SUPPORT VECTOR MACHINES (SVM)

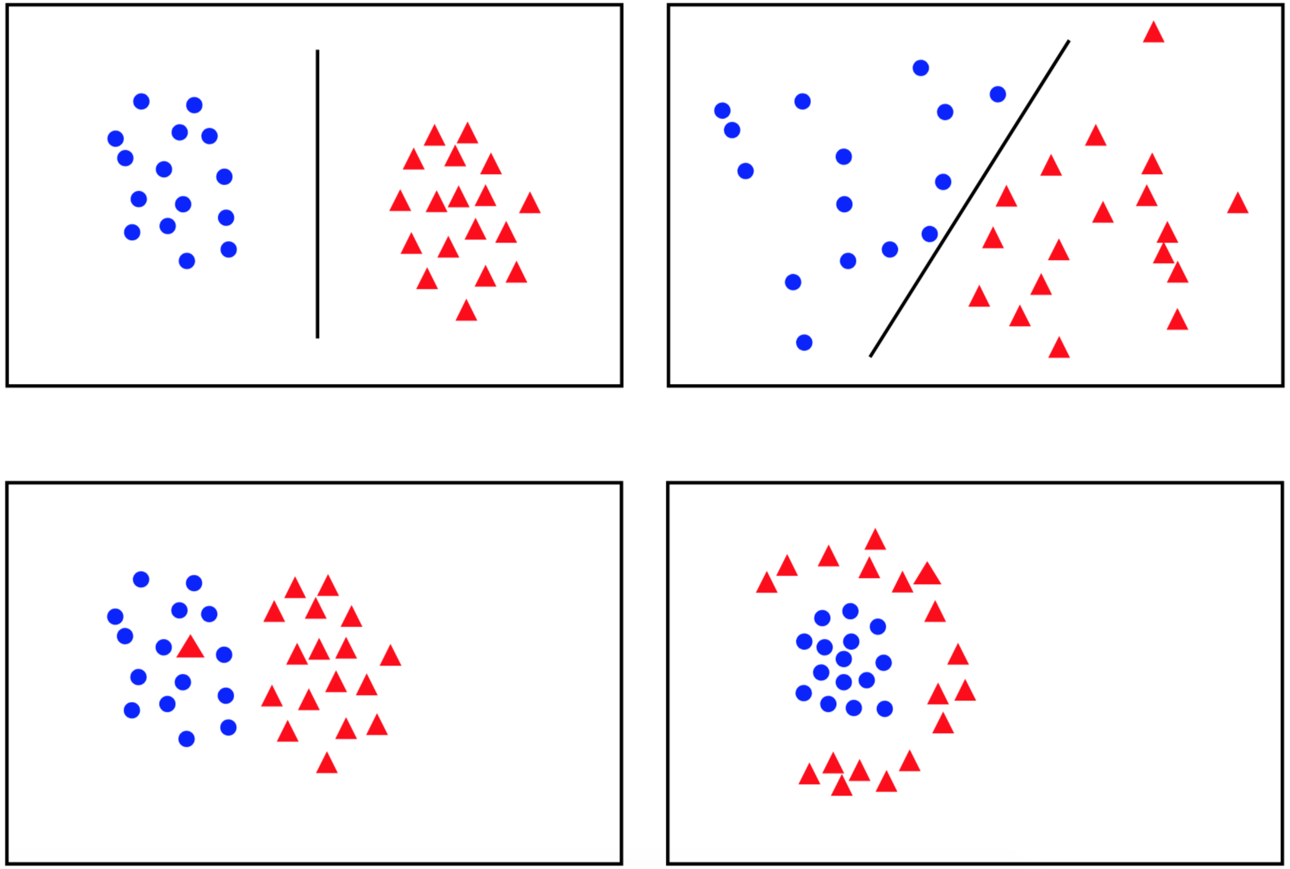
Introduction

- SVM is one of the most efficient machine learning algorithms
- A method used for regression (SVR) and classification (SVC)
  - mostly used in classification problems
- SVM is fundamentally a binary classifier, but can be extended for multiclass problems
- Classification performed by learning a linear separator of the data

Linear Separability

linearly separable

not linearly separable



Linear Classifier

Given training data $\{(x_i, y_i), 0 \leq i \leq n \; and \; x_i \epsilon \mathbb{R}^d\}$ and $y_i \epsilon \{-1,1\}$, learn a classifier $f(x)$ such that

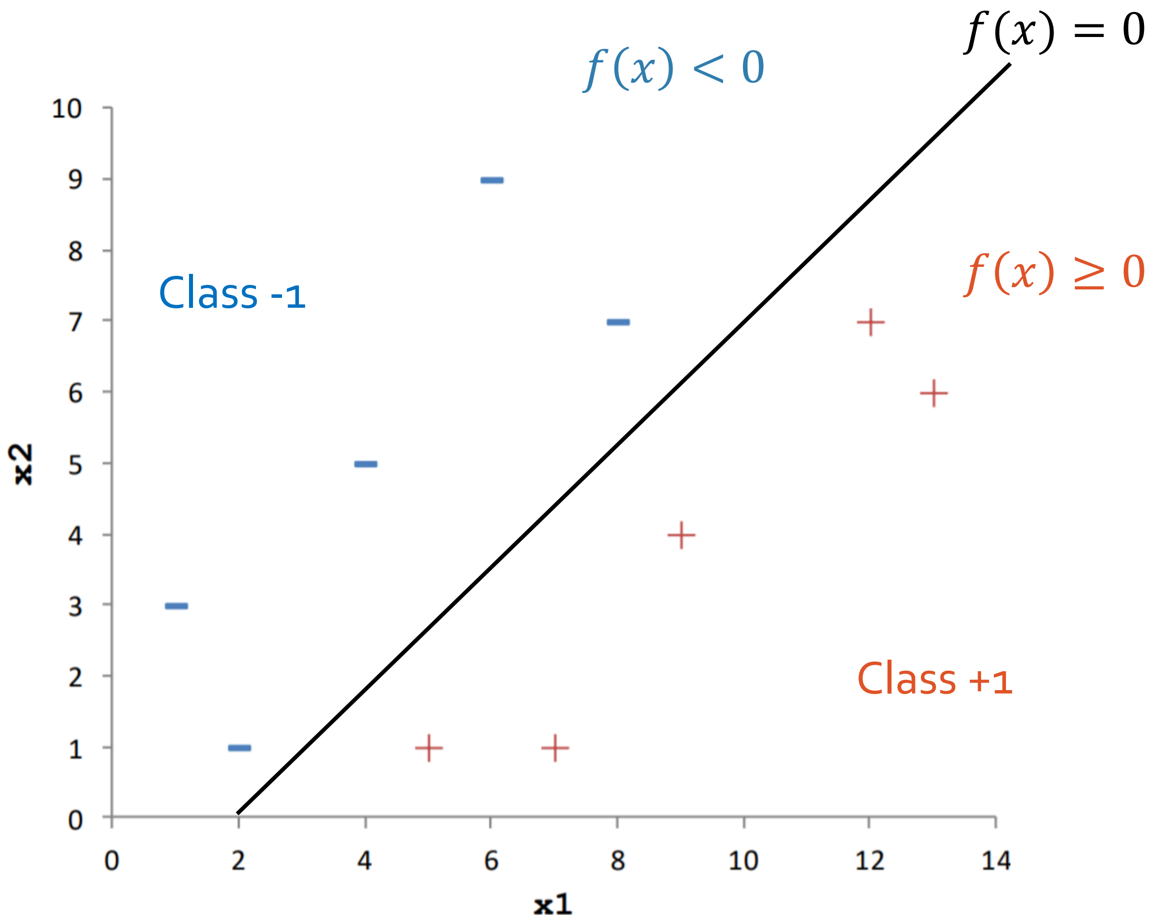
$$f(x_i) = \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

A linear classifier has the form (hyperplan)

$$f(x) = w^T x + b$$

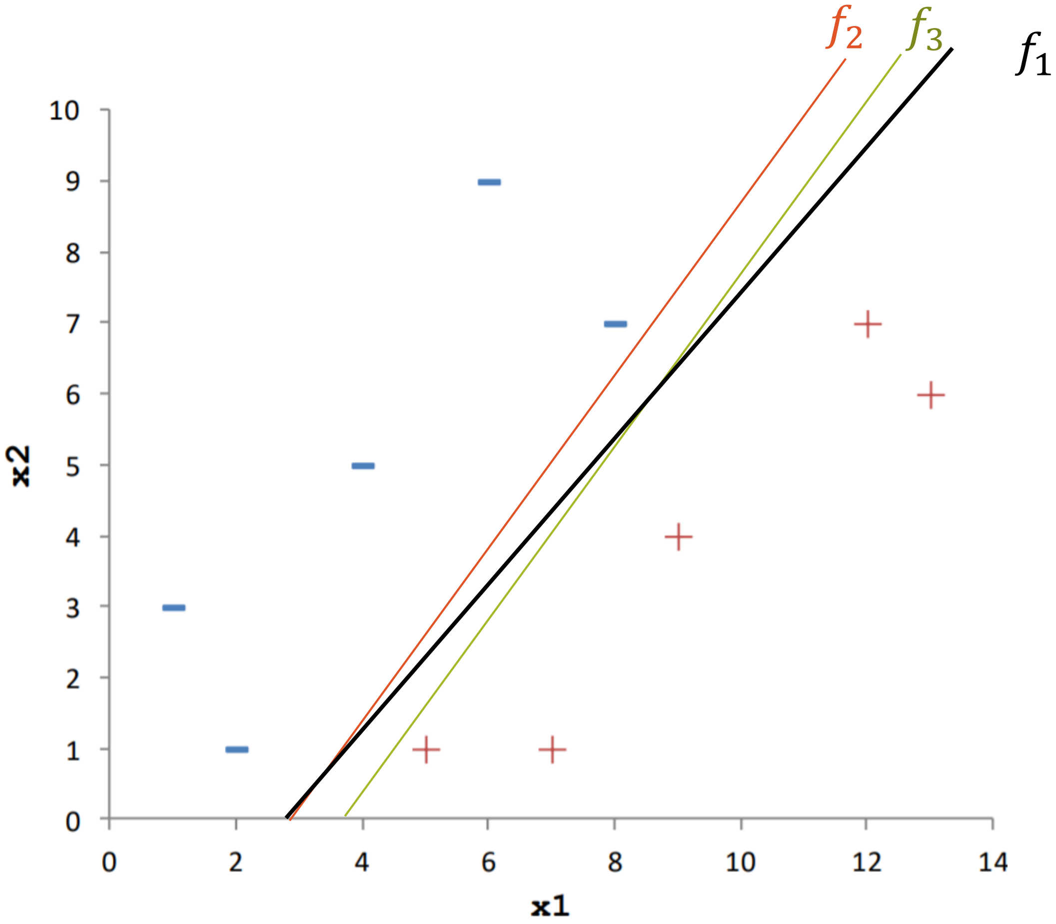
- w is the vector of weights and b is the bias

The goal is to find the vector of weights w that satisfies $y_i(w^T x_i + b) \geq 1$



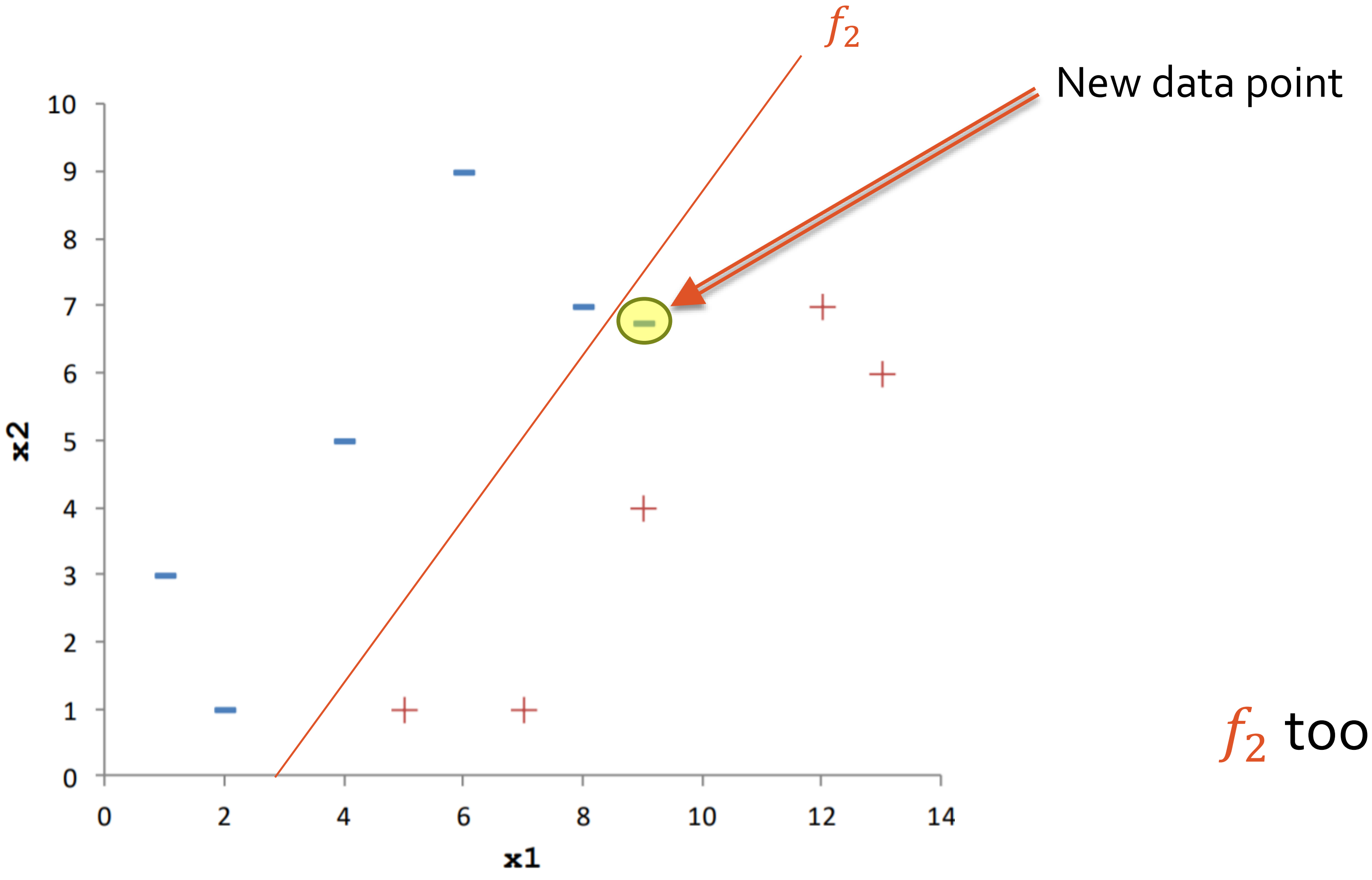
Linear Classifier

- There exists an infinite number of linear separators which one is optimal?



Linear Classifier

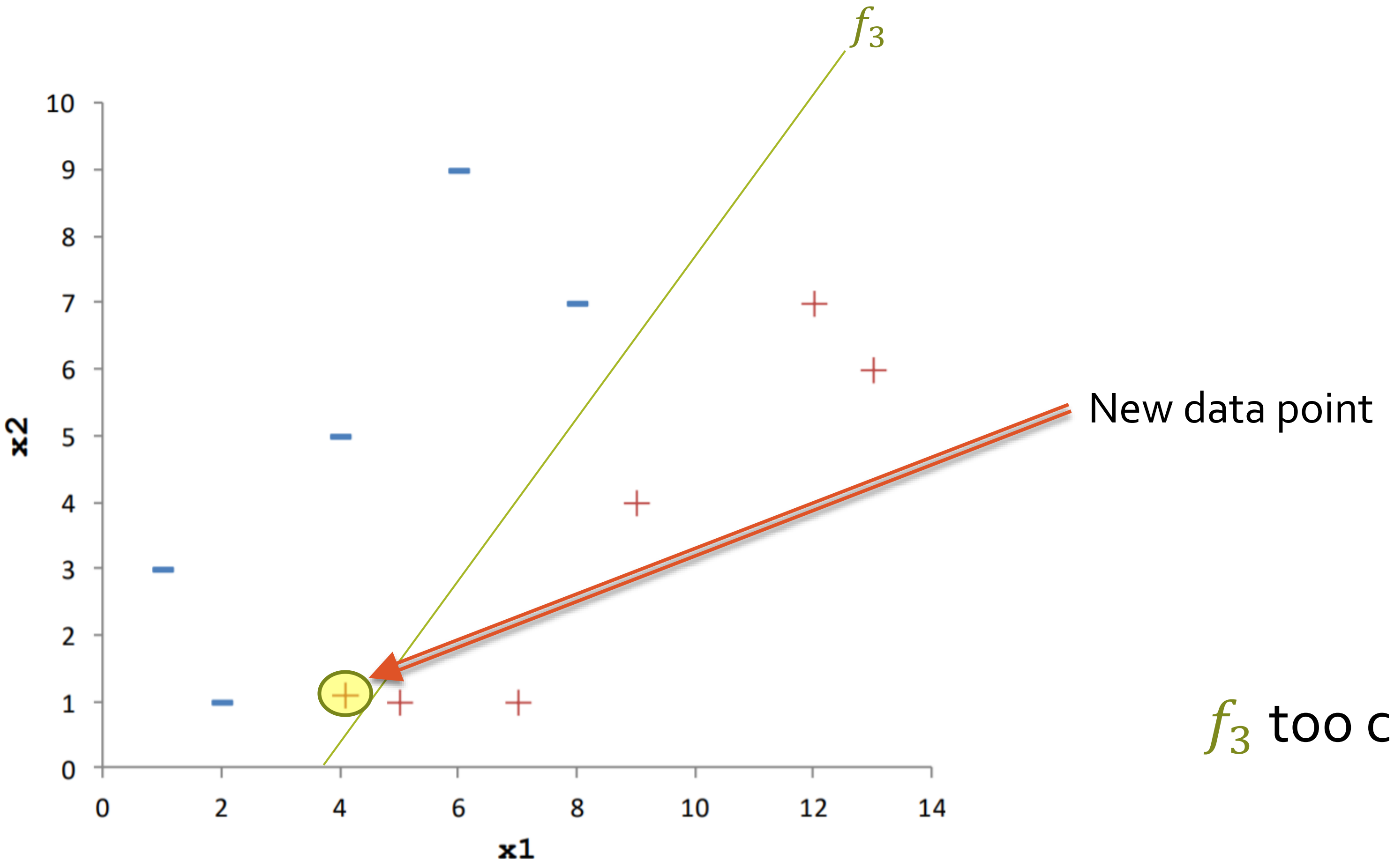
- There exists an infinite number of linear separators which one is optimal?



f_2 too close to training data

Linear Classifier

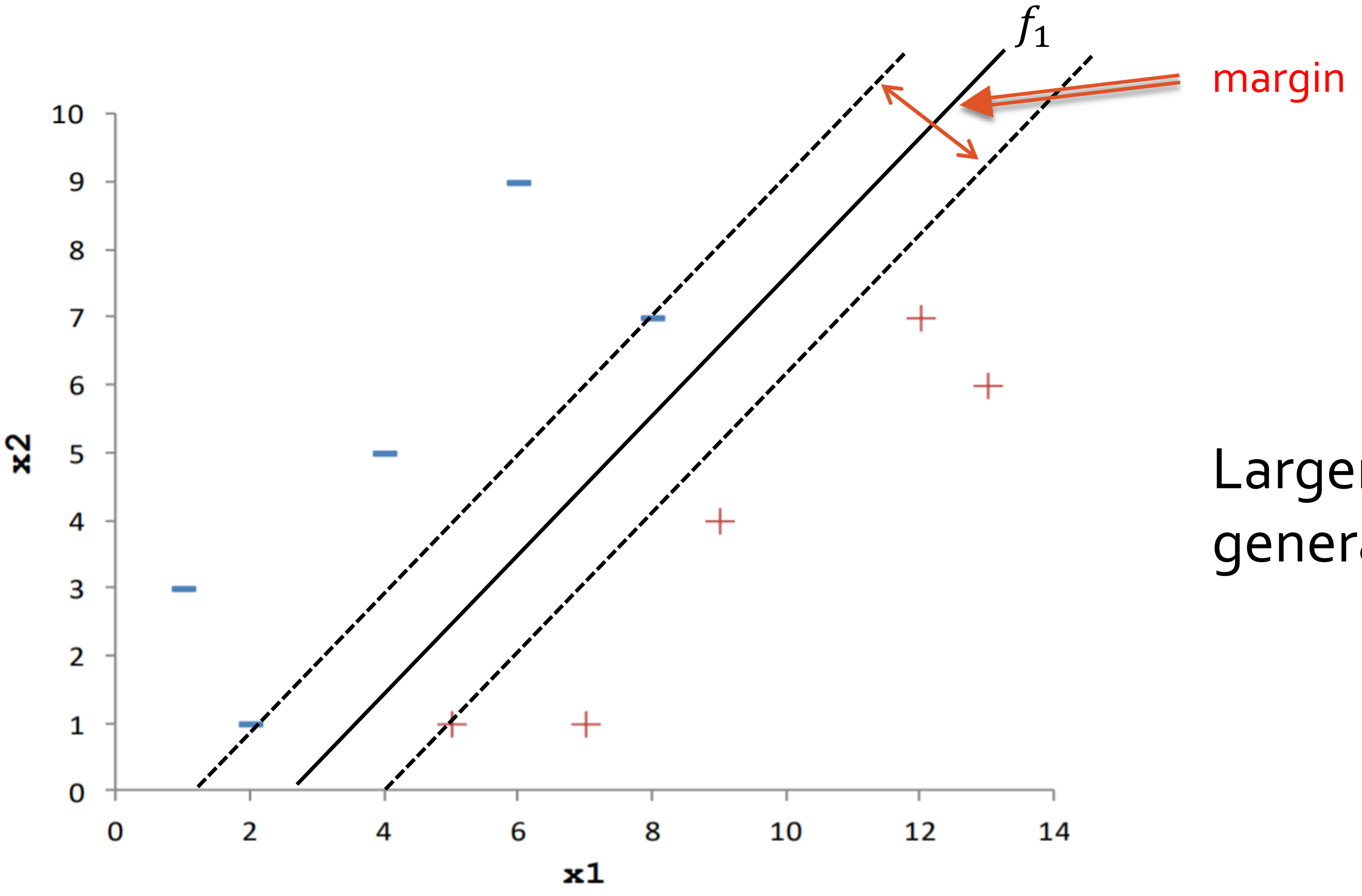
- There exists an infinite number of linear separators which one is optimal?



f_3 too close to training data

Linear Classifier

- There exists an infinite number of linear separators which one is optimal?

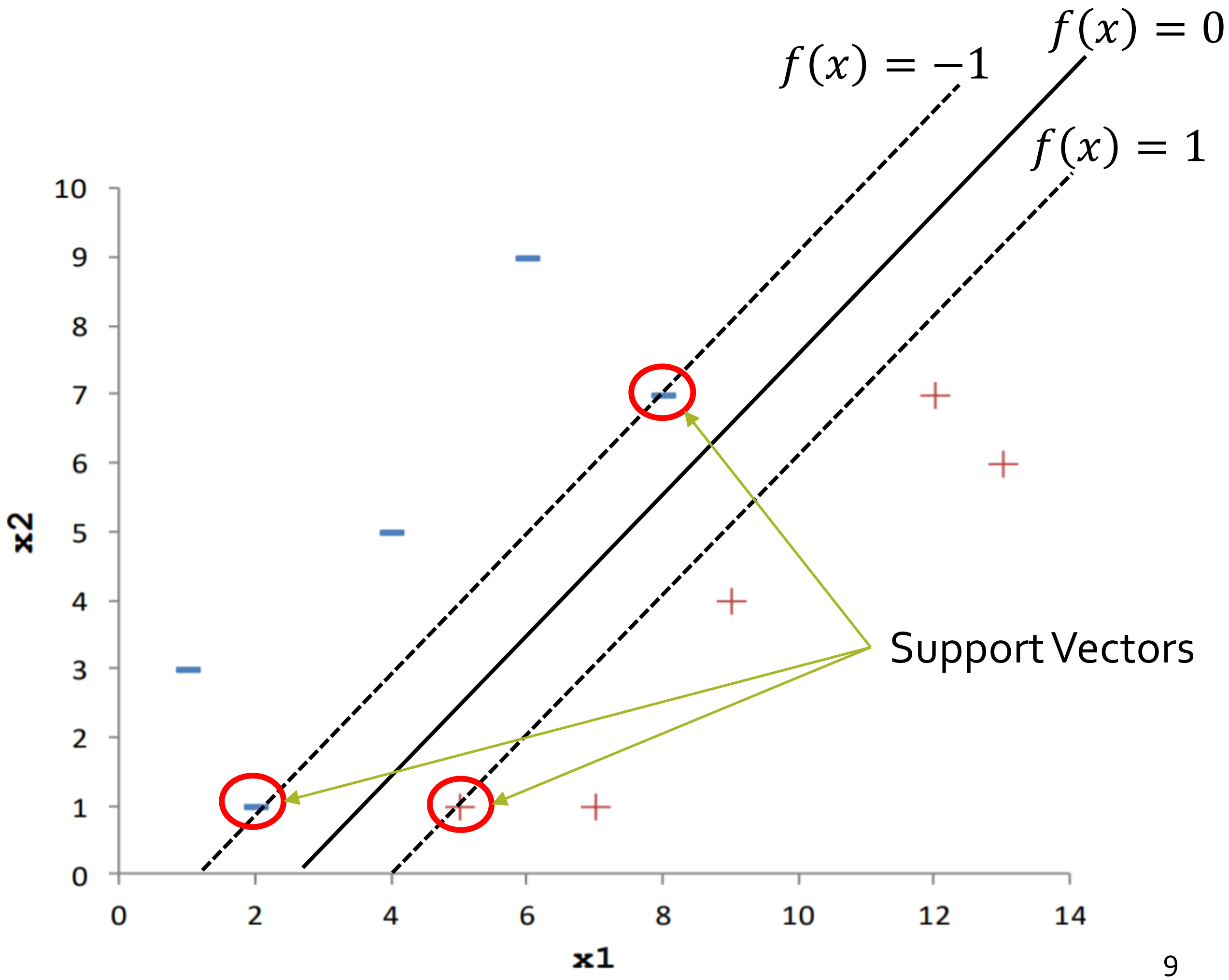


Larger margin is preferred for generalization

Support Vector Machines

- Find the optimal hyperplane that maximize the margin
- Distance from any point x_i to the hyperplane f is $d = \frac{|f(x_i)|}{||w||}$ (orthogonal projection)
- Width of the margin: $\gamma = \frac{2}{||w||}$
- For all x_i,

$$y_i f(x_i) = y_i(w^T x_i + b) \geq 1$$



Learning SVM

- Formulated as an optimization problem:

$$\max_{w} \frac{2}{||w]]} \qquad \mathit{subject\ to} \quad y_i(w^T x_i + b) \geq 1$$

- Equivalent to

$$\min_{w} \frac{1}{2} ||w||^2 \qquad \mathit{subject\ to} \quad y_i(w^T x_i + b) \geq 1$$

- This is a quadratic optimization problem subject to linear constraints and there is a unique minimum

Solving SVM

- **Primal problem:** for $w \in \mathbb{R}^d$ (d is the dimension of the feature vector x)

$$\min_w \frac{1}{2}||w||^2 \quad subject\ to \quad y_i(w^T x_i + b) \geq 1$$

$$\mathrm{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_i \alpha_i(y_i(w^T x_i + b) - 1)$$

$$(\frac{\partial L}{\partial b} = 0 \rightarrow \sum_i \alpha_i y_i = 0 \text{ , } \frac{\partial L}{\partial w} = 0 \rightarrow \mathrm{w} = \sum_i \alpha_i y_i x_i)$$

- **Dual problem** (lagrangien dual): for $\alpha \in \mathbb{R}^n$ (n is the number of training points)

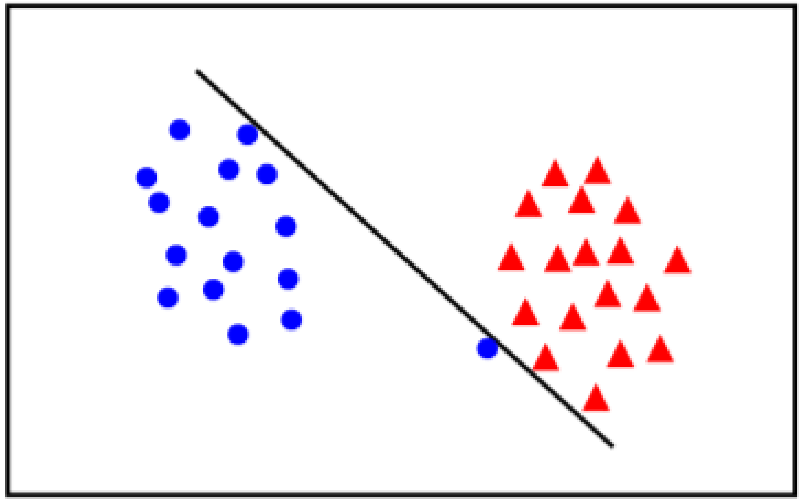
$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \quad subject\ to \quad \sum_i \alpha_i y_i = 0 \quad , \alpha_i \geq 0$$

Solving SVM

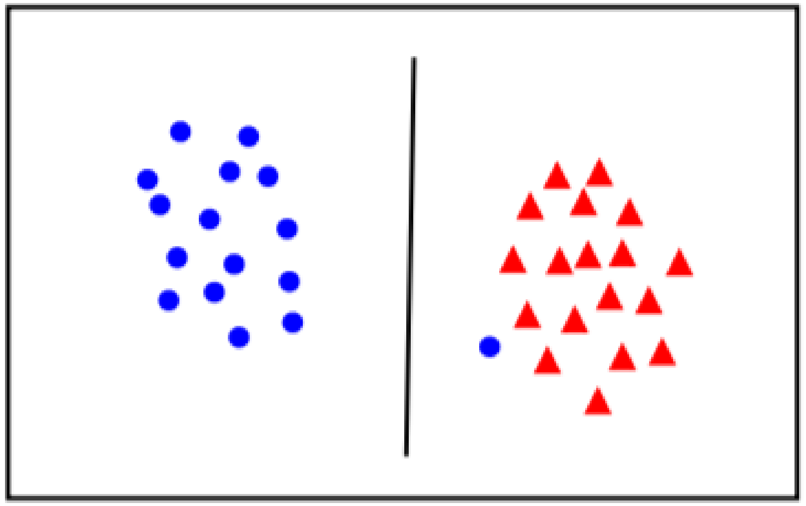
- The optimization problem is solved
  - for w in primal formulation
  - for α in dual formulation
- If $N << d$ then more efficient to solve for α than w
- Dual form only involves $(x_i x_j)$, which is very useful when working with kernels.

Linear separability

- What about mislabelled data and outliers?



Linearly separable but narrow margin



A large margin may be preffered even though some points are misclassified (the constraint is not satisfied)

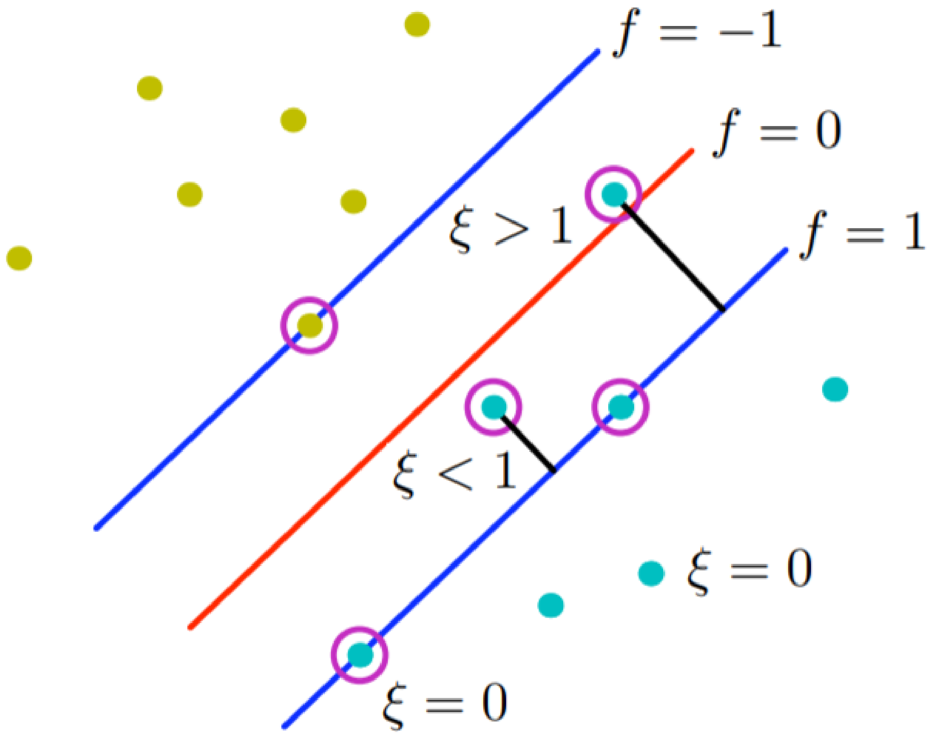
Solution: Loosen some of the constraints by introducing slack variables (soft margin)

Soft-margin classification

- Slack variable $\xi_i \geq 0$ for each datapoint i

$$y_i(w^T x_i + b) - (1 - \xi_i) \geq 0$$

- Errors occur if $\xi_i \geq 1$
  - $0 < \xi_i \leq 1$: point i is between margin and correct side of hyperplane. This is a margin violation
  - $\xi_i > 1$: point i is misclassified



Soft-margin classification

- Introduce a penalty for the errors
  - Primal form

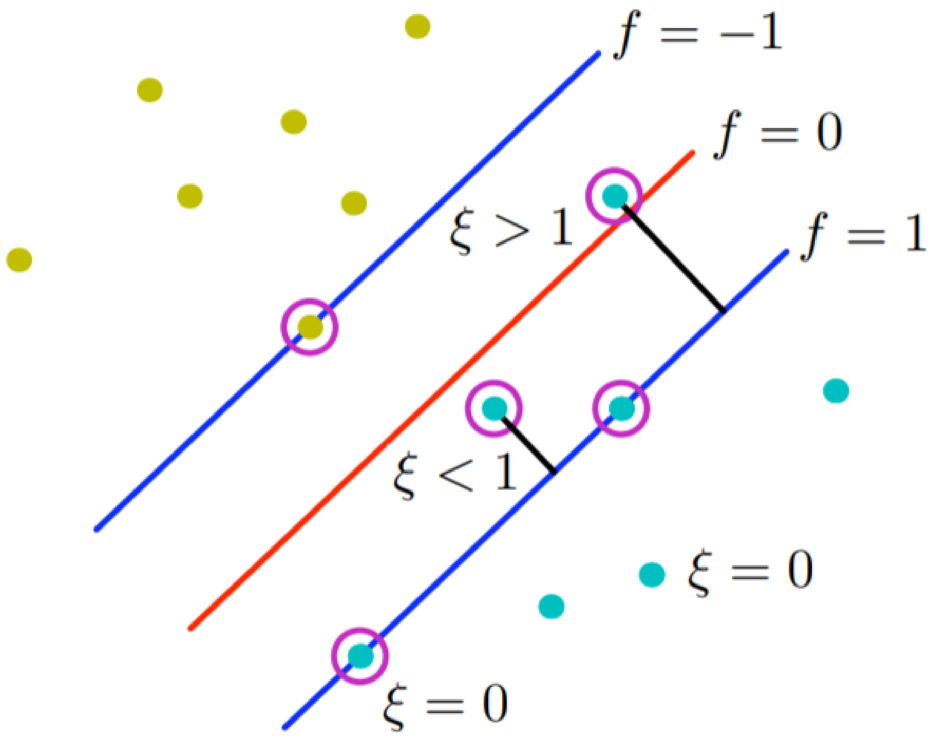
$$\min_{w\epsilon\mathbb{R}^d,\xi_i\epsilon\mathbb{R}^+} \frac{1}{2}||w||^2 + C\sum_{1}^{n}\xi_i$$

$$subject\ to\ y_i(w^T x_i + b) - (1 - \xi_i) \geq 0$$

  - Dual form

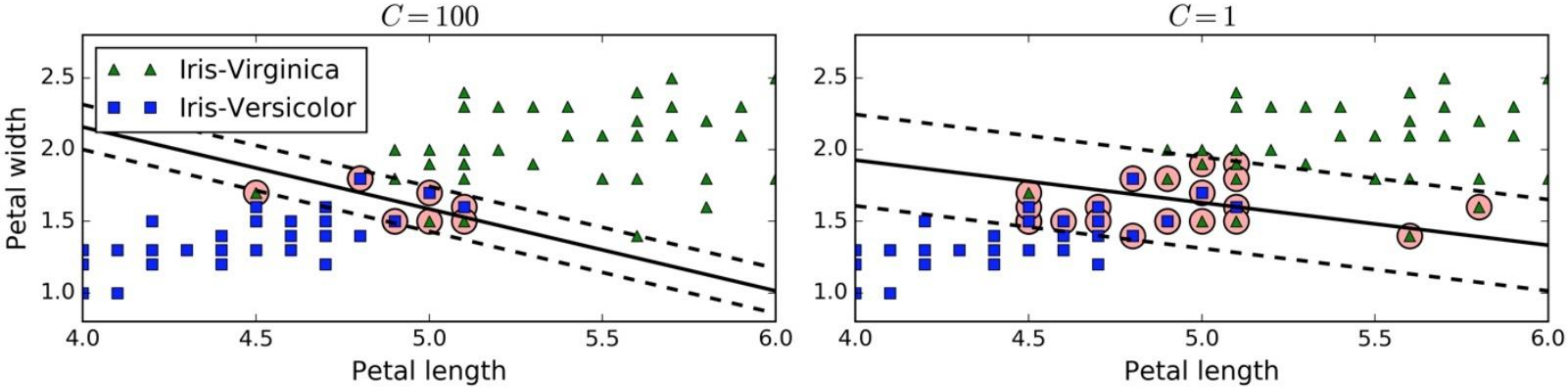
$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j$$

$$subject\ to\ \sum_i \alpha_i y_i = 0\ \ , C \geq \alpha_i \geq 0$$



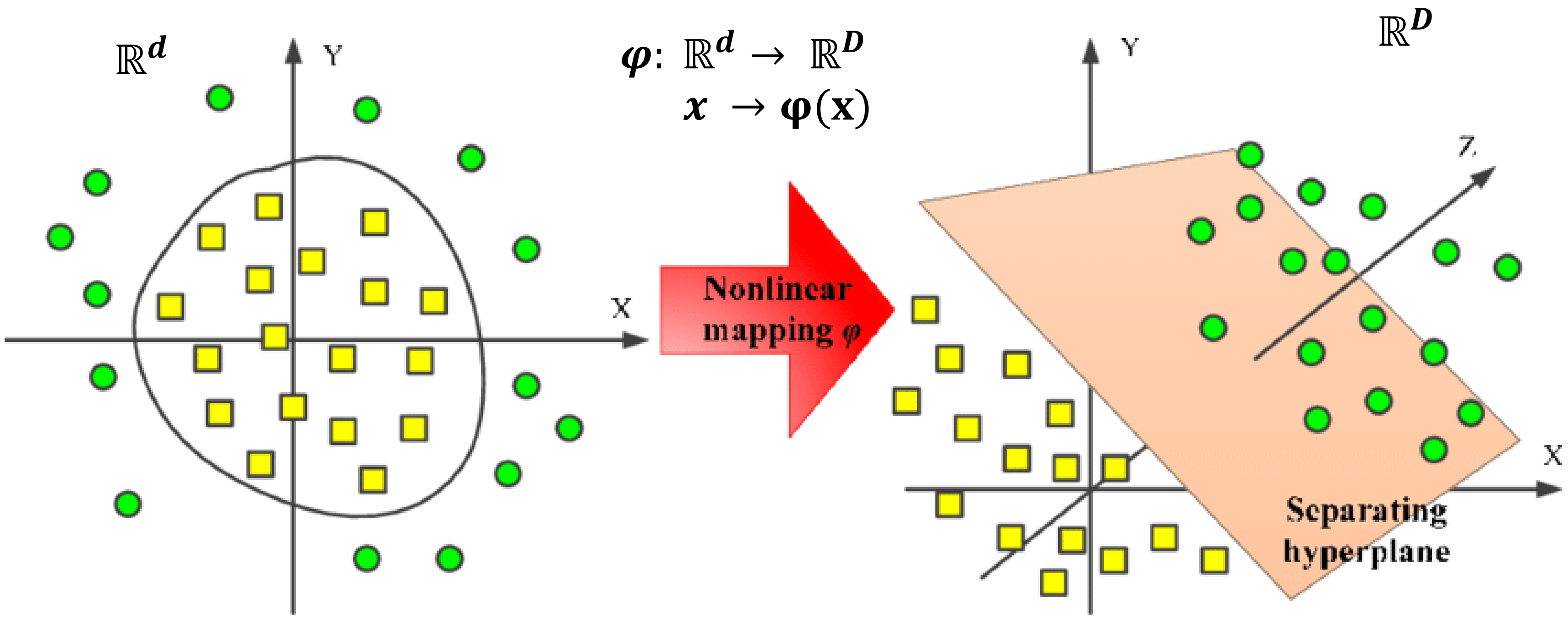
The hyperparameter C

- C is the penality parameter



- Small margins for higher values of C: better classification but may overfit
- Large margins for lower values of C: makes some prediction error but generalize well

Nonlinear problems

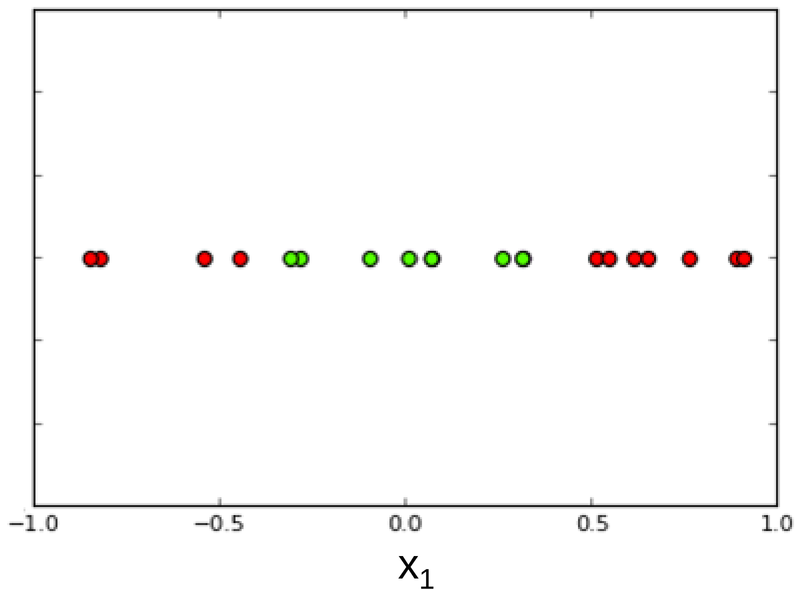


- Transform the data to a higher dimensional space where it can be separated by a linear hyperplane

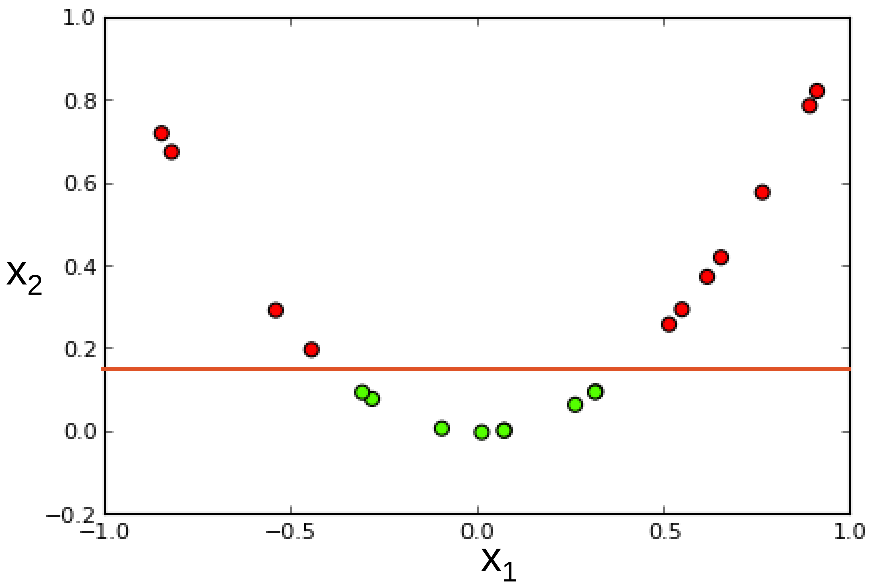
- Learn a linear classifier for the new space : $f(x) = w^T\varphi(x) + b$

Nonlinear problems

$\mathbb{R}^{\mathbf{1}}$



$\mathbb{R}^{\mathbf{2}}$



Add a feature $x_2=(x_1)^2$ to make the dataset linearly separable

The kernel trick

- Learning classifiers in high dimensions is very expensive
- Dual classifier in the transformed feature space:

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \varphi(x_i)\varphi(x_j) \quad subject\ to\ \sum_i \alpha_i y_i = 0\,, \qquad C \geq \alpha_i \geq 0$$

- $\varphi(x)$ only occurs as a product $\varphi(x_i)\varphi(x_j)$
- Classifier can be learnt without explicitly computing $\varphi(x)$
- Write $\mathrm{K}(x_j, x_i) = \varphi(x_j)\varphi(x_i)$. This is known as a Kernel

Common kernels

- Linear kernels

$$k(x, x_0) = x^T x_0$$

- Polynomial kernels

$$k(x, x_0) = (x^T x_0 + 1)^d \; for \; any \; d > 0$$

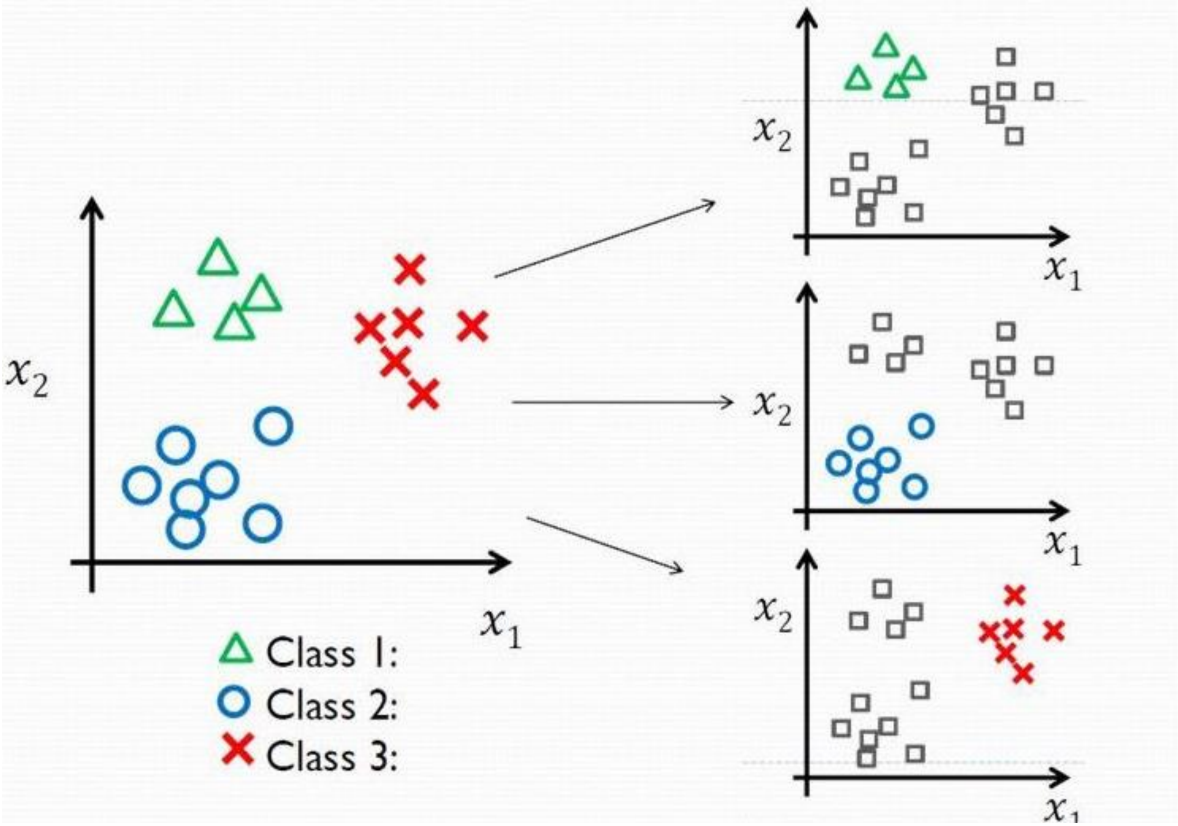
- Gaussian kernels

$$k(x, x_0) = \exp\left(-\frac{||x - x_0||^2}{2\sigma^2}\right) \; for \; \sigma > 0$$

Multiclass Classification

- **One against all SVM**
  - Train one binary SVM by class
  - For prediction, evaluate $w^T x + b$ and pick the largest.

$$y = \underset{j}{\operatorname{argmax}}\, {w_j}^T x + b$$



Multiclass Classification

- **Multiclass SVM**

$$\min_{w_j \epsilon \mathbb{R}^d, \xi_i \epsilon \mathbb{R}^+} \frac{1}{2} \sum_{j=1}^{L} ||w_j||^2 + C \sum_{1}^{n} \xi_i$$

$$subject\ to\ w_{y_i} x_i + b_{y_i} + \xi_i \geq w_j x_i + b_j + 1 \quad \forall i \epsilon \{1, .., n\}, \forall j \neq y_i$$

$$\xi_i \geq 0$$

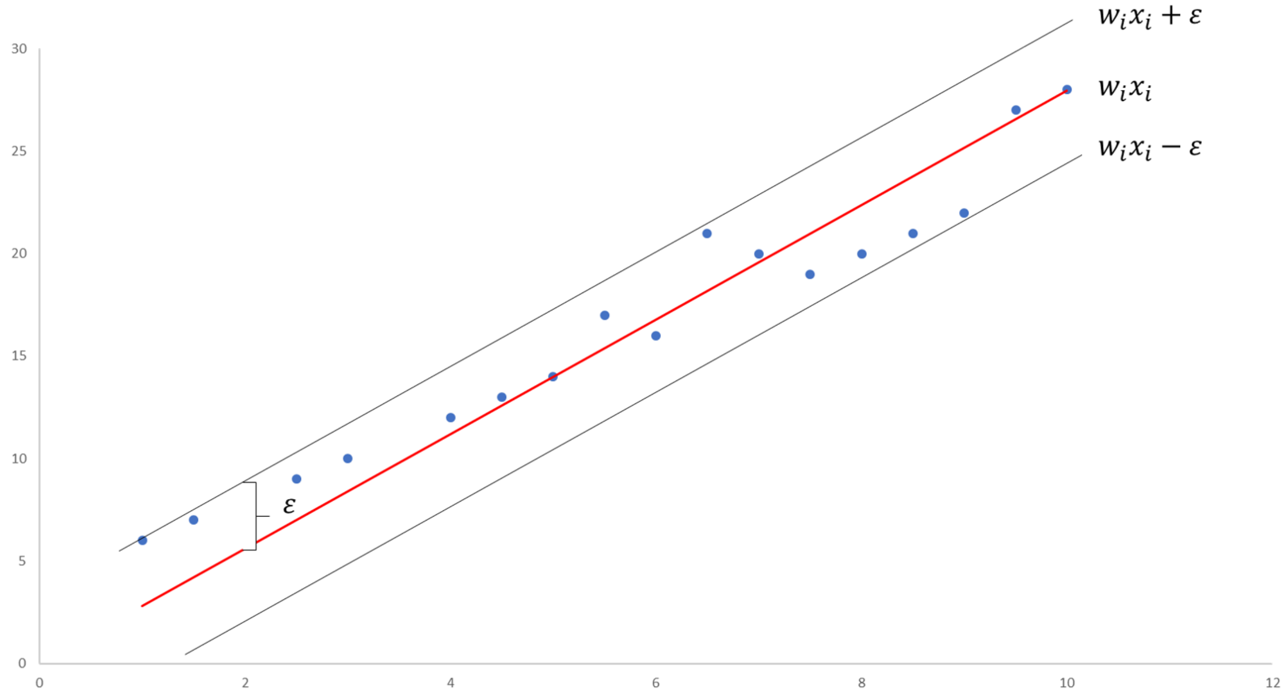
Key idea: suppose that point i belongs to class y_i. Then, for all $j \neq y_i$ we should make sure that $w_{y_i} x_i + b_{y_i}$ (the classifier for class y_i) is greater than $w_j x_i + b_j$ (the classifier for class j) by the largest margin

SVM regression

- Fit data points inside the margin. The width of the margin is controlled by ε

$$\min_{w} \frac{1}{2} ||w||^2$$

$$subject\ to \quad |y_i - w^T x_i| \leq \varepsilon$$



Conclusions

- Two key points of SVM:
  - Maximize the margin between classes using actual data points
  - Project the data into a higher dimensional space in which the data is linearly separable
- Hard margin vs soft margin
  - Soft margin makes SVM more robust to outliers
- SVM is model for classification and for regression
  - In classification, find an hyperplan that separate the classes with the largest margin
  - In regression, find an hyperplan that fit the data within a margin of a given width
- SVM is inherently a two class model but can be extended for multiclass problems
  - one vs all
  - Multiclass SVM