TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

Prof. Dr. Björn Sprungk
Faculty of Mathematics and Computer Science
Institute of Stochastics

Mathematics of machine learning

Chapter 4: Support Vector Machines and Kernel Methods

Winter term 2024/25

Contents

Chapter 4: Support Vector Machines and Kernel Methods

What it's about?

1. Get to know further important milestones of machine learning:

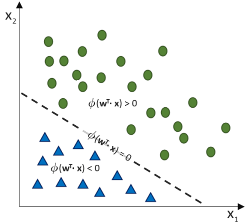
   - Support vector machines (SVM) (since 1970s)
   - Kernel methods (since 1990s)

   which in combination were the dominant supervised learning ansatz in the 1990s and early 2000s.

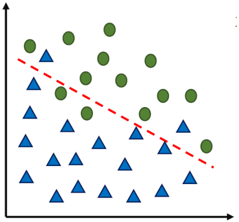
2. Understand the advantage of the SVM ansatz in comparison to other linear methods

3. Encounter the universal approximation theorem for kernels — our way to control ϵ_{app}

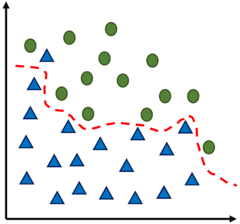
In the following three sections we discuss different approaches of support vector machines applicable in different situations:



Hard SVM
for linearly separable data



Soft SVM
for non-linearly separable data



Kernel SVM
for nonlinear decision boundaries

4.1 Hard SVM

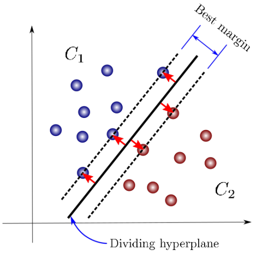
Introduction

- In the following two sections we will again consider linear hypotheses $h_{\mathbf{w},b} \in \mathcal{L}_d$ which separate $\mathcal{X} = \mathbb{R}^d$ into two halfspaces.

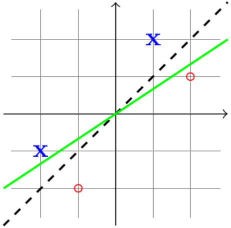
- **Objective:** Find a separating hyperplane that has largest possible distance (margin) to the data.

- **Motivation:**
  This hyperplane separates the data most "clearly" (black separates the data points more clearly than green).

- We again distinguish whether the sample is linearly separable or not – and start with the simpler case.



Source: towardsdatascience.com



Source: "Understanding Machine Learning" (2014)

The Margin

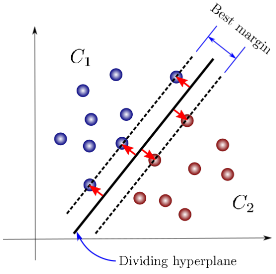
- The margin of a sample

$$s = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$$

to a hyperplane

$$H_{\mathbf{w},b} := \{\mathbf{x} \in \mathbb{R}^d \colon \mathbf{w} \cdot \mathbf{x} + b = 0\}$$

is the smallest distance of a point $\mathbf{x}_i \in \mathbb{R}^d$ to $H_{\mathbf{w},b}$.



Source: towardsdatascience.com

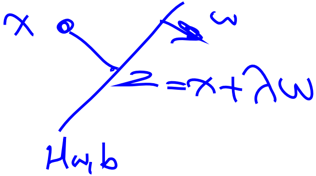
Proposition 4.1:

Let $\mathbf{w} \in \mathbb{R}^d$ be normalized, i. e. $\|\mathbf{w}\| = 1$. Then the margin of an $\mathbf{x} \in \mathbb{R}^d$ to the hyperplane $H_{\mathbf{w},b}$ is given by

$$d(\mathbf{x}, H_{\mathbf{w},b}) := \min_{\mathbf{z} \in H_{\mathbf{w},b}} \|\mathbf{x} - \mathbf{z}\| = |\mathbf{w} \cdot \mathbf{x} + b|.$$

Proof of proposition 4.7:

shortest distance between hyperplane $H_{w,b}$ and $x \in \mathbb{R}^d$ is along direction $w \in \mathbb{R}^d$ (normal dir)



Thus $d(x, H_{w,b}) = \min_{z \in H_{w,b}} \|x - z\|$

$$= \|x - (\underbrace{x + \lambda w}_{\text{orthogonal projection of } x \text{ onto } H_{w,b}})\|$$

$$= \|\lambda w\| = |\lambda| \times \underbrace{|w|}_{=1} = |\lambda|$$

it suffices to show that for

$$+b \stackrel{?}{=} 0 \to wx + \lambda \frac{w^2}{1} = \pm b \; (wx+b) + b \; \text{we} + \lambda = 0$$

$x + \lambda w \in H_{w,b}$:

$w(x + \lambda w)$.

largest distance to closest training points

= argmax = argmin margin = hard SVM

$\to$ or argmax of argmin of $R_S(h)$

Remark: For an arbitrary vector $\mathbf{w} \neq \mathbf{0}$ we have $H_{\mathbf{w},b} = H_{\mathbf{w}/\|\mathbf{w}\|, b/\|\mathbf{w}\|}$ and, hence,

$$d(\mathbf{x}, H_{\mathbf{w},b}) = |\mathbf{w} \cdot \mathbf{x} + b| \; / \; \|\mathbf{w}\|, \qquad \mathbf{x} \in \mathbb{R}^d.$$

Definition 4.2:

For a hyperplane $H_{\mathbf{w},b} \subset \mathbb{R}^d$, $\mathbf{w} \neq \mathbf{0}$, the margin to a sample $s = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ is defined as

$$\gamma_{\mathbf{w},b}(s) := \frac{1}{\|\mathbf{w}\|} \min_{i=1,\ldots,m} |\mathbf{w} \cdot \mathbf{x}_i + b|.$$

Goal

Among all hyperplanes $H_{\mathbf{w},b}$ separating a sample s find the one that has the largest margin $\gamma_{\mathbf{w},b}(s)$:

$$(\mathbf{w}_s, b_s) \in \underset{(\mathbf{w},b)\in\mathbb{R}^{d+1}}{\operatorname{argmax}} \; \gamma_{\mathbf{w},b}(s) \qquad \text{subject to:} \quad \underbrace{y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 0 \quad \forall i.}_{\text{then training data is seprated}}$$

$$\Rightarrow \max_{w,b} \left(\min_i \frac{|wx+b|}{\|w\|} \right)$$

The Hard SVM rule

- The problem

||w||^2 is smallest possible w, making int differentiable and convex which makes it easier to solve using quadratic optimization methods.

$$(\mathbf{w}_s, b_s) \in \underset{(\mathbf{w},b)\in\mathbb{R}^{d+1}}{\operatorname{argmax}} \; \gamma_{\mathbf{w},b}(s) \qquad \text{subject to:} \qquad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) > 0 \quad \forall i. \tag{$\star$}$$

has infinitely many solutions, since $\lambda(\mathbf{w}_s, b_s)$, $\lambda > 0$, produces the same hyperplane. One therefore often adds $\|\mathbf{w}\| = 1$ as a constraint.

- The optimization problem $(\star)$ can be conveniently solved by quadratic optimization:

Hard SVM rule

Given: nontrivial, linearly separable sample s with m pairs of data $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$.

Compute: $h_{\mathbf{w}_s,b_s} = \mathrm{SVM}_{\textsf{hard}}(s) \in \mathcal{L}_d$ given by

$$(\mathbf{w}_s, b_s) = \underset{(\mathbf{w},b)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \; \|\mathbf{w}\|^2 \qquad \text{subject to:} \qquad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \forall i.$$

Theorem 4.3:

If s is linearly separable and nontrivial, i.e., there exist i, j with $y_i \neq y_j$, then the hard SVM rule solves the optimization problem $(\star)$ and the largest possible margin is

$$\gamma^\star(s) = \gamma_{\mathbf{w}_s, b_s}(s) = \frac{1}{\|\mathbf{w}_s\|}.$$

considerig the $wx + b = 1$

since $wx + b \geq 1$

Remarks: goal = min of $\|w\|$ ← norm

- The hard SVM rule is a convex quadratic optimization task and has a unique solution $(\mathbf{w}_s, b_s)$ provided s is linearly separable and nontrivial.

- The (hard) SVM rule yields a particular minimizer of the empirical risk:

empirical risk

The smaller the w, the wider the margin.

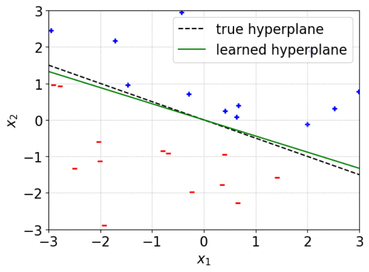
$$h_{\mathbf{w}_s, b_s} = \operatorname{argmax}\left\{\gamma_{\mathbf{w},b}(S) : h_{\mathbf{w},b} \in \underset{h \in \mathcal{L}_d}{\operatorname{argmin}}\, \mathcal{R}_s(h)\right\}$$

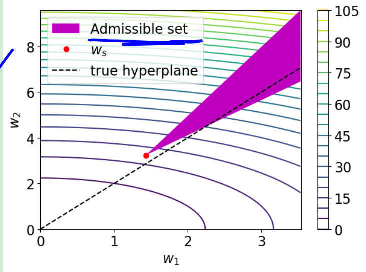
→ ERM

- It holds for the value B in Theorem 4.?? on the convergence of the perceptron algorithm that $B = \|\mathbf{w}_s\| = 1/\gamma^\star(s)$.

Example: Synthetic dataset

- For $\mathcal{X} = \mathbb{R}^2$ we want to learn an $h_{\mathbf{w},0} \in \mathcal{L}_d$ by the hard SVM rule.
- The $m = 25$ training data was generated using a true hypothesis from $\mathcal{L}_d$ with $\mathbf{w}^\dagger = (1,2)^\top$, $b^\dagger = 0$.
- The hard SVM rule yields $\mathbf{w}_s \approx (1.43, 3.24)$ with a margin $\gamma_{\mathbf{w}_S,0}(s) \approx 0.28$.
- The true separating hyperplane, on the other hand, has a margin of $\gamma_{\mathbf{w}^\dagger,0}(S) \approx 0.20$.
- This example can again be reproduced by a Jupyter notebook





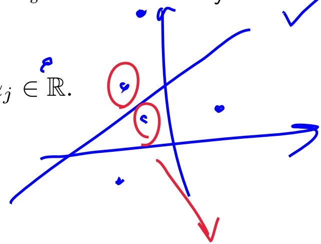
$\{(w,b) : y_i(wx_i + b) \geq 1 \ \forall i = 1, \dots, m\}$

Why is it called "support vector" machine?

it its larger than number of traing data then w_s

- The name *support vector machine* comes from the fact that the weight vector $\mathbf{w}_s \in \mathbb{R}^d$ learned by the hard SVM rule is composed of very special data points $\mathbf{x}_j \in \mathbb{R}^d$:

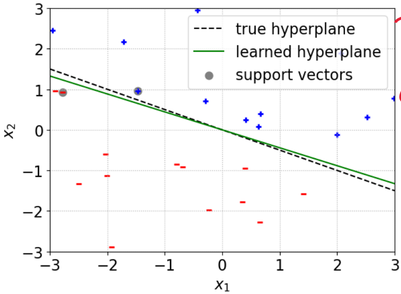
$$\mathbf{w}_s = \sum_{j \in J} \alpha_j \mathbf{x}_j, \qquad j \in J := \{i : y_i(\mathbf{w}_s \cdot \mathbf{x}_i + b_s) = 1\}, \quad \alpha_j \in \mathbb{R}.$$



- The vectors $\mathbf{x}_j$ with $y_j(\mathbf{w}_s \cdot \mathbf{x}_j + b_s) = 1$ are called support vectors of $\mathbf{w}_s$.

- The support vectors are exactly those data points $\mathbf{x}_j$ which have the smallest distance to the hyperplane $H_{\mathbf{w}_s, b_s}$:

$$y_i\left(\mathbf{w}_s \cdot \mathbf{x}_i + b_s\right) = 1$$
$$\Longleftrightarrow$$
$$d(\mathbf{x}_i, H_{\mathbf{w}_s, b_s}) = \gamma_{\mathbf{w}_s, b_s}(s)$$



they are closest margin so they are support vector

Mathematical background

Theorem 4.4: (Karush–Kuhn–Tucker conditions)

coefficiency

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be differentiable, $g_i(\mathbf{w}) = \mathbf{a}_i^\top \mathbf{w} + c_i$, $\mathbf{a}_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}$, for $i = 1, \ldots, m$ and consider

$$\mathbf{w}^* \in \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) \quad \text{subject to:} \quad g_i(\mathbf{w}) \le 0 \quad \forall i = 1, \ldots, m.$$

Then there exist coefficients $\alpha_i \ge 0$, $i = 1, \ldots, m$, such that

$$\nabla f(\mathbf{w}^*) + \sum_{i=1}^m \alpha_i \nabla g_i(\mathbf{w}^*) = \mathbf{0} \qquad \text{und} \quad \alpha_i g_i(\mathbf{w}^*) = 0 \qquad \forall i = 1, \ldots, m.$$

Consider now the SVM rule as a special case of the above optimization task:

$$f(\mathbf{w}, b) = \|\mathbf{w}\|^2, \qquad \Rightarrow \quad \nabla f(\mathbf{w}, b) = (2\mathbf{w}, 0)$$
$$g_i(\mathbf{w}, b) = 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \qquad \Rightarrow \quad \nabla g_i(\mathbf{w}, b) = -y_i(\mathbf{x}_i, 1)^\top.$$

then the KKT conditions yield $\quad \mathbf{w}_s = \sum_{j\in J} \alpha_j \mathbf{x}_j$

Advantage of the hard SVM rule

- The quantitative fundamental theorem as well as Theorem 4.?? yield for the 0-1 loss and under the realizability assumption that with probability of at least $1-\delta$

$$\mathcal{R}_\mu(\mathrm{SVM}_{\textsf{hard}}(S)) \leq \sqrt{\frac{c}{m}\left(d + \ln\left(\frac{1}{\delta}\right)\right)},$$

ok only for finite data

for a random sample S of size m

VCD of that is $d+1 \sim d$

- **Note:** The realizability assumption ensures the linear separability of S almost surely
- The same PAC condition applies to the output of the Perceptron algorithm.
- Notice, the bound grows with the feature dimension d.

PAC condition: $P\left(R_\mu(A(s)) \leq \inf_{h\in H} R_\mu(h) + \epsilon\right) \geq 1-\delta$

$m = m_H(\epsilon,\delta) \leq \frac{c}{\epsilon^2}\left(VCD(H) + \ln\left(\frac{1}{\delta}\right)\right) \rightarrow \epsilon \leq \sqrt{\frac{c}{m}\, VCD(H) + \ln\left(\frac{1}{\delta}\right)}$

since we determine ϵ such that for fixed $\delta \in (0,1)$

$\rightarrow P_{\mu^m}\left(R_\mu(A(s)) \leq \inf_{h\in H} R_\mu(h) + \epsilon\right) \geq 1-\delta$

$\rightarrow$ FTL: $P_{\mu^m}\left(R_\mu(A(s))\right) \leq \inf_{h\in H} R_\mu(h) + \underbrace{\sqrt{\frac{c}{m}\left(VCD(H) + \ln\left(\frac{1}{\delta}\right)\right)}}_{\text{tends to } \infty}$

$\Downarrow$

Fundamental theory

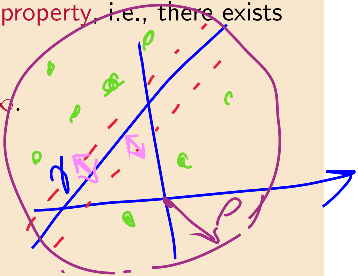
- Using refined techniques we can improve the bound for the hard SVM rule for particular μ:

no training data can be outside of circle

Theorem 4.5:

Let μ be a distribution on $\mathbb{R}^d \times \{-1,+1\}$ with the so-called (γ,ρ)-separability property, i.e., there exists $(\mathbf{w}^*, b^*) \in \mathbb{R}^{d+1}$ with $\|\mathbf{w}^*\| = 1$ and such that for $(\mathbf{X}, Y) \sim \mu$ almost surely

$$Y\left(\mathbf{w}^* \cdot \mathbf{X} + b^*\right) \geq \gamma > 0 \qquad \text{and} \qquad \|\mathbf{X}\| \leq \rho < \infty.$$



Then we have with probability at least $1-\delta$ that

it may hold for infinite purposes

$$\mathcal{R}_\mu(\mathrm{SVM}_{\mathsf{hard}}(S)) \leq \frac{1}{\sqrt{m}} \left(\frac{2\rho}{\gamma} + \sqrt{2 \ln\left(\frac{2}{\delta}\right)} \right).$$

Proof: See Chapter 26 in "Understanding Machine Learning" (2014)

- This yields, the error of the (hard) SVM rule is dimension independent for such distributions μ.

The larger the radius & smaller the margin, the larger the bound