Statistical Machine Learning (BE4M33SSU)
Lecture 3: Empirical Risk Minimization II

Czech Technical University in Prague
V.Franc

Linear classifier with minimal classification error

- $\mathcal{X}$ is a set of observations and $\mathcal{Y} = \{+1, -1\}$ a set of hidden labels
- $\phi\colon \mathcal{X} \to \mathbb{R}^n$ is fixed feature map embedding $\mathcal{X}$ to $\mathbb{R}^n$
- **Task:** find linear classification strategy $h\colon \mathcal{X} \to \mathcal{Y}$

$$h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b) = \begin{cases} +1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b \geq 0 \\ -1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b < 0 \end{cases}$$

with minimal expected risk

$$R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}\Big(\ell^{0/1}(y, h(x))\Big) \quad \text{where} \quad \ell^{0/1}(y, y') = [\![y \neq y']\!]$$

- We are given a set of training examples

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m\}$$

drawn from i.i.d. with the distribution $p(x, y)$.

ERM learning for linear classifiers

- The Empirical Risk Minimization principle leads to solving

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) \tag{1}$$

where the empirical risk is

$$R^{0/1}_{\mathcal{T}^m}(h(\cdot;\boldsymbol{w},b)) = \frac{1}{m}\sum_{i=1}^{m} [\![y^i \neq h(x^i;\boldsymbol{w},b)]\!]$$

In this lecture we address the following issues:

1. The statistical consitency of the ERM for hypothesis space containing linear classifiers.

2. Algorithmic issues: in general, there is no known algorithm solving the task (1) in time polynomial in m.

Vapnik-Chervonenkis (VC) dimension

Definition 1. *Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $\{x^1,\ldots,x^m\} \in \mathcal{X}^m$ be a set of m input observations. The set $\{x^1,\ldots,x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $\boldsymbol{y} \in \{+1,-1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i) = y^i$, $i \in \{1,\ldots,m\}$.*

Definition 2. *Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $\mathcal{X}$ which can be shattered by $\mathcal{H}$.*

Vapnik-Chervonenkis (VC) dimension

Definition 1. *Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $\{x^1,\ldots,x^m\} \in \mathcal{X}^m$ be a set of m input observations. The set $\{x^1,\ldots,x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $\boldsymbol{y} \in \{+1,-1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i) = y^i$, $i \in \{1,\ldots,m\}$.*

Definition 2. *Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $\mathcal{X}$ which can be shattered by $\mathcal{H}$.*

Theorem 1. *The VC-dimension of the hypothesis space of all linear classifiers operating in n-dimensional feature space $\mathcal{H} = \{h(x;\boldsymbol{w},b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w},b) \in (\mathbb{R}^n \times \mathbb{R})\}$ is $n+1$.*

Consistency of prediction with two classes and $0/1$-loss

Theorem 2. *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$ and $\mathcal{T}^m = \{(x^1,y^1),\ldots,(x^m,y^m)\} \in (\mathcal{X}\times\mathcal{Y})^m$ a training set draw from i.i.d. rand vars with distribution $p(x,y)$. Then, for any $\varepsilon > 0$ it holds*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h)\right| \geq \varepsilon\right) \leq 4\left(\frac{2\,e\,m}{d}\right)^d e^{-\frac{m\,\varepsilon^2}{8}}\,.$$

Consistency of prediction with two classes and $0/1$-loss

Theorem 2. *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$ and $\mathcal{T}^m = \{(x^1,y^1),\ldots,(x^m,y^m)\} \in (\mathcal{X}\times\mathcal{Y})^m$ a training set draw from i.i.d. rand vars with distribution $p(x,y)$. Then, for any $\varepsilon > 0$ it holds*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h)\right| \geq \varepsilon \right) \leq 4\left(\frac{2\,e\,m}{d}\right)^d e^{-\frac{m\,\varepsilon^2}{8}} .$$

Corollary 1. *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$. Then ERM is statistically consistent in $\mathcal{H}$ w.r.t $\ell^{0/1}$ loss function.*

Corollary 2. *Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis space with VC dimension $d < \infty$. Then, for any $0 < \delta < 1$ the inequality*

$$R^{0/1}(h) \leq R^{0/1}_{\mathcal{T}^m}(h) + \sqrt{\frac{8\big(d\log(2\,m)+1\big)+\log\frac{4}{\delta}\big)}{m}}$$

holds for any $h \in \mathcal{H}$ with probability $1-\delta$ at least.

Training linear classifier from separable examples

Definition 3. *The examples* $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ *are linearly separable w.r.t. feature map* $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ *if there exists* $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ *such that*

$$y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) > 0\,, \qquad i \in \{1, \ldots, m\} \tag{2}$$

Perceptron algorithm:

Input: linearly separable examples $\mathcal{T}^m$

Output: linear classifier with $R^{0/1}_{\mathcal{T}^m}(h(\cdot; \boldsymbol{w}, b)) = 0$

step 1: $\boldsymbol{w} \leftarrow \mathbf{0}$, $b \leftarrow 0$

step 2: find $(\boldsymbol{x}^i, y^i)$ such that $y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) \le 0$.
If not found exit, the current $(\boldsymbol{w}, b)$ solves the problem.

step 3: $\boldsymbol{w} \leftarrow \boldsymbol{w} + y^i\, \boldsymbol{\phi}(x^i)$, $b \leftarrow b + y^i$ and goto to step 2.

Training linear classifier from NON-separable examples

- The intractable ERM problem we wish to solve

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} \frac{1}{m}\sum_{i=1}^{m} \underbrace{[\![y^i \neq h(x^i;\boldsymbol{w},b))]\!]}_{\ell^{0/1}(y^i,h(x^i;\boldsymbol{w},b))}$$

where $h(x;\boldsymbol{w},b) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b)$.

Training linear classifier from NON-separable examples

- The intractable ERM problem we wish to solve

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} \frac{1}{m}\sum_{i=1}^{m} \underbrace{[\![y^i \neq h(x^i;\boldsymbol{w},b)]\!]}_{\ell^{0/1}(y^i,h(x^i;\boldsymbol{w},b))}$$

where $h(x;\boldsymbol{w},b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b)$.

- The ERM problem is approximated by a tractable **convex problem**

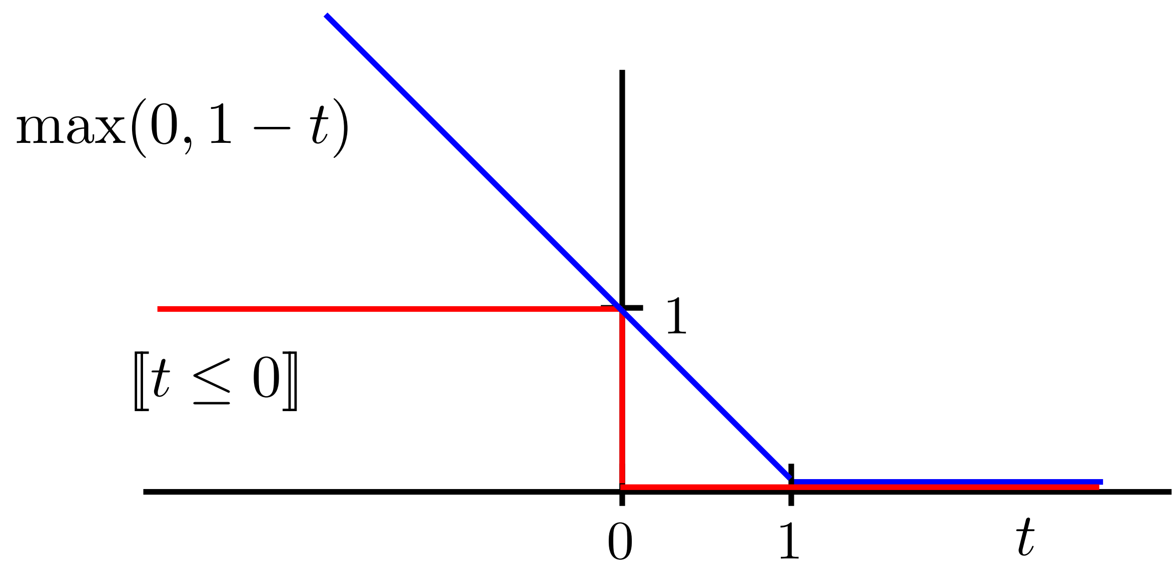
$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} \frac{1}{m}\sum_{i=1}^{m} \underbrace{\max\{0, 1 - y^i f(x^i;\boldsymbol{w},b)\}}_{\psi(y^i,f(x^i;\boldsymbol{w},b))}$$

where $f(x;\boldsymbol{w},b) = \langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b$ and $\psi(y, f(x))$ is so called Hinge-loss.

The hinge-loss upper bounds the 0/1-loss

- The hinge-loss is an upper bound of the $0/1$-loss evaluated for the predictor $h(x) = \mathrm{sign}(f(x))$:

$$\underbrace{[\![\mathrm{sign}(f(x)) \neq y]\!]}_{\ell^{0/1}(y, f(x))} = [\![\, y\, f(x) \leq 0]\!] \leq \underbrace{\max\{0, 1 - y\ f(x)\}}_{\psi(y, f(x))}$$



Support Vector Machines

- Find linear classifier $h(x; \boldsymbol{w}, b) = \mathrm{sign}(\langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b)$ by solving

$$(\boldsymbol{w}^*, b^*) = \mathop{\mathrm{argmin}}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \left(\underbrace{\frac{\lambda}{2} \|\boldsymbol{w}\|^2}_{\substack{\text{penalty} \\ \text{term}}} + \underbrace{\frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i) \rangle + b)\}}_{\text{empirical error}} \right)$$

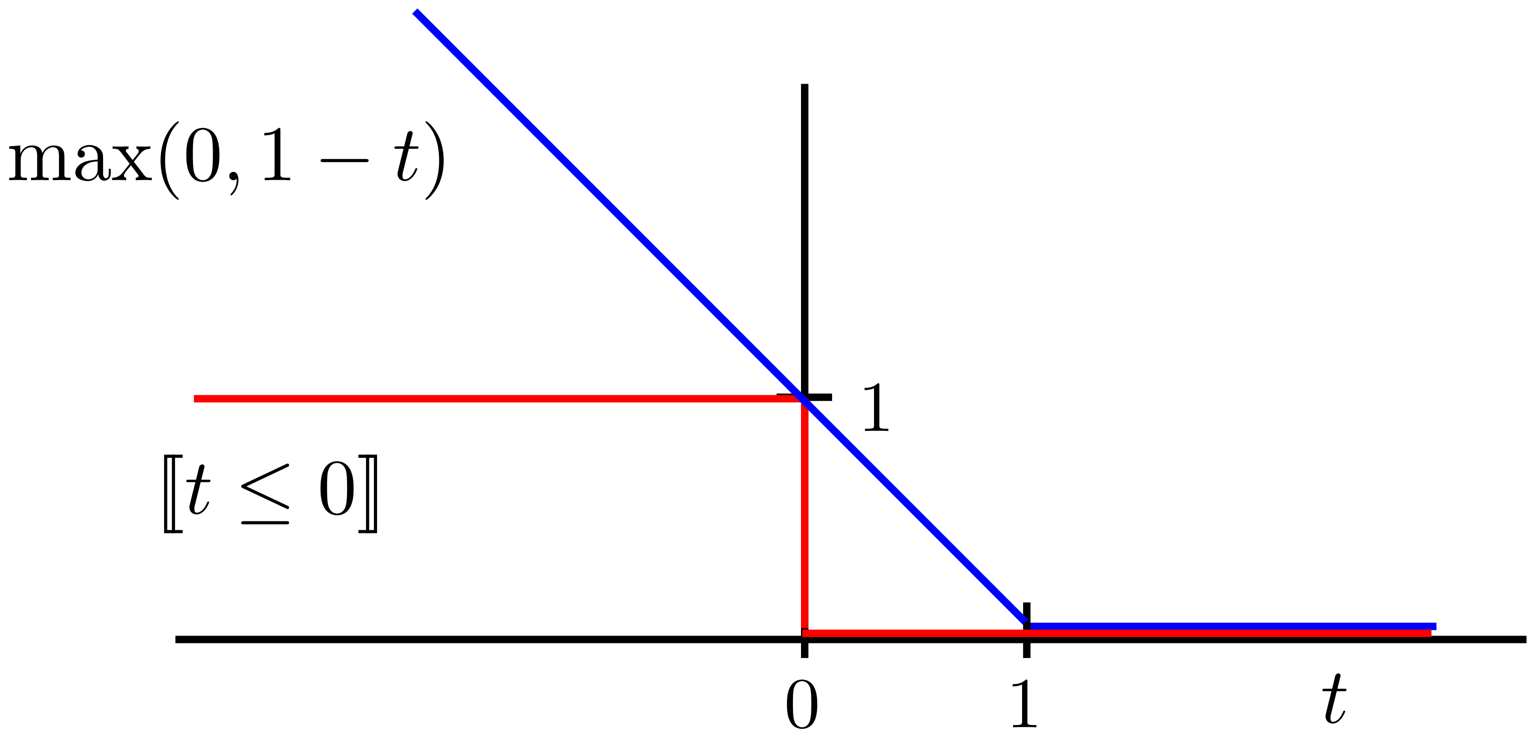
- The regularization constant $\lambda \geq 0$ helps to prevent overfitting (i.e. high estimation error) by constraining the parameter space.
  - $\lambda_1 > \lambda_2$ implies $\|\boldsymbol{w}_1^*\| \leq \|\boldsymbol{w}_2^*\|$

- Small $\|\boldsymbol{w}\|$ implies score $f(x; \boldsymbol{w}, b) = \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b$ varies slowly.
  - Cauchy inequality:
    $(\langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle - \langle \boldsymbol{\phi}(x'), \boldsymbol{w} \rangle)^2 \leq \|\boldsymbol{\phi}(x) - \boldsymbol{\phi}(x')\|^2 \|\boldsymbol{w}\|^2$

Summary

Topics covered in the lecture

- Linear classifier
- Vapnik-Chervonenkis dimension
- Consistency + generalization bound for two-class prediction and $0/1$-loss
- ERM problem for linear classifiers
- Perceptron for separable examples
- SVM for non-seperable examples


$\max(0, 1 - t)$
$[\![t \leq 0]\!]$
1
0
1
t