tuning parameter
  - Bias-variance trade-off
  - Choose the value of *C* via cross-validation

- **Note**: in James et al. the *C* parameter is not the standard one, but inversely proportional!!!

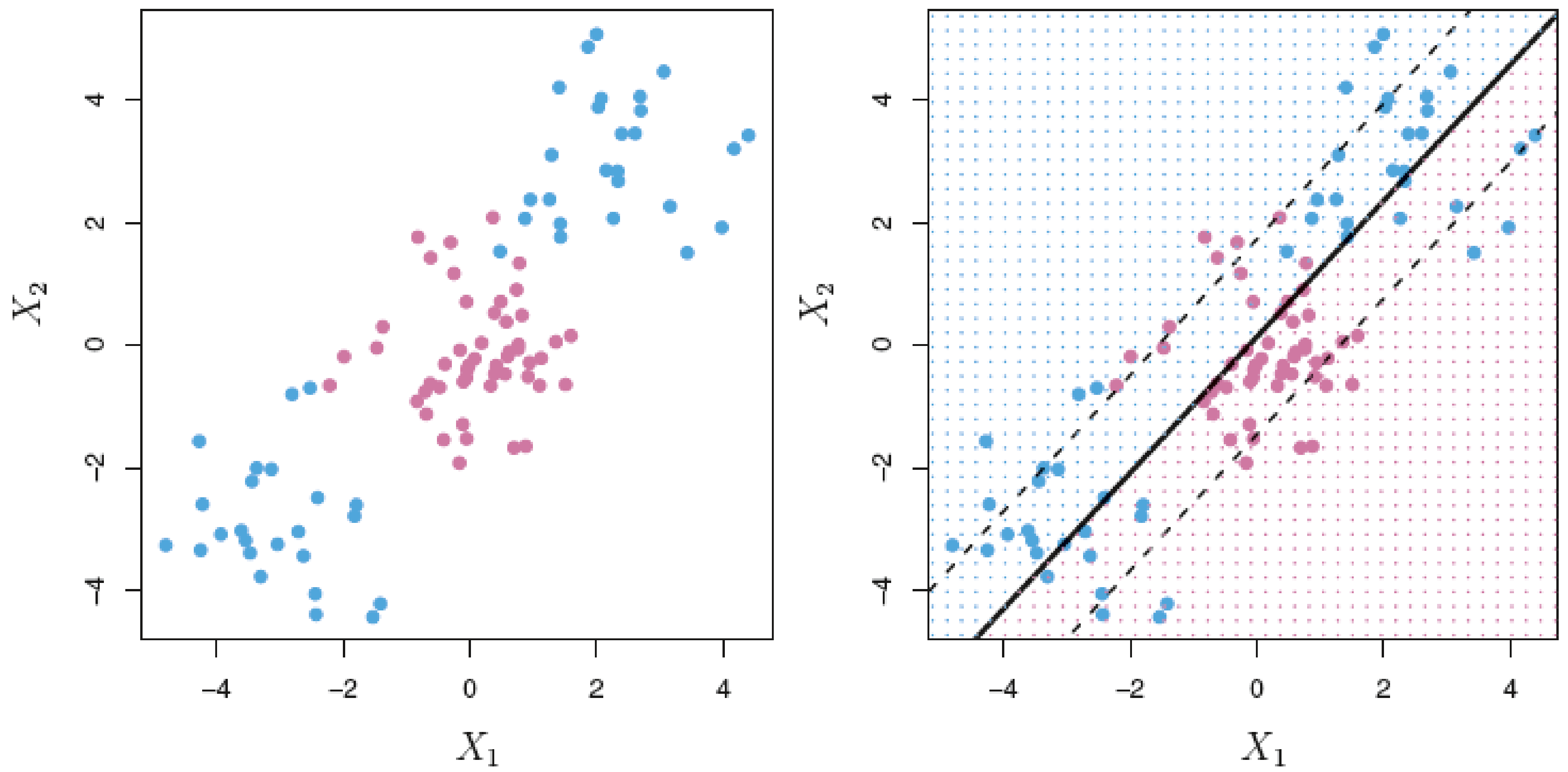
Small C



Large C

Support Vector Machines

- Non-linear class boundaries
- Enlarge the feature space



Support Vector Machines

- Feature space enlarged with functions of the predictors
  - Huge number of possible features
- SVM enlarge the feature space using kernels

- Support vector classifier: inner products of the observations
  - $G(x) = \text{sign}[x^T \beta + \beta_0]$ $\qquad$ $\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$
  - $\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$ $\qquad$ $f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i \langle x, x_i \rangle y_i$

Support Vector Machines

- Transformed feature vectors *h*(*x*):
  - Solution function:
    $$\begin{aligned} f(x) &= h(x)^T\beta + \beta_0 \\ &= \sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i)\rangle + \beta_0 \end{aligned}$$
    - Cheap computations of the inner products for particular choices of *h*
  - All we need are inner products
    - To represent the linear classifier *f*(x)
    - To compute its coefficients
- Need not to specify *h*(*x*), but the kernel function:
  $$K(x, x') = \langle h(x), h(x')\rangle$$
  - Similarity between two observations
- Solution: $$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$$

Support Vector Machines

- Kernel vs. enlarging the feature space using functions
  - Computational advantage: n(n-1)/2 inner products
  - Without explicitly working in the enlarged feature space
- Linear kernel: SVC

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

- Combination of SVC with a non-linear kernel: SVM
- Polynomial kernel of degree *d*:

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

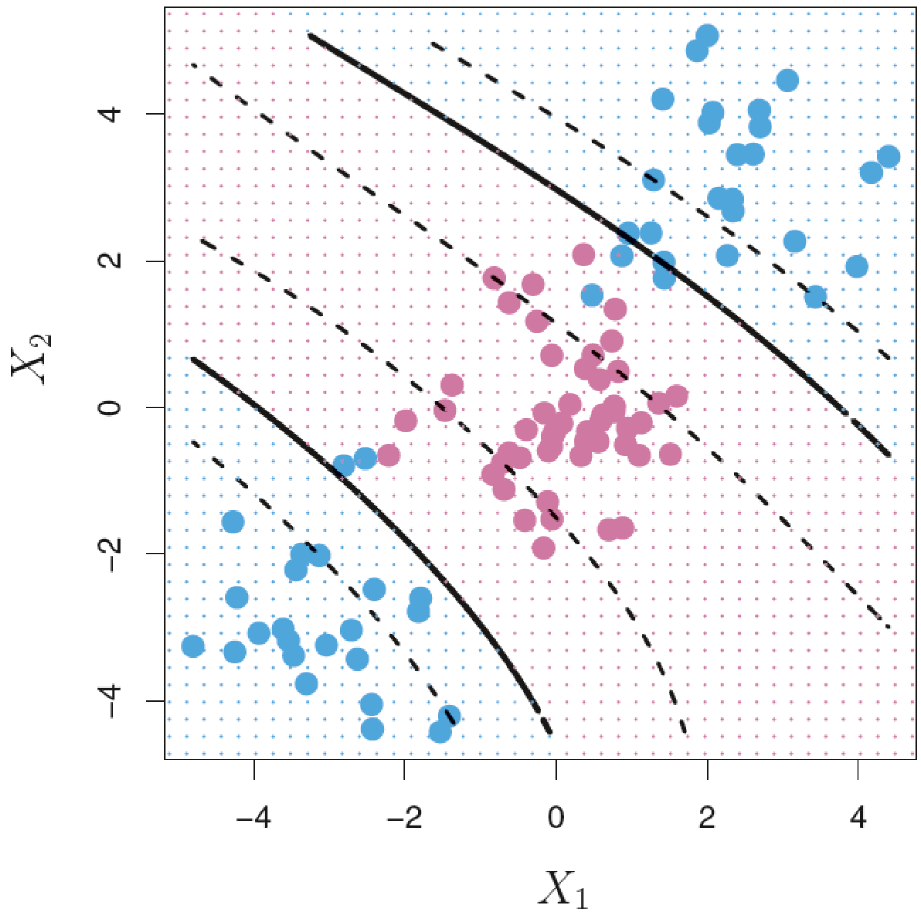
- Radial kernel:

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

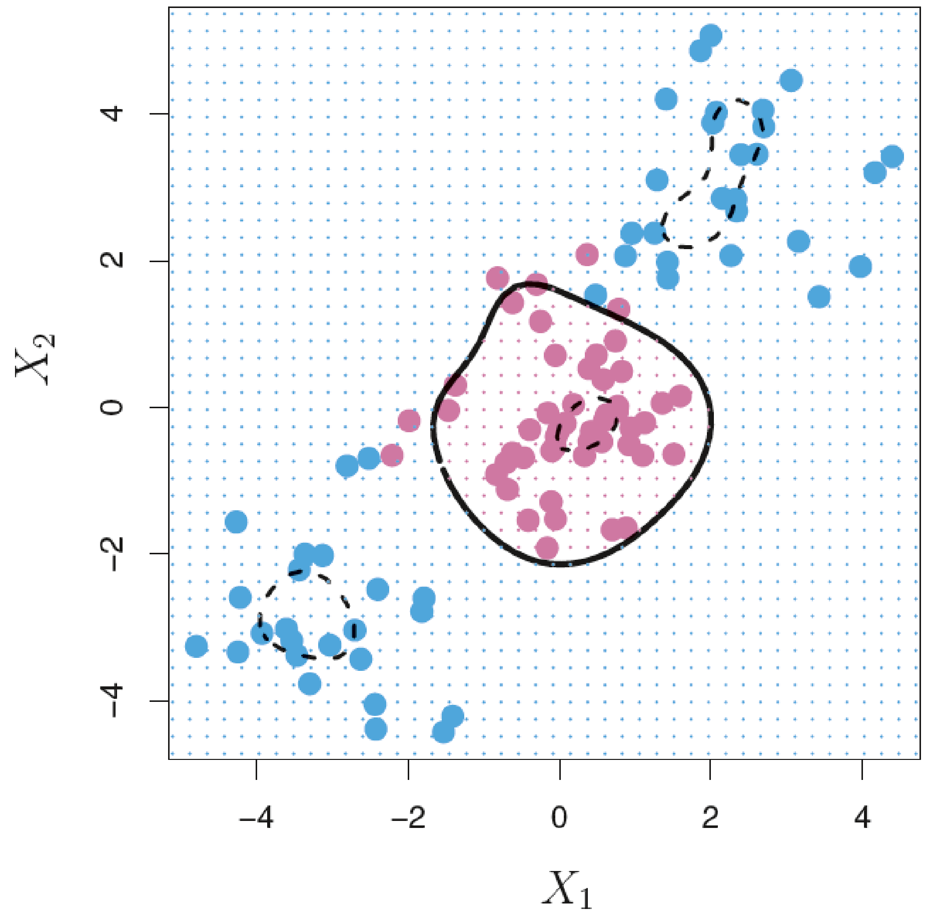
  - Very local behavior: training observations far from x play no role

Support Vector Machines

Polynomial kernel (d=3)

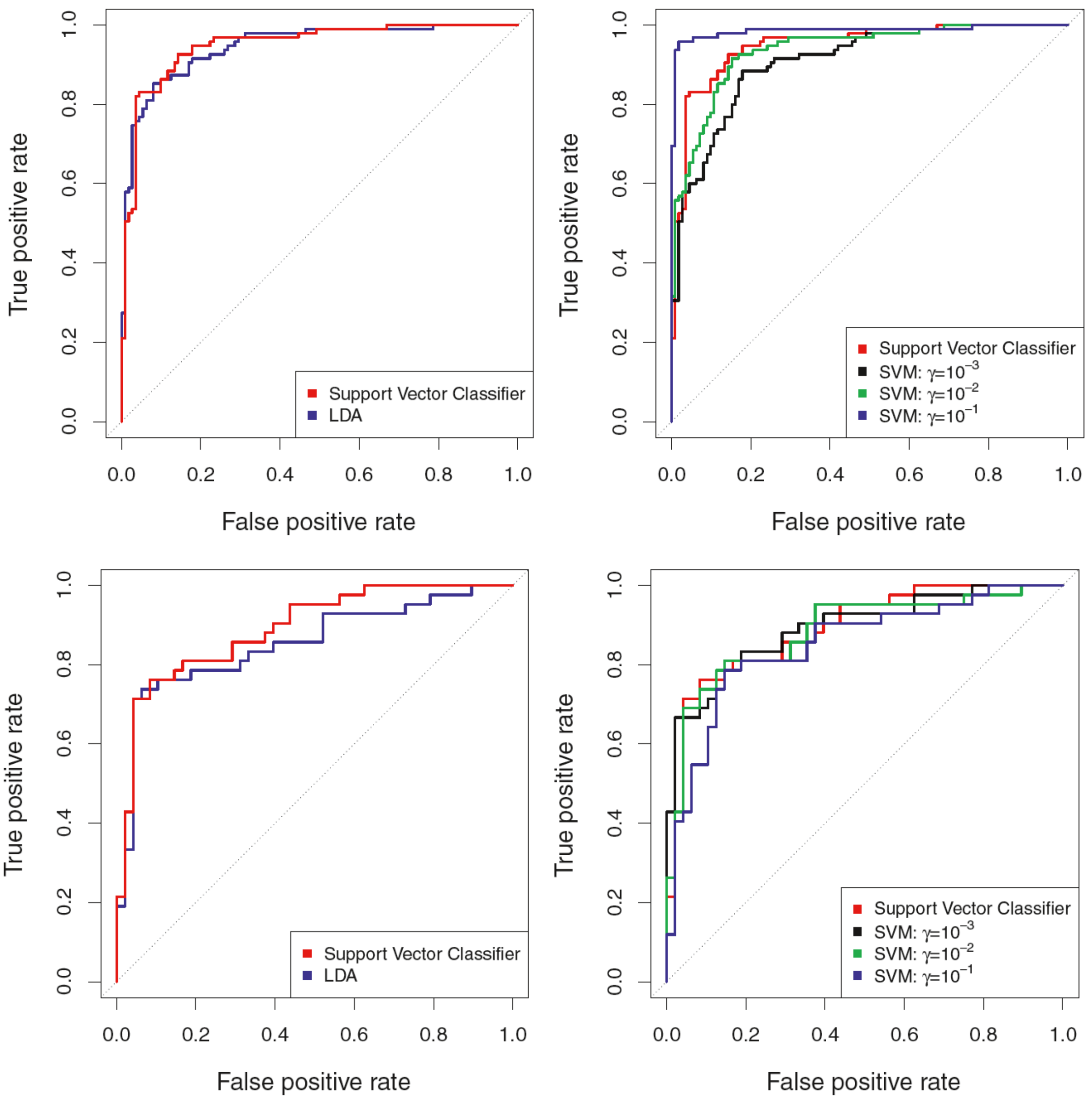


Radial kernel



Example: Heart dataset

- Training (upper row)
  - 207 observations
  - Best: SVM-$\gamma=10^{-1}$
- Test (lower row)
  - 90 observations
  - Best: SVC, SVM-$\gamma=10^{-2}$, SVM-$\gamma=10^{-3}$



SVMs with more than Two Classes

- K classes

- One vs. One
  - Learn K(K-1)/2 (all the pairs) of classifiers
  - Test: count the number of times that the observation is assigned to each of the K classes

- One vs. All
  - Learn K classifiers: k-th class vs. remaining K-1 classes
  - Test: assign the observation to the class with largest $f(x)$ (highest level of confidence)

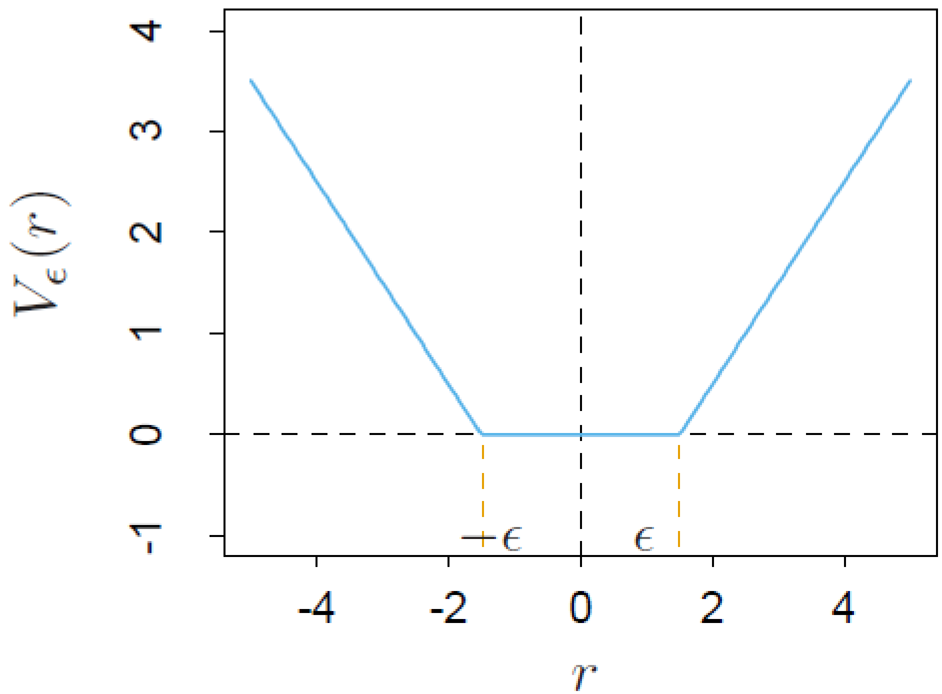
Support Vector Regression

- Minimize:

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2$$

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon, & \text{otherwise} \end{cases}$$

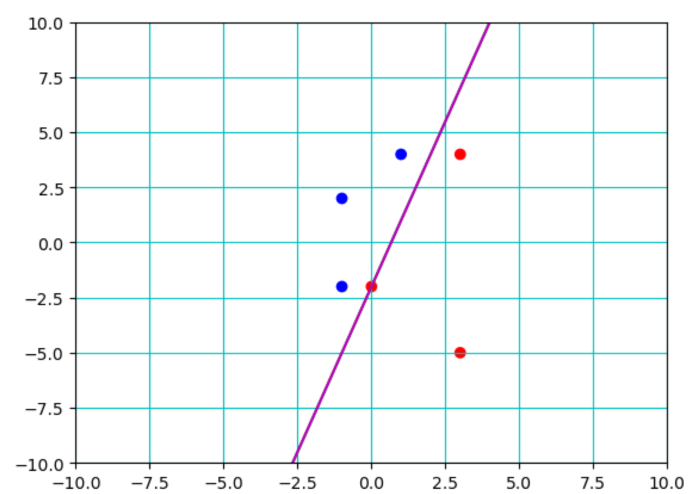
  - λ: regularization parameter
- SVR not as good for regression as SVMs for classification



Exercise

- Given the following classification data set with 6 examples, 2 input variables and 1 output variable, using a linear SVM with C=1, we have obtained the corresponding alpha values indicated in the last column.
  - What are the support vectors and which of them are in the margin boundary?
  - What are the hyperplane coefficients (beta, beta_0) and the value of M?
  - What are the values of epsilon?
  - Which examples are incorrectly classified?

Example	X_1	X_2	Y	alpha
1	-1	-2	+1	0.944
2	-1	+2	+1	0
3	+1	+4	+1	0.111
4	+3	+4	-1	0.056
5	0	-2	-1	1
6	+3	-5	-1	0



Exercise

- Given the following classification data set with 16 examples, 2 input variables and 1 output variable, using a linear SVM with C=1, we have obtained the corresponding alpha values indicated in the last column.
  - What are the support vectors and which of them are in the margin boundary?
  - What are the hyperplane coefficients (beta, beta_0) and the value of M?
  - What are the values of epsilon?
  - Which examples are incorrectly classified?

Example	X_1	X_2	Y	alpha
1	2	6	1	0
2	4	3	1	1
3	4	4	1	0,3333
4	4	6	1	0
5	6	3	1	1
6	7	7	1	0,1667
7	8	4	1	1
8	9	8	1	1
9	2	1	-1	1
10	6	2	-1	0,5
11	7	4	-1	1
12	8	8	-1	1
13	9	1	-1	0
14	10	3	-1	0
15	10	6	-1	1
16	12	4	-1	0

Bibliography

- G. James, D. Witten, T. Hastie, y R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer, 2021.
  - Chapter 9
- T. Hastie, R. Tibshirani, y J. Friedman, The elements of statistical learning. Springer, 2009.
  - Chapter 12