

Învățare Automată (Machine Learning)



Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

Master Informatică, anul I, 2018-2019, cursul 13

Administrative

- final exam on 20th of June, 9 AM, amf. Titeica
- you can bring your laptop:
 - Internet connection allowed
 - cannot talk to your colleagues
- duration = 2 hours
- exercises similar to those presented in lectures, seminars, assignments
- final exam: 3 points + 1 point bonus = 4 points

Recap - Assignment 2

- modified version of the assignment from last week
- 3 problems = 3.5 points
- 1 bonus problem = 1 point
- ***deadline: Wednesday, 5th of June 2019, 23:59***
 - late submission policy: maximum 3 days allowed, -10% (= 0.35 points) for each day
 - submit hard copy
- OR
- send a pdf with a scan of your solution to bogdan.alexu@fmi.unibuc.ro
- for every problem write clear explanations, proofs to justify your answer (if you write just some indications you will not get too many points)
- do not share/copy the solution with/from your colleagues: you + your colleague/s will get 0 points

Recap - SVM

SVM Definition

- A **Support Vector Machine (SVM)** is a *non-probabilistic binary linear classifier*.
 - **Classifier** – A supervised learning method which predicts a categorical class.
 - **Linear** – The decision boundary is an n-dimensional hyperplane.
 - **Binary** – It can learn to predict one of two classes (a “+” class and a “-” class).
 - **Non-probabilistic** – The output of its *decision function* is not bounded, so it cannot be interpreted as probability.

There are methods of adapting an SVM to be used for **regression** problems and for making it **non-linear**, **multiclass** and/or **probabilistic**.

- An SVM tries to find a *separating* hyperplane, which is as far away from all training points at the same time (a **maximum-margin hyperplane**)
 - A point is classified as “+” or “-”, depending on which part of the hyperplane it lies.

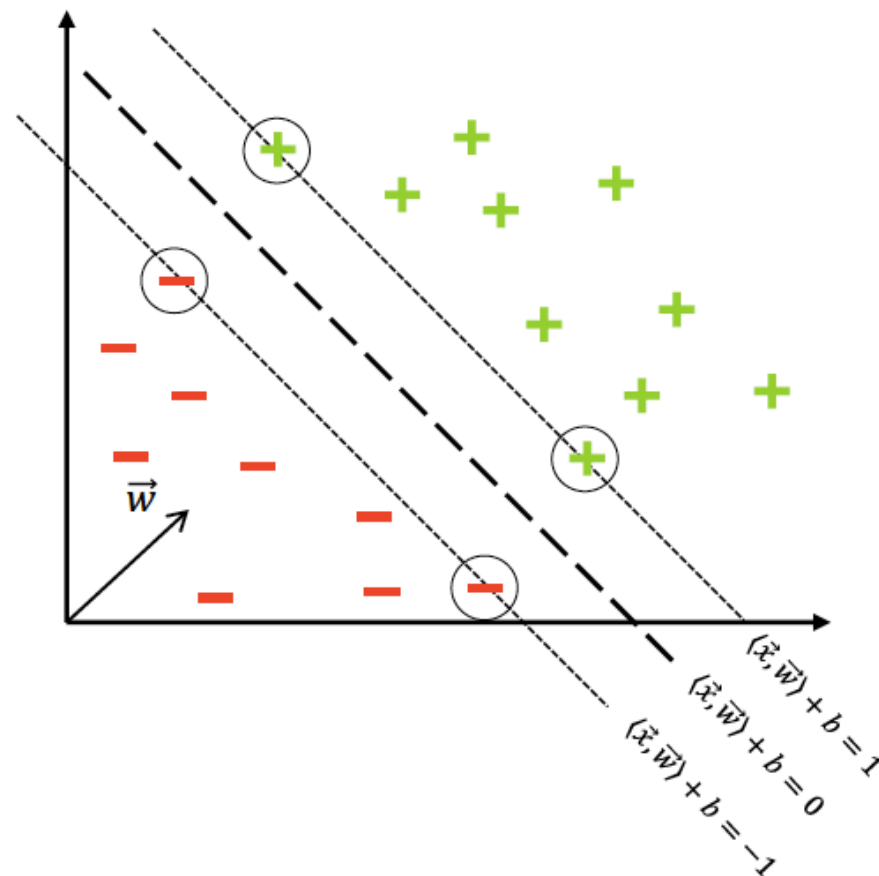
Recap - SVM

Learning a “good” separating hyperplane

- For training examples $(\vec{x}^{(i)}, y^{(i)})$, which lie exactly on the edges of the gap:

$$y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 = 0$$

- We call these examples “**Support Vectors**”



Recap - SVM

Making the margin “wide”

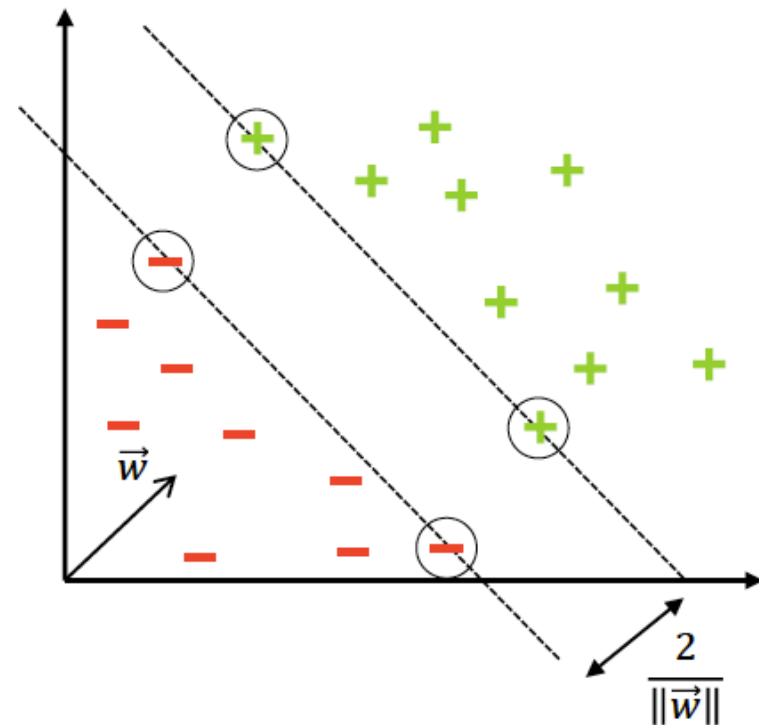
- How do we express the width of the gap?

$$g = \frac{2}{\|\vec{w}\|}$$

- We want to *maximize the gap*:

$$\text{maximize } \frac{2}{\|\vec{w}\|} \Rightarrow \text{minimize } \|\vec{w}\| \Rightarrow$$

$$\text{minimize } \frac{\|\vec{w}\|^2}{2}$$



Recap - SVM

What we have so far

SVM Primal Form

- The decision rule is:

\vec{x} is a “+” sample if $\langle \vec{x}, \vec{w} \rangle + b \geq 0$

- In order to obtain \vec{w} and b we need to:


$$\begin{aligned} & \text{minimize} \quad \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to} \quad y^{(i)} (\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$

Hard SVM

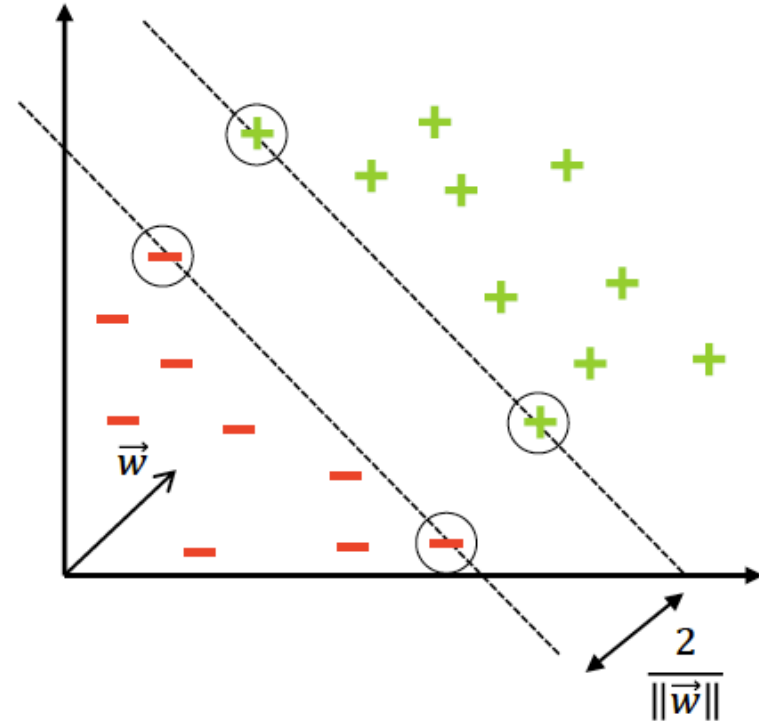
Hard SVM

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ with the largest possible margin.

$$\begin{aligned} & \text{minimize } \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to } y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$


 $\gamma = \frac{1}{\|w\|}$

$$\begin{aligned} & \text{minimize } \frac{1}{2\gamma^2} \\ & \text{subject to } y^{(i)} \left(\left\langle x^{(i)}, \frac{w}{\|w\|} \right\rangle + \frac{b}{\|w\|} \right) - \frac{1}{\|w\|} \geq 0 \end{aligned}$$



Hard SVM

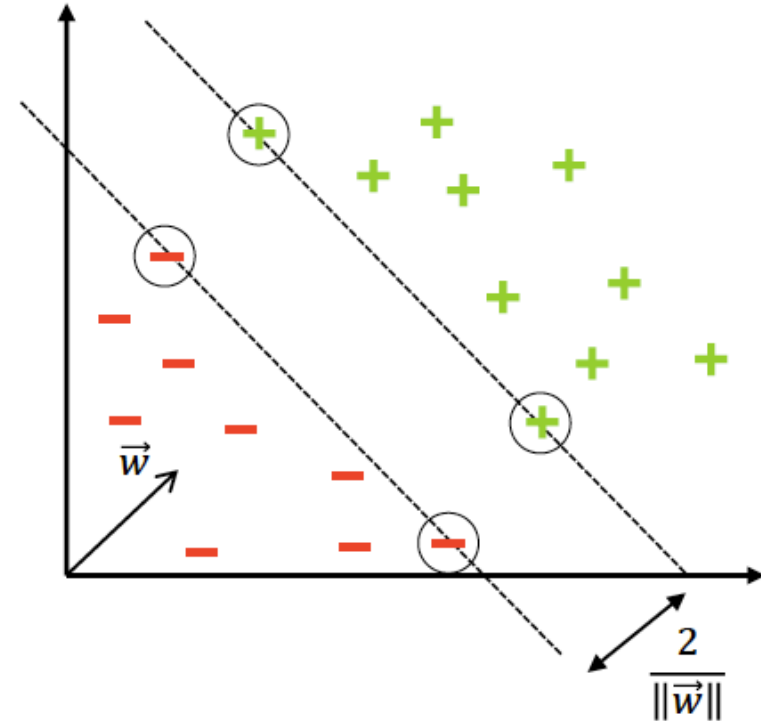
Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{aligned} & \text{minimize } \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to } y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$

$\Updownarrow \quad \gamma = \frac{1}{\|\vec{w}\|}$

$$\text{maximize } \gamma^2$$


$$\text{subject to } y^{(i)} \left(\left\langle x^{(i)}, \frac{w}{\|w\|} \right\rangle + \frac{b}{\|w\|} \right) \geq \gamma$$



Hard SVM

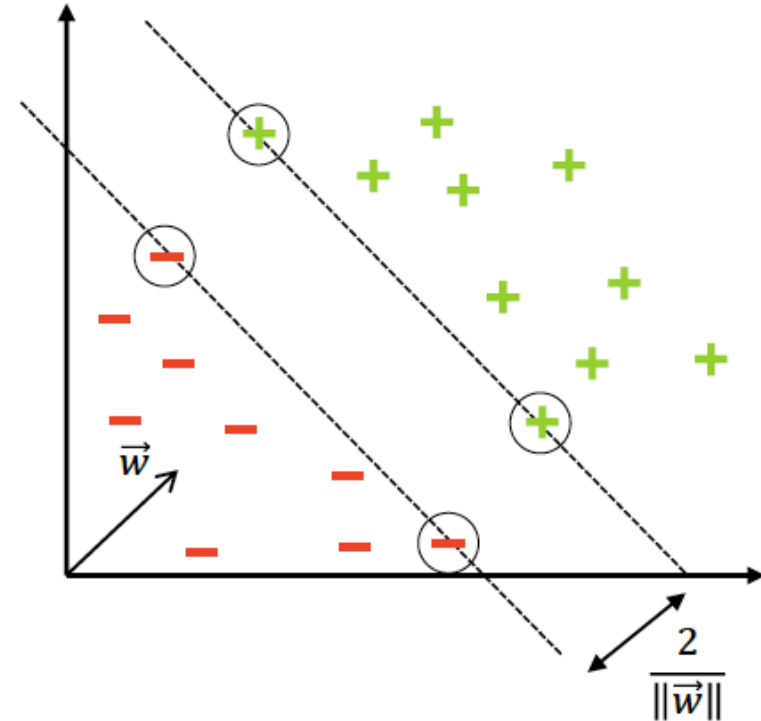
Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{aligned} & \text{minimize } \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to } y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$


 $\gamma = \frac{1}{\|\vec{w}\|}$

maximize γ

subject to $y^{(i)} \left(\left\langle x^{(i)}, \frac{w}{\|w\|} \right\rangle + \frac{b}{\|w\|} \right) \geq \gamma$



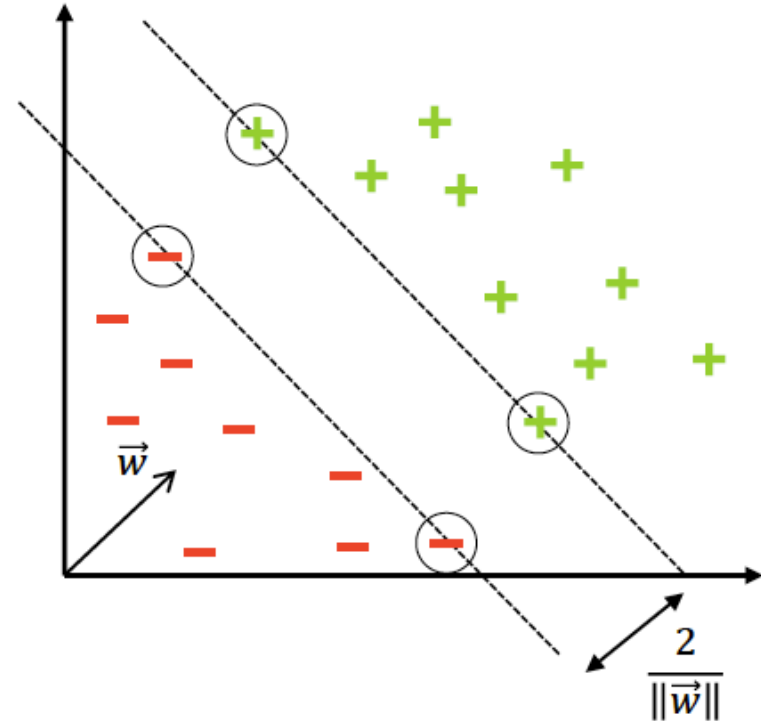
Hard SVM

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{aligned} & \text{minimize } \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to } y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$

$\Updownarrow \gamma = \frac{1}{\|\vec{w}\|}$

$$\operatorname{argmax}_{w,b} \min_{i=1,m} y^{(i)} \left(\left\langle x^{(i)}, \frac{w}{\|w\|} \right\rangle + \frac{b}{\|w\|} \right)$$



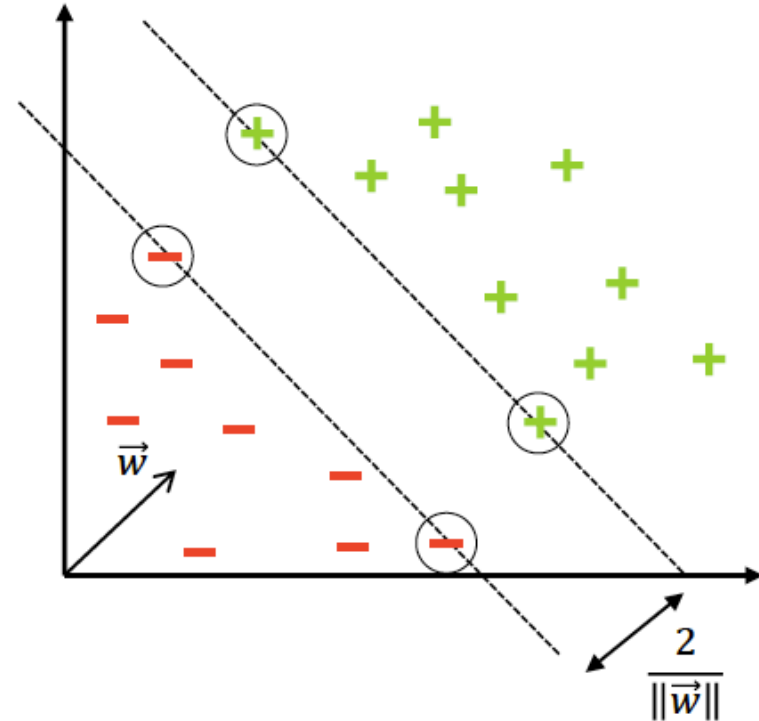
Hard SVM

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{aligned} & \text{minimize } \frac{\|\vec{w}\|^2}{2} \\ & \text{subject to } y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{aligned}$$

$\Updownarrow \gamma = \frac{1}{\|\vec{w}\|}$

$$\operatorname{argmax}_{w_0, b_0} \min_{i=1, m} y^{(i)} \left(\left\langle x^{(i)}, \frac{w_0}{\|w_0\|} \right\rangle + \frac{b_0}{\|w_0\|} \right)$$



Hard SVM

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{array}{ll} \text{minimize} & \frac{\|\vec{w}\|^2}{2} \\ \text{subject to} & y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{array}$$

$\gamma = \frac{1}{\|\vec{w}\|}$



Hard-SVM

input: $(x_1, y_1), \dots, (x_m, y_m)$

solve:

$$(w_0, b_0) = \underset{(w, b)}{\operatorname{argmin}} \|w\|^2 \text{ s.t. } \forall i, y_i(\langle w, x_i \rangle + b) \geq 1$$

output: $\hat{w} = \frac{w_0}{\|w_0\|}, \hat{b} = \frac{b_0}{\|w_0\|}$

$$\operatorname{argmax}_{w_0, b_0} \min_{i=1, m} y^{(i)} \left(\left\langle x^{(i)}, \frac{w_0}{\|w_0\|} \right\rangle + \frac{b_0}{\|w_0\|} \right)$$

Hard SVM

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set $S = \{(x_1, y_1), (x_1, y_1), \dots, (x_m, y_m)\}$ with the largest possible margin.

$$\begin{array}{ll} \text{minimize} & \frac{\|\vec{w}\|^2}{2} \\ \text{subject to} & y^{(i)}(\langle \vec{x}^{(i)}, \vec{w} \rangle + b) - 1 \geq 0 \end{array}$$

$\gamma = \frac{1}{\|w\|}$

Hard-SVM

input: $(x_1, y_1), \dots, (x_m, y_m)$

solve:

$$(w_0, b_0) = \underset{(w, b)}{\operatorname{argmin}} \|w\|^2 \text{ s.t. } \forall i, y_i(\langle w, x_i \rangle + b) \geq 1$$

output: $\hat{w} = \frac{w_0}{\|w_0\|}, \hat{b} = \frac{b_0}{\|w_0\|}$

$$\operatorname{argmax}_{\|w\|=1, b} \min_{i=1, m} y^{(i)}(\langle x^{(i)}, w \rangle + b)$$

The sample complexity of Hard SVM

The sample complexity of Hard SVM

Consider $\mathcal{H} = \mathcal{HS}^d$ be the set of halfspaces (linear classifiers) in \mathbf{R}^d

$$\mathcal{H} = \mathcal{HS}^d = \{h_{w,b}: \mathbf{R}^d \rightarrow \{-1, 1\}, h_{w,b}(x) = \text{sign}\left(\sum_{i=1}^d w_i x_i + b\right) \mid w \in \mathbf{R}^d, b \in \mathbf{R}\}$$

“Homogenous” linear classifiers: $b = 0$.

$$\mathcal{HS}_0^d = \{h_{w,0}: \mathbf{R}^d \rightarrow \{-1, 1\}, h_{w,0}(x) = \text{sign}\left(\sum_{i=1}^d w_i x_i\right) \mid w \in \mathbf{R}^d\}$$

$$\text{VCdim}(\mathcal{HS}_0^d) = d \text{ (proof in Lecture 6)}$$

$$\text{VCdim}(\mathcal{HS}^d) = d + 1 \text{ (proof in the book, easy to extend from the one given in the lecture)}$$

The sample complexity of Hard SVM

$\mathcal{H} = \mathcal{H}S^d$, $\text{VCdim}(\mathcal{H}) = d < \infty$. From the fundamental theorem of statistical learning we have that there are absolute constants C_1, C_2 such that:

\mathcal{H} is PAC learnable with sample complexity:

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

The sample complexity of learning halfspaces $m_{\mathcal{H}}(\epsilon, \delta)$ grows with the dimensionality d of the problem. Furthermore, the fundamental theorem of learning tells us that if the number of examples is significantly smaller than d/ϵ then no algorithm can learn an accurate halfspace. This is problematic when d is very large.

Text classification with bag-of-words

- classify a short text document according to its topic, say, whether the document is about sports or not.
- represent documents as vectors.
 - effective way is to use a bag-of-words representation.
 - define a dictionary of words and set the dimension d to be the number of words in the dictionary.
 - we represent a document as a vector $\mathbf{x} \in \{0,1\}^d$, where $x_i = 1$ if the i -th word in the dictionary appears in the document and $x_i = 0$ otherwise.
- a halfspace for the problem of text classification assigns weights to words
- common to have $d >$ number of training examples. In practice the problem is solvable, we can categorize text based on BOW.

Text Categorization with Support Vector Machines: Learning with Many Relevant Features

Thorsten Joachims

Universität Dortmund
Informatik LS8, Baroper Str. 301
44221 Dortmund, Germany

Abstract. This paper explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyzes the particular properties of learning with text data and identifies why SVMs are appropriate for this task. Empirical results support the theoretical findings. SVMs achieve substantial improvements over the currently best performing methods and behave robustly over a variety of different learning tasks. Furthermore, they are fully automatic, eliminating the need for manual parameter tuning.

https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

Sample complexity for separability case

- if the number of examples is significantly smaller than d/ε then no algorithm can learn an accurate halfspace.
 - this is problematic when d is very large.
- make an additional assumption on the underlying data distribution
 - define a “separability with margin γ ” assumption
 - if the data follows this assumption (is separable with margin γ) then the sample complexity is bounded from above by a function of $1/\gamma^2$.
 - even if the dimensionality d is very large (or even infinite), as long as the data adheres to the separability with margin assumption we can still have a small sample complexity.
 - *the sample complexity will not depend on d .*
 - there is no contradiction to the lower bound given in the fundamental theorem of learning because we are now making an additional assumption on the underlying data distribution.

Sample complexity for separability case

Definition 15.3. Let \mathcal{D} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that \mathcal{D} is separable with a (γ, ρ) -margin if there exists (\mathbf{w}^*, b^*) such that $\|\mathbf{w}^*\| = 1$ and such that with probability 1 over the choice of $(\mathbf{x}, y) \sim \mathcal{D}$ we have that $y(\langle \mathbf{w}^*, \mathbf{x} \rangle + b^*) \geq \gamma$ and $\|\mathbf{x}\| \leq \rho$. Similarly, we say that \mathcal{D} is separable with a (γ, ρ) -margin using a homogenous halfspace if the preceding holds with a halfspace of the form $(\mathbf{w}^*, 0)$.

$$\begin{aligned} & \text{minimize} \quad \frac{\|\bar{\mathbf{w}}\|^2}{2} \\ & \text{subject to} \quad y^{(i)}(\langle \bar{\mathbf{x}}^{(i)}, \bar{\mathbf{w}} \rangle + b) - 1 \geq 0 \end{aligned}$$

$$\gamma = \frac{1}{\|\mathbf{w}\|}$$



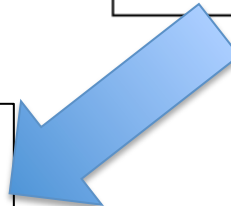
Hard-SVM

input: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$

solve:

$$(\mathbf{w}_0, b_0) = \underset{(\mathbf{w}, b)}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } \forall i, y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$$

output: $\hat{\mathbf{w}} = \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}, \hat{b} = \frac{b_0}{\|\mathbf{w}_0\|}$



$$\underset{\|\mathbf{w}\|=1, b}{\operatorname{argmax}} \min_{i=1, m} y^{(i)}(\langle \mathbf{x}^{(i)}, \mathbf{w} \rangle + b)$$

Sample complexity for separability case

Definition 15.3. Let \mathcal{D} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that \mathcal{D} is separable with a (γ, ρ) -margin if there exists (\mathbf{w}^*, b^*) such that $\|\mathbf{w}^*\| = 1$ and such that with probability 1 over the choice of $(\mathbf{x}, y) \sim \mathcal{D}$ we have that $y(\langle \mathbf{w}^*, \mathbf{x} \rangle + b^*) \geq \gamma$ and $\|\mathbf{x}\| \leq \rho$. Similarly, we say that \mathcal{D} is separable with a (γ, ρ) -margin using a homogenous halfspace if the preceding holds with a halfspace of the form $(\mathbf{w}^*, 0)$.

Theorem 15.4. Let \mathcal{D} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (γ, ρ) -separability with margin assumption using a homogenous halfspace. Then, with probability of at least $1 - \delta$ over the choice of a training set of size m , the 0-1 error of the output of Hard-SVM is at most

$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}.$$

Sample complexity for separability case

Theorem 15.4. *Let \mathcal{D} be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (γ, ρ) -separability with margin assumption using a homogenous halfspace. Then, with probability of at least $1 - \delta$ over the choice of a training set of size m , the 0-1 error of the output of Hard-SVM is at most*

$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}.$$

In practice, use bag-of-words with norm = 1, so $\rho = 1$.

The sample complexity does not depend on d .