Statistical Machine Learning (BE4M33SSU)
Lecture 3: Support Vector Machines I

Czech Technical University in Prague

Linear classifier with minimal classification error

- $\mathcal{X}$ is a set of observations and $\mathcal{Y} = \{+1, -1\}$ is a set of hidden labels
- $\phi\colon \mathcal{X} \to \mathbb{R}^n$ is fixed feature map embedding observations from $\mathcal{X}$ to $\mathbb{R}^n$
- Task: we search for a 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

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

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

ERM for hypothesis space containing 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

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

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

- Correctness: is the ERM algorithm using the hypothesis space $\mathcal{H} = \{h(x;\boldsymbol{w},b) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w},b) \in (\mathbb{R}^n \times \mathbb{R})\}$ statistically consistent? ... Yes.

Generalization bound for prediction with two classes and 0/1-loss

Theorem 1. *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. random variables with distribution $p(x,y)$. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ the inequality*

$$R^{0/1}(h) \leq R^{0/1}_{\mathcal{T}^m}(h) + \sqrt{\frac{D(\log \frac{2m}{D} + 1) + \log \frac{1}{\delta}}{m}}$$

holds for any $h \in \mathcal{H}$.

- Unlike the finite hypothesis case the cardinality of $\mathcal{H}$ is replaced by the VC-dimension of $\mathcal{H}$ define even if $|\mathcal{H}|$ is infinite.
- As in the finite case, the bound holds for any $p(x,y)$ and the confidence interval can be decreased either by increasing m or decreasing D.

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

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* $\phi\colon \mathcal{X} \to \mathbb{R}^n$ *if there exists* $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ *such that*

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

- Implementation of the ERM for linearly separable examples $\mathcal{T}^m$ leads to solving (2) which provides a classifier $h(x; \boldsymbol{w}, b)$ with zero empirical risk $R^{0/1}_{\mathcal{T}^m}(h(\cdot; \boldsymbol{w}, b)) = 0$.

- Note that $y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) > 0$ implies

$$h(x^i) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) = y^i$$

- The task (2) can be dealt with by linear programming solvers or special solvers like the Perceptron algorithm.

Maximum margin classifier

Definition 4. *Given linearly separable examples $\mathcal{T}^m$, the maximum margin classifier is a linear classifier $h(\cdot; \boldsymbol{w}^*, b^*)$ with parameters*

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmax}_{\substack{\boldsymbol{w}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\} \\ b\in\mathbb{R}}} \gamma(\boldsymbol{w}, b) \qquad (3)$$

where the margin is defined as

$$\gamma(\boldsymbol{w}, b) = \min_{i\in\{1,\dots,m\}} \frac{y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b)}{\|\boldsymbol{w}\|}$$

- The problem (3) is equivalent to a convex quadratic program

$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{(\boldsymbol{w},b)\in\mathbb{R}^{n+1}} \frac{1}{2}\|\boldsymbol{w}\|^2$$

  subject to

$$y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) \geq 1\,, \qquad i \in \{1,\dots,m\}$$

Definition 5. *Given (possibly non-separable) examples $\mathcal{T}^m$, the parameters of the linear SVM classifier are obtained as the solution of a convex QP*

$$(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) = \mathop{\mathrm{argmin}}_{\substack{(\boldsymbol{w},b)\in\mathbb{R}^{n+1} \\ \boldsymbol{\xi}\in\mathbb{R}^m}} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \right)$$

subject to

$$\begin{aligned} y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) &\geq 1 - \xi_i\,, \quad i \in \{1,\dots,m\} \\ \xi_i &\geq 0\,, \qquad\quad i \in \{1,\dots,m\} \end{aligned}$$

- The (regularization) constant $\lambda > 0$ is a hyper-parameter controlling the trade-off between the quadratic term $\frac{1}{2}\|\boldsymbol{w}\|^2$ and the sum of slack variables.

Equivalent formulations of linear SVM

- The linear SVM is equivalent to an unconstrained convex problem

$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{(\boldsymbol{w},b)\in\mathbb{R}^{n+1}} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b)\} \right)$$

  following from the observation that for given $(\boldsymbol{w}, b)$ the optimal value of the slack variable is $\xi^i(\boldsymbol{w}, b) = \max\{0, 1 - y^i(\langle \boldsymbol{x}^i, \boldsymbol{w}\rangle + b\}$

- The linear SVM problem is further equivalent to

$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{\|\boldsymbol{w}\|\leq R, b\in\mathbb{R}} \left(\frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b)\} \right)$$

  where $R = r(\lambda)$ and $r\colon \mathbb{R} \to \mathbb{R}$ is a non-increasing function of λ.

Linear SVM implements ERM of an auxiliary problem

- $\mathcal{X}$, $\mathcal{Y} = \{+1, -1\}$ and $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ defined as before.
- The goal of the auxiliary problem is to find a decision function $f\colon \mathcal{X} \to \mathbb{R}$ minimizing the expectation of the hinge loss:

$$R^{\psi}(f) = \mathbb{E}_{(x,y)\sim p}(\psi(y, f(x))) \quad \text{where} \quad \psi(y,t) = \max\{0, 1 - y\, t\}$$

- Assuming the hypothesis space which contains the linear functions

$$\mathcal{F}_R = \left\{ f(x) = \langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b \mid (\boldsymbol{w}, b) \in \mathbb{R}^{n+1}, \|\boldsymbol{w}\| \leq R \right\}$$

the ERM principle leads to solving

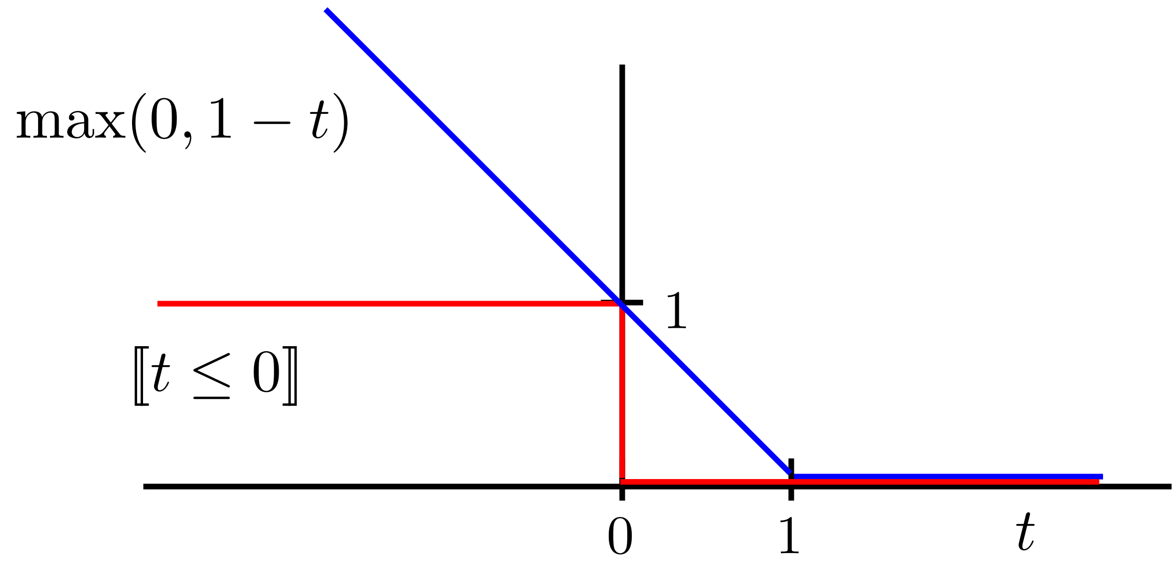
$$f^* = \operatorname*{Argmin}_{f \in \mathcal{F}_R} R^{\psi}_{\mathcal{T}^m}(f) \quad \text{where} \quad R^{\psi}_{\mathcal{T}^m}(f) = \frac{1}{m} \sum_{i=1}^{m} \psi(y^i, f(x^i))$$

which is exactly the task solved by SVM algorithm.

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 $h(x) = \text{sign}(f(x))$:

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



- Therefore 0/1-risk of $h(x) = \text{sign}(f(x))$ is upper-bounded by ψ-risk:

$$R^{0/1}(\text{sign}(f)) \leq R^{\psi}(f) \qquad \text{for any} \qquad f \colon \mathcal{X} \to \mathbb{R}$$

Upper bound on the excess error

- The best attainable 0/1-risk is $R^{0/1}_* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R^{0/1}(h)$.
- The best attainable ψ-risk is $R^\psi_* = \inf_{f \in \mathbb{R}^\mathcal{X}} R^\psi(f)$

Theorem 3. *The inequality*

$$\underbrace{R^{0/1}(\mathrm{sign}(f)) - R^{0/1}_*}_{\substack{\text{excess error}\\ \text{of original task}}} \leq \underbrace{R^\psi(f) - R^\psi_*}_{\substack{\text{excess error}\\ \text{of auxiliary task}}}$$

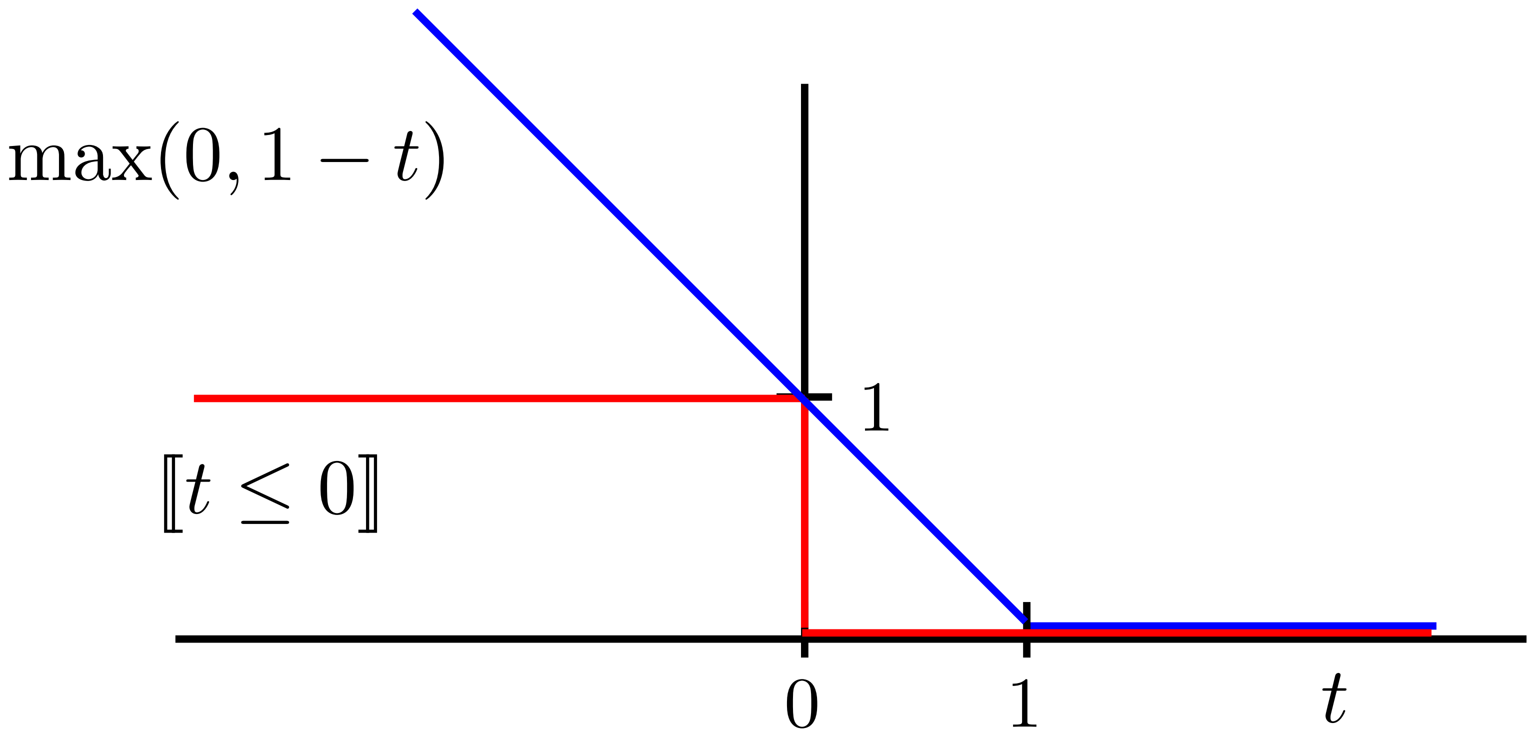
holds for all $f \colon \mathcal{X} \to \mathbb{R}$

Corollary 1. *Let* $\mathcal{F} \subseteq \{f \colon \mathcal{X} \to \mathbb{R}\}$ *be such that the approximation error of the auxiliary task is zero, that is,* $\inf_{f \in \mathcal{F}} R^\psi(f) = R^\psi_*$. *Then any minimizer of the* ψ*-risk* $R^\psi(f)$ *is a minimizer of the 0/1-risk* $R^{0/1}(\mathrm{sign}(f))$.

Summary

Topics covered in the lecture

- Generalization bound for two-class classifiers and $0/1$-loss
- Vapnik-Chervonenkis dimension for linear classifier
- Linear Support Vector Machines
- SVMs implement ERM for an auxiliary problem
- Excess error of ψ-risk upper bounds the excess error of $0/1$-risk


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