Faculty of Science

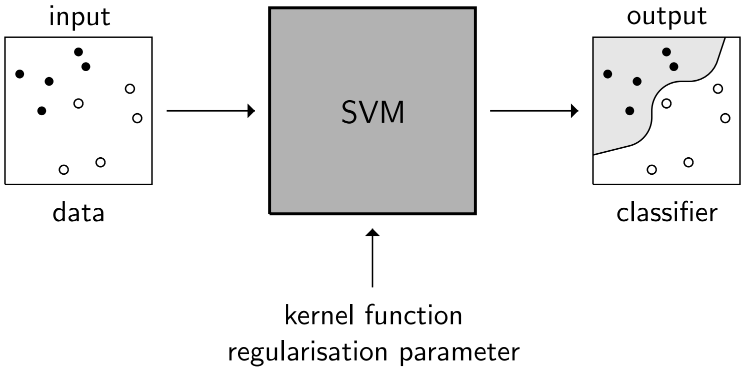
Support Vector Machines

Statistical Methods for Machine Learning

Christian Igel
igel@diku.dk

Department of Computer Science
University of Copenhagen

Binary Support Vector Machines



Support Vector Machines

We proceed in three steps:

1. Linear hard margin SVMs: large margin classification of linearly separable data
2. Non-linear hard margin SVMs: large margin classification of linearly separable data in feature space
3. Linear and non-linear soft margin SVMs: large margin classification of general data

Recall: Margins

Definition

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

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

The geometric margin of an example $(\boldsymbol{x}_i, y_i)$ with respect to a hyperplane $(\boldsymbol{w}, b)$ 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 geometric margin ρ_S of a hyperplane $(\boldsymbol{w}, b)$ with respect to a training set S is $\min_i \rho_i$.
The functional margin γ_S of a hyperplane $(\boldsymbol{w}, b)$ with respect to a training set S is $\min_i \gamma_i$.

Recall: Separable data

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

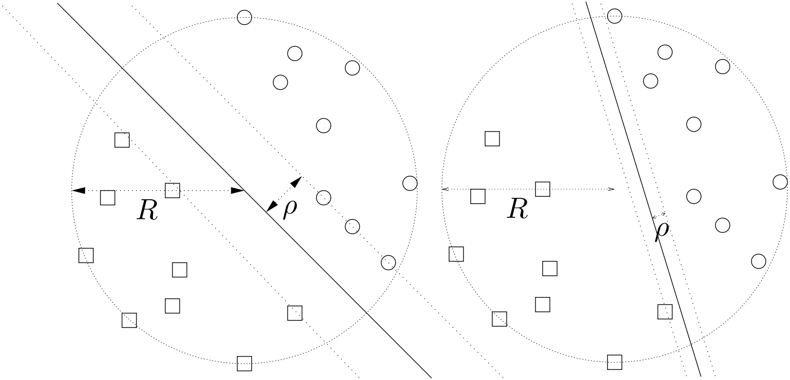
$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0$$

which implies

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

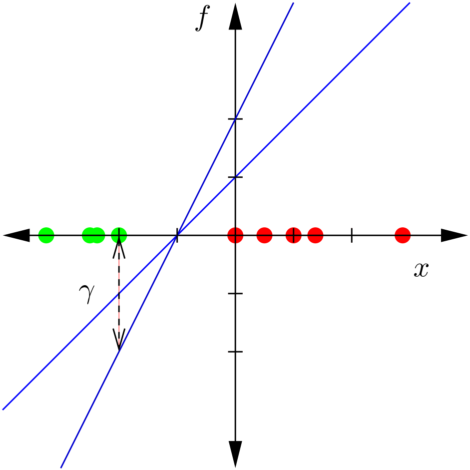
for some $\gamma > 0$

Large margins



"Inherent degree of freedom"

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



Linear large margin classifier for separable data

Given linearly separable training data $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)\}$

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

Getting rid of inherent degree of freedom by fixing $\gamma = 1$ (alternatively $\|\boldsymbol{w}\| = 1$)

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

is equal to

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

Linear hard margin SVM, primal form

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

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

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

Linear hard margin SVM, dual form

Primal form:

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

Dual form:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha}} \quad & \inf_{\boldsymbol{w},b} \quad L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \\ \text{subject to} \quad & \alpha_i \geq 0 \;,\; i = 1, \dots, \ell \end{aligned}$$

with Lagrangian:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{\ell} \alpha_i [y_i(\langle \boldsymbol{w}, x_i \rangle + b) - 1]$$

Linear hard margin SVM, KKT

$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{\ell} \alpha_i [y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1]$

Karush-Kuhn-Tucker (KKT) theorem requires

$$\frac{\partial}{\partial \boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = 0 \qquad \frac{\partial}{\partial b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = 0$$

yielding

$$\frac{\partial}{\partial \boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \boldsymbol{w} - \sum_{i=1}^{\ell} \alpha_i y_i \boldsymbol{x}_i \quad \text{and} \quad \frac{\partial}{\partial b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i y_i$$

implying

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i y_i \boldsymbol{x}_i \quad \text{and} \quad 0 = \sum_{i=1}^{\ell} \alpha_i y_i$$

Linear hard margin SVM

using $\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i y_i \boldsymbol{x}_i$ gives

$$
\begin{aligned}
L(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle - \sum_{i=1}^{\ell} \alpha_i [y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1] \\
&= \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \sum_{i=1}^{\ell} \alpha_i \\
&= \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle
\end{aligned}
$$

Linear hard margin SVM

Linear Hard Margin SVM: For linearly separable data $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)\}$ the solution of

$$\text{maximize}_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

$$\text{subject to} \quad \sum_{i=1}^{\ell} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \ , \quad i = 1, \ldots, \ell$$

leads to the maximal margin hyperplane with margin $\rho = 1/\|\boldsymbol{w}^*\|$ using

$$\boldsymbol{w}^* = \sum_{i=1}^{\ell} \alpha_i y_i \boldsymbol{x}_i$$

$$b^* = -\frac{\max_{y_i=-1}(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle) + \min_{y_i=1}(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle)}{2} \ .$$

KKT complementarity condition I

- Karush-Kuhn-Tucker (KKT) complementarity condition requires
$$\alpha_i^*[y_i(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^*) - 1] = 0$$
for $i = 1, \ldots, \ell$
- KKT condition can be used to compute b^*
- Solution is sparse

$$\mathsf{SV} = \{\boldsymbol{x}_i \,|\, \alpha_i^* \neq 0\}$$

$$f(\boldsymbol{x}, \boldsymbol{\alpha}^*, b^*) = \sum_{i=1}^{\ell} y_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b^* = \sum_{\boldsymbol{x}_i \in \mathsf{SV}} y_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b^*$$

KKT complementarity condition II

For $\boldsymbol{x}_j \in \mathsf{SV}$

$$y_j f(\boldsymbol{x}_j, \boldsymbol{\alpha}^*, b^*) = y_j \left(\sum_{\boldsymbol{x}_i \in \mathsf{SV}} y_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + b^* \right) = 1$$

and therefore

$$\begin{aligned}
\langle \boldsymbol{w}^*, \boldsymbol{w}^* \rangle &= \sum_{i,j=1}^{\ell} y_i y_j \alpha_i^* \alpha_j^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\
&= \sum_{\boldsymbol{x}_j \in \mathsf{SV}} \alpha_j^* y_j \sum_{\boldsymbol{x}_i \in \mathsf{SV}} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\
&= \sum_{\boldsymbol{x}_j \in \mathsf{SV}} \alpha_j^* (1 - y_j b^*) = \sum_{\boldsymbol{x}_j \in \mathsf{SV}} \alpha_j^*
\end{aligned}$$

Recall: Kernel trick

Kernel trick

Given an algorithm formulated in terms of a positive definite kernel k (e.g., the std. scalar product $\langle ., . \rangle$), one can construct an alternative algorithm by replacing k by an alternative kernel.

Hard margin SVM

Hard Margin SVM: For training data $S = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$ linearly separable in the feature space defined by the kernel k the solution $\boldsymbol{\alpha}^*$, b^* of

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

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

leads to the decision rule $\operatorname{sign}(f(x))$ with

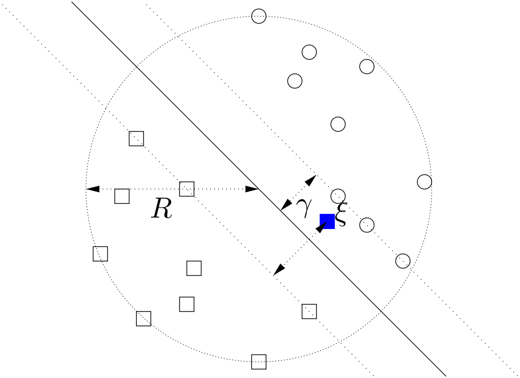
$$f(x) = \sum_{i=1}^{\ell} y_i \alpha_i^* k(x_i, x) + b^*$$

that is equivalent to the maximal margin hyperplane in the feature space defined by kernel k with margin $\rho = 1/\|\boldsymbol{w}^*\| = 1/\sqrt{\sum_{x_j \in \mathsf{SV}} \alpha_j^*}$.

Slack variables

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

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



Linear soft margin SVM, primal form

Primal form of hard margin SVM

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

turns into

$$\text{minimize}_{\boldsymbol{\xi},\boldsymbol{w},b} \quad \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + \frac{C}{2}\sum_{i=1}^{\ell} \xi_i^2$$

$$\text{subject to} \quad y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i \ , \ i = 1, \dots, \ell$$

or

$$\text{minimize}_{\boldsymbol{\xi},\boldsymbol{w},b} \quad \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{\ell} \xi_i$$

$$\text{subject to} \quad y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i \ , \ i = 1, \dots, \ell$$

$$\xi_i \geq 0 \ , \ i = 1, \dots, \ell$$

2-Norm soft margin SVM

2-Norm Soft Margin SVM: For training data $S = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$ and kernel k the solution $\boldsymbol{\alpha}^*$, b^* of

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

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

leads to the decision rule $\operatorname{sign}(f(x))$ with $f(x) = \sum_{i=1}^{\ell} y_i \alpha_i^* k(x_i, x) + b^*$, where b^* is chosen so that $y_i f(x_i) = 1 - \alpha_i^*/C$ for any i with $\alpha_i \neq 0$ and the slack variables of the "corresponding hyperplane" in the feature space defined by kernel k are defined relative to the *geometric* margin $\rho = 1/\|\boldsymbol{w}^*\| = 1/\sqrt{\sum_{x_j \in \mathrm{SV}} \alpha_j^* - \frac{1}{C}\langle \boldsymbol{\alpha}^*, \boldsymbol{\alpha}^* \rangle}$.

1-norm soft margin SVM

1-norm soft margin SVM: For training data $S = \{x_1, y_1), \dots, (x_\ell, y_\ell)\}$ and kernel k the solution $\boldsymbol{\alpha}^*$, b^* of

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

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

leads to the decision rule $\mathrm{sign}(f(x))$ 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 the feature space defined by kernel k are defined relative to the *geometric* margin $\rho = 1/\|\boldsymbol{w}^*\| = 1/\sqrt{\sum_{x_j \in \mathrm{SV}} \alpha_j^*}$.

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 $L_{\text{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y}) \qquad (y \in \{-1, 1\}, \hat{y} \in \mathbb{R})$
- 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

- Loss $L_{\mathsf{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$
- 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

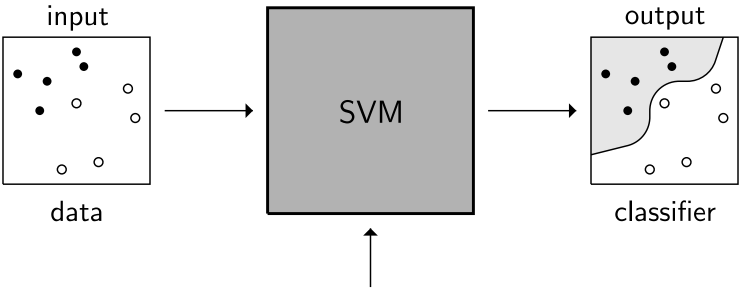
$$\begin{aligned} \text{minimize}_{\boldsymbol{\xi},\boldsymbol{w},b} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i \ , \ \ i = 1, \dots, \ell \\ & \xi_i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

corresponds to

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

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

Binary SVMs



prior knowledge:
kernel function
regularisation parameter

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