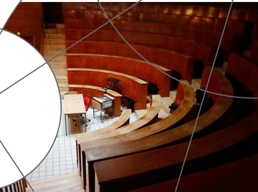
# Support Vector Machines

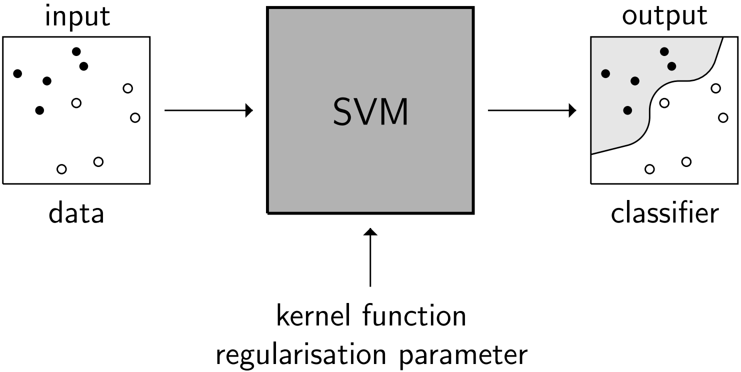
## Statistical Methods for Machine Learning

Christian Igel
Department of Computer Science

igel@diku.dk



# Binary Support Vector Machines

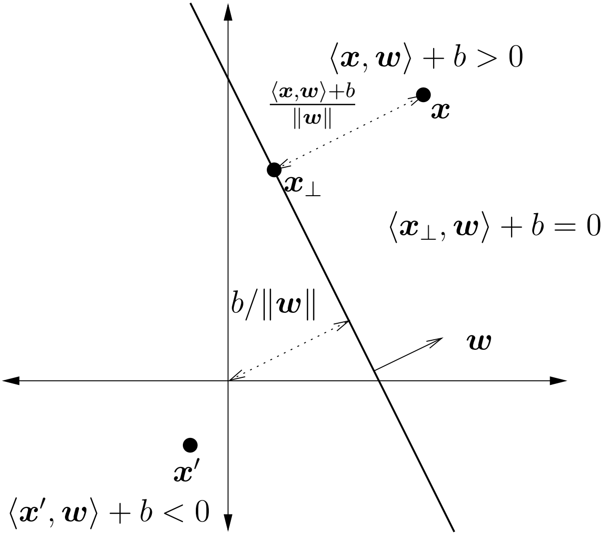


# Outline

1. Large margin classification
2. Linear soft-margin SVMs
3. Non-linear SVMs
4. Regularization and SVMs
5. Solving the SVM learning problem

# Recall: Linear decision functions

$f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + b$



# Recall: Margins

The *functional margin* of an example $(\boldsymbol{x}_i, y_i)$ with respect to a hyperplane $(\boldsymbol{w}, b)$ is

$$\gamma_i := y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \ .$$

and its *geometric margin* is

$$\rho_i := y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)/\|\boldsymbol{w}\| = \gamma_i/\|\boldsymbol{w}\| \ .$$

A positive margin implies correct classification.

The functional margin $\gamma_S$ of a hyperplane $(\boldsymbol{w}, b)$ with respect to a training set $S$ is $\min_i \gamma_i$.

# Recall: Separable data

$S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_\ell, y_\ell)\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ is linearly separable if there exists a hyperplane $(\boldsymbol{w}, b)$ such that for all $i = 1, \dots, \ell$

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0$$

which implies

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \gamma$$

for some $\gamma > 0$.

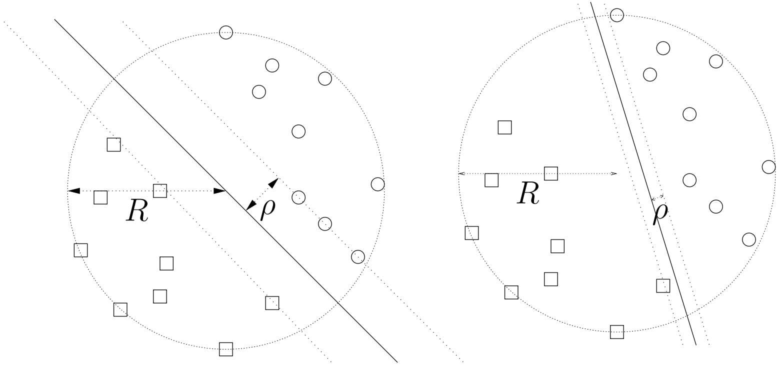
# Outline

# Support Vector Machines

We proceed in three steps:

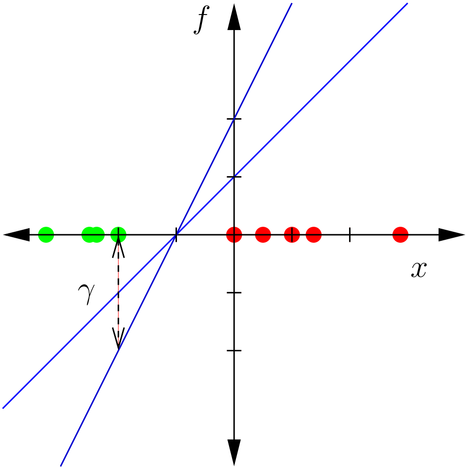
1. Linear hard margin SVMs: Large margin classification of linearly separable data
2. Soft margin SVMs: Dealing with outliers
3. Non-linear hard and soft margin SVMs: Using kernel trick to do classification in a feature space

# Large margins



# "Inherent degree of freedom"

Inherent degree of freedom: $(c\boldsymbol{w}, cb)$ leads to same decision boundary for all $c \in \mathbb{R}^+$



# Large margin classifier for separable data

Given linearly separable training data $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)\}$, we get rid of the inherent degree of freedom in

$$\begin{aligned} \text{maximize}_{\boldsymbol{w},b} \quad & \rho = \gamma / \|\boldsymbol{w}\| \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq \gamma \;\;, \;\; i = 1, \ldots, \ell \end{aligned}$$

by fixing $\gamma = 1$ (alternatively $\|\boldsymbol{w}\| = 1$)

$$\begin{aligned} \text{maximize}_{\boldsymbol{w},b} \quad & \rho = 1 / \|\boldsymbol{w}\| \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 \;\;, \;\; i = 1, \ldots, \ell \end{aligned}$$

is equal to:

$$\begin{aligned} \text{minimize}_{\boldsymbol{w},b} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 \;\;, \;\; i = 1, \ldots, \ell \end{aligned}$$

# Linear hard margin SVM primal

Given linearly separable data $S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_\ell, y_\ell)\}$ the hyperplane $(\boldsymbol{w}, b)$ solving

$$\begin{aligned} \text{minimize}_{\boldsymbol{w},b} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 \;\;, \;\; i = 1, \dots, \ell \end{aligned}$$

realizes the maximal margin hyperplane with margin $\rho = 1/\|\boldsymbol{w}\|$.

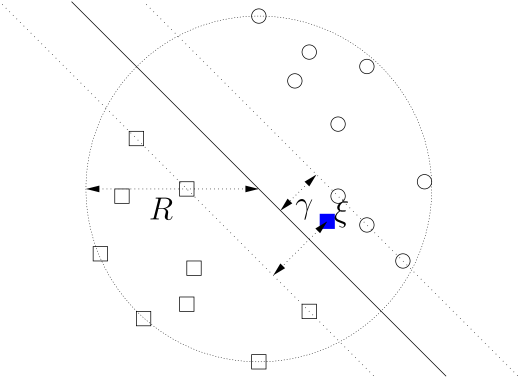
# Outline

1. Large margin classification
2. Linear soft-margin SVMs
3. Non-linear SVMs
4. Regularization and SVMs
5. Solving the SVM learning problem

# Tolerating margin violations: Slack variables

For a fixed value $\gamma > 0$, we can define the margin *slack variable* $\xi_i$ of an example $(\boldsymbol{x}_i, y_i)$ with respect to the hyperplane $(\boldsymbol{w}, b)$ and target margin $\gamma$ as

$$\xi((\boldsymbol{x}_i, y_i), (\boldsymbol{w}, b), \gamma) = \xi_i := \max(0, \gamma - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)) \ .$$



# 2-norm linear soft margin SVM primal

A quadratic penalty turns the hard margin SVM primal into:

Given $S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_\ell, y_\ell)\} \in (\mathbb{R}^d \times \{-1, 1\})^\ell$ and a regularization parameter $C \geq 0$, a 2-norm linear soft margin SVM computes an affine linear decision function $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + b$ by solving:

$$\begin{aligned} \text{minimize}_{\boldsymbol{\xi}, \boldsymbol{w}, b} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + \frac{C}{2} \sum_{i=1}^{\ell} \xi_i^2 \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i \;, \;\; i = 1, \dots, \ell \end{aligned}$$

# 1-norm linear soft margin SVM primal

Penalizing the absolute values of the slack variables gives:

Given $S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_\ell, y_\ell)\} \in (\mathbb{R}^d \times \{-1, 1\})^\ell$ and a regularization parameter $C \geq 0$, a 1-norm linear soft margin SVM computes an affine linear decision function $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + b$ by solving:

$$
\begin{aligned}
\text{minimize}_{\boldsymbol{\xi}, \boldsymbol{w}, b} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{\ell} \xi_i \\
\text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i \;, \;\; i = 1, \dots, \ell \\
& \xi_i \geq 0 \;, \;\; i = 1, \dots, \ell
\end{aligned}
$$

# Outline

1. Large margin classification
2. Linear soft-margin SVMs
3. Non-linear SVMs
4. Regularization and SVMs
5. Solving the SVM learning problem

# Non-linear SVMs

- In SVMs elements from $\mathcal{X}$ occur only in scalar products – we can apply the "kernel trick"!
- Consider an arbitrary input space $\mathcal{X}$ and a feature map $\Phi : \mathcal{X} \to \mathcal{H}_k$, where $\mathcal{H}_k$ is the RKHS induced by kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ .
- The non-linear SVM learns decision functions of the form

  $$f(x) = \langle \Phi(x), \boldsymbol{w} \rangle + b \ .$$

  Here $\boldsymbol{w} \in \mathcal{H}_k$ and $\Phi : x \mapsto k(x, \cdot)$.
- The scalar product $\langle \Phi(x), \boldsymbol{w} \rangle$ will be computed using the kernel trick.

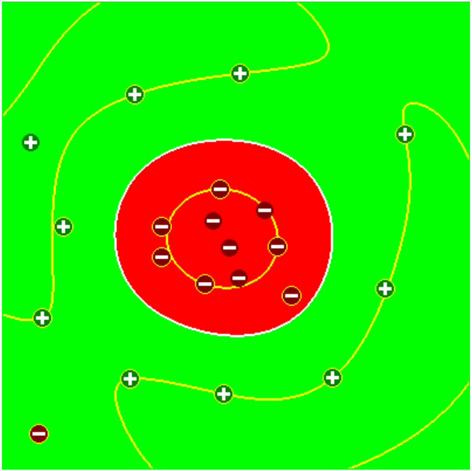
# 1-norm non-linear soft margin SVM primal

Given $S = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\} \in (\mathcal{X} \times \{-1, 1\})^\ell$ a regularization parameter $C \geq 0$, and a kernel $k$ on $\mathcal{X}$, a 1-norm soft margin SVM computes a linear decision function $f(x) = \langle \Phi(x), \boldsymbol{w} \rangle + b$ by solving:

$$\begin{aligned}
\text{minimize}_{\boldsymbol{\xi}, \boldsymbol{w}, b} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{\ell} \xi_i \\
\text{subject to} \quad & y_i(\langle \boldsymbol{w}, \Phi(x_i) \rangle + b) \geq 1 - \xi_i \ , \quad i = 1, \dots, \ell \\
& \xi_i \geq 0 \ , \quad i = 1, \dots, \ell \ ,
\end{aligned}$$

where $\Phi(x) = k(x, \cdot)$.

# Regularization and kernel representation



Kernel $k$: Represent data for linear classification (ideally, $h^{\text{Bayes}} \in \mathcal{H}_k^b$)
Slack variables: Deal with noise and outliers (i.e., $\mathcal{R}_p^{\text{Bayes}} > 0$)

# Outline

1. Large margin classification
2. Linear soft-margin SVMs
3. Non-linear SVMs
4. Regularization and SVMs
5. Solving the SVM learning problem

# 1-norm soft margin SVM and regularization I

- 1-norm soft margin SVM, primal

$$\begin{aligned} \text{minimize}_{\boldsymbol{\xi},\boldsymbol{w},b} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i \ , \ \ i = 1,\dots,\ell \\ & \xi_i \geq 0 \ , \ \ i = 1,\dots,\ell \end{aligned}$$

- For fixed $\boldsymbol{w}$ optimal slack variables are

$$\xi_i = \max(0, 1 - y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b))$$

- Loss $L_{\text{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$, $y \in \{-1, 1\} \subset \mathbb{R}$, $\hat{y} \in \mathbb{R}$
- Hypothesis classes
  - $\mathcal{H}_k$: RKHS induced by $k$
  - $\mathcal{H}_k^b = \{f(x) = g(x) + b \,|\, g \in \mathcal{H}_k, b \in \mathbb{R}\}$

# 1-norm soft margin SVM and regularization II

- Consider loss $L_{\text{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$ and hypothesis classes $\mathcal{H}_k$ and $\mathcal{H}_k^b = \{f(x) = g(x) + b \,|\, g \in \mathcal{H}_k, b \in \mathbb{R}\}$
- 1-norm soft margin SVM

$$\begin{aligned} \text{minimize}_{\boldsymbol{\xi}, \boldsymbol{w}, b} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{\ell} \xi_i \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i) \rangle + b) \geq 1 - \xi_i \ , \quad i = 1, \dots, \ell \\ & \xi_i \geq 0 \ , \quad i = 1, \dots, \ell \end{aligned}$$

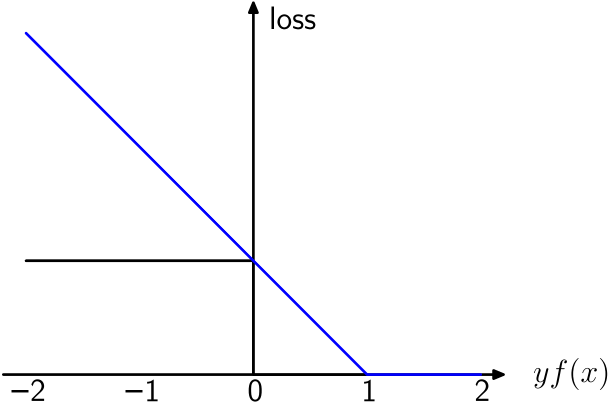
corresponds to

$$\text{minimize}_{f \in \mathcal{H}_k^b} \quad \frac{1}{\ell} \sum_{i=1}^{\ell} L_{\text{hinge}}(y_i, f(x_i)) + \gamma_\ell \|f\|_k^2$$

where $\gamma_\ell = (2\ell C)^{-1}$ and $\|.\|_k$ inherited from $\mathcal{H}_k$ to $\mathcal{H}_k^b$ is only a semi-norm

# Hinge loss as convex surrogate for 0-1 loss

$$L_{\mathsf{hinge}}(y, f(x)) = [1 - yf(x)]_+ = \max(0, 1 - yf(x))$$




0-1 loss applied to $\operatorname{sgn}(f(x))$ and hinge loss

# Inspecting the SVM solution I

- Representer theorem can be applied to SVMs and tells us that the solution must have the form

$$f(x) = \sum_{i=1}^{\ell} \beta_i k(x_i, x) + b \ .$$

- We have $\boldsymbol{w} = \sum_{i=1}^{\ell} \beta_i k(x_i, \cdot)$ and $\langle \Phi(x), \boldsymbol{w} \rangle = \langle k(x, \cdot), \boldsymbol{w} \rangle = \sum_{i=1}^{\ell} \beta_i k(x_i, x)$.
- Typically, many $\beta_i$ are zero. The training patterns corresponding to the non-zero coeffs are the support vectors. With $\mathsf{SV} = \{i \mid \beta_i \neq 0\}$ the decision function is

$$f(x) = \sum_{i \in \mathsf{SV}} \beta_i k(x_i, x) + b \ .$$

- Each coefficient can be written as $\beta_i = y_i \alpha_i$, with $\alpha_i = |\beta_i|$.

# Outline

1. Large margin classification
2. Linear soft-margin SVMs
3. Non-linear SVMs
4. Regularization and SVMs
5. Solving the SVM learning problem

# Solution strategy

Typically, we do not solve the non-linear SVM problems directly, but the corresponding dual problems.

1. We derive the Lagrangian. The constraints give rise to Lagrange multipliers $\alpha_1, \dots, \alpha_\ell$, the dual variables.
2. The Karush-Kuhn-Tucker (KKT) theorem gives us necessary and sufficient conditions for an optimum.
3. We set the derivatives of the Lagrangian w.r.t. the primal ("original") variables to zero; solve analytically w.r.t. primal variables; and substitue primal variables into Lagrangian.
4. The Lagrangian is maximized with w.r.t. dual variables.

# 1-norm soft margin SVM, dual form

For $\{x_1, y_1), \dots, (x_\ell, y_\ell)\}$ and kernel $k$ solving

$$\text{maximize}_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

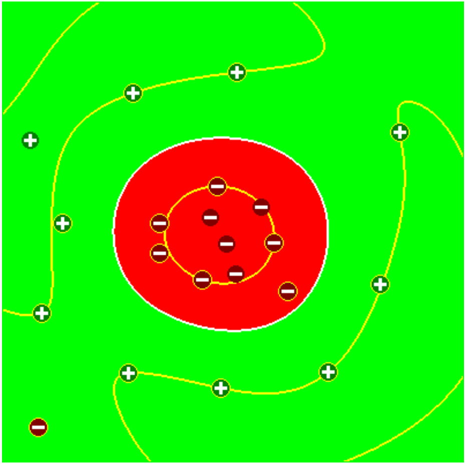
$$\text{subject to} \quad \sum_{i=1}^{\ell} \alpha_i y_i = 0 \;, \quad C \geq \alpha_i \geq 0 \;, \quad i = 1, \dots, \ell$$

leads to the decision rule $h(x) = \operatorname{sgn}(f(x))$ with $f(x) = \sum_{i=1}^{\ell} y_i \alpha_i^* k(x_i, x) + b^*$, where $b^*$ is chosen so that $y_i f(x_i) = 1$ for any $i$ with $C > \alpha_i > 0$ and the slack variables of the "corresponding hyperplane" in $\mathcal{H}_k^b$ are defined relative to the margin $\rho = 1/\|\boldsymbol{w}^*\| = 1/\sqrt{\sum_{x_i, x_j \in \text{SV}} y_i y_j \alpha_i^* \alpha_j^* k(x_i, x_j)}$.

# Notes on SVM optimization and solution

- SVM training is a convex constraint optimization problem, more precisely a quadratic program, with $\ell$ variables.
- Training always converges to an optimal solution.
- The solution is sparse, because of zero coefficients $\alpha_i$ in the kernel expansion. The coefficients of training patterns that lie directly on the margin or are misclassified are non-zero.
- Optimization time scales between quadratically and cubically in $\ell$.
- For 1-norm soft-margin SVMs, the parameter $C$ is an upper bound on the magnitude of the coefficients in the kernel expansion.

# Inspecting the solution II



Bounded SV: $\alpha_i = C,\ \xi_i \geq 0,\ y_i f(x_i) \leq 1$

Free SV: $0 < \alpha_i < C,\ \xi_i = 0,\ y_i f(x_i) = 1$

Non-SV: $\alpha_i = 0,\ \xi_i = 0,\ y_i f(x_i) > 1$

# Binary SVMs

input

output

Cortes, Vapnik: Support-Vector Networks, *Machine Learning* 20(3):273–297, 1995

$\{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$

$$f(x) = \sum_{i=1}^{\ell} \beta_i k(x_i, x) + b$$

kernel function $k$

regularisation parameter $C$

$\beta_i, \dots, \beta_\ell \ , b \in \mathbb{R}$

$$\underset{f \in \mathcal{H}_k^b}{\text{minimise}} \frac{1}{\ell} \sum_{i=1}^{\ell} L_{\text{hinge}}(y_i, f(x_i)) + \frac{1}{2C\ell} \|f\|_{\mathcal{H}_k}^2$$