Statistical Machine Learning (BE4M33SSU)
Lecture 4: Support Vector Machines

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

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

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.

Training linear classifier from separable examples

Definition 1. *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 $y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) \leq 0$.
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 $h(x;\boldsymbol{w},b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b)$.

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

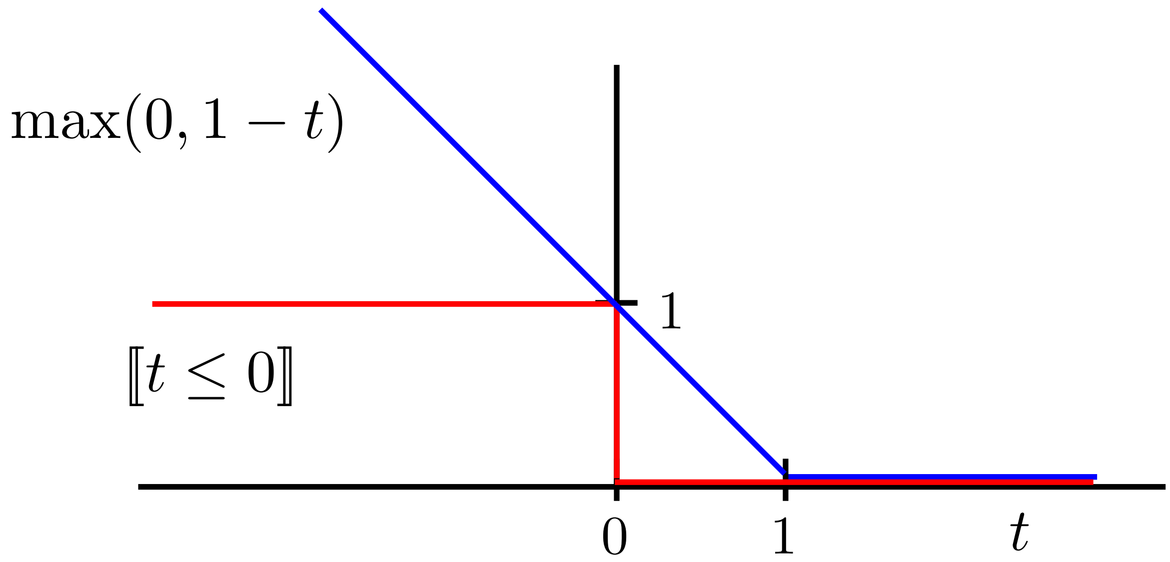
$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n\times\mathbb{R})} \frac{1}{m}\sum_{i=1}^{m} \underbrace{\max\{0, 1 - y^i f(x^i;\boldsymbol{w},b)\}}_{\psi(y^i,f(x^i;\boldsymbol{w},b))}$$

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

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



Support Vector Machines

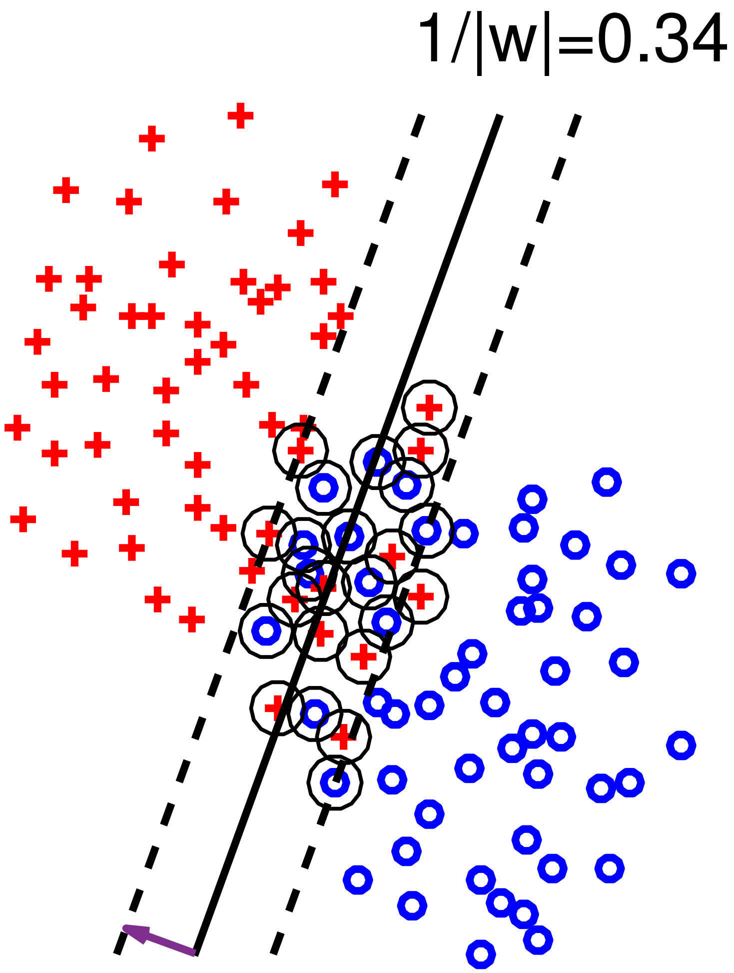
- Find linear classifier $h(x; \boldsymbol{w}, b) = \mathrm{sign}(\langle \boldsymbol{\phi}(x), \boldsymbol{w}\rangle + b)$ by solving

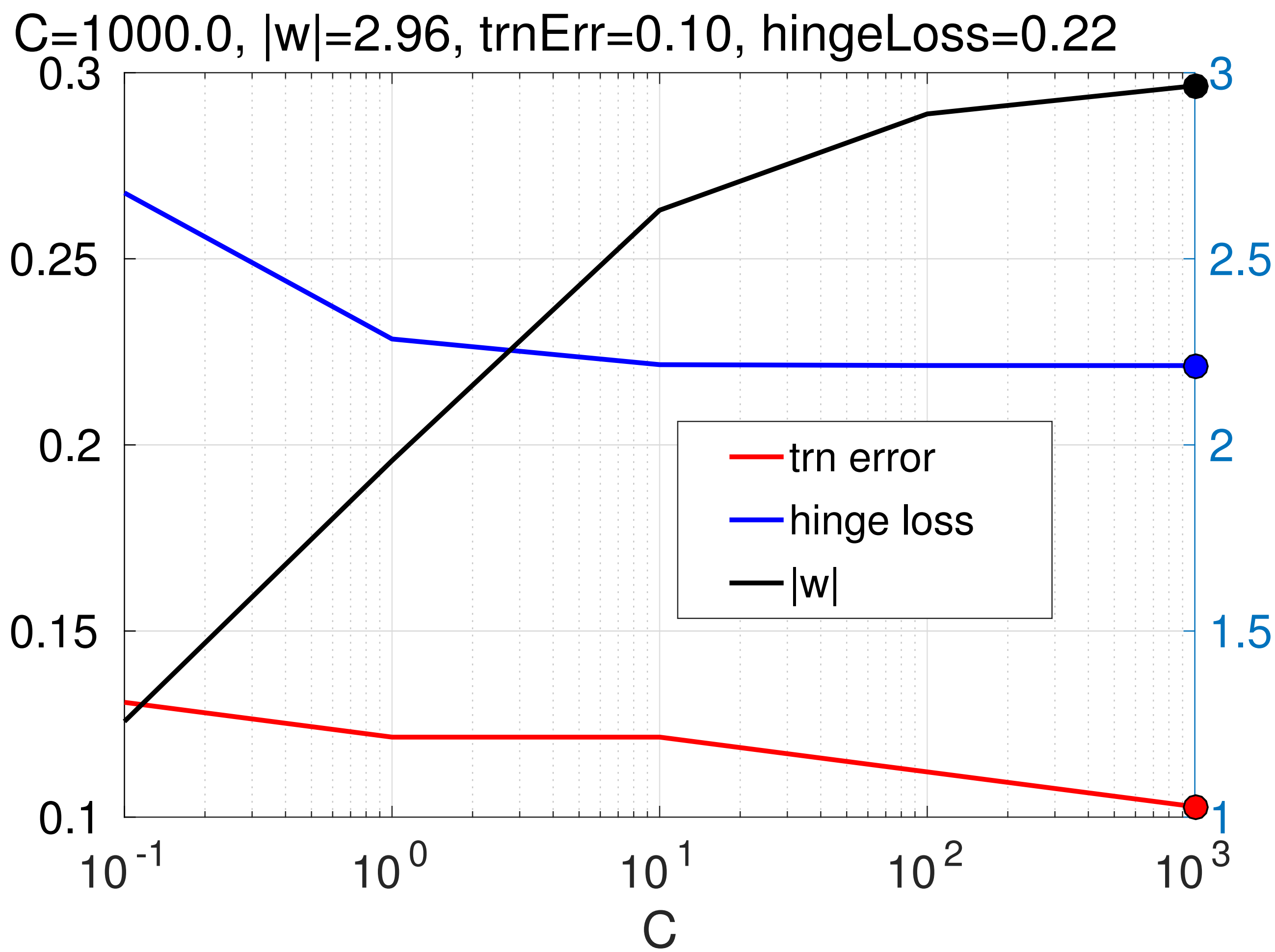
$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}} \Bigg(\underbrace{\frac{1}{2}\|\boldsymbol{w}\|^2}_{\text{penalty term}} + \underbrace{C\sum_{i=1}^m \max\{0, 1 - y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b)\}}_{\text{empirical error}}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

- The regularization constant $C \geq 0$ helps to prevent overfitting (i.e. high estimation error) by constraining the parameter space.
  - $C_1 < C_2$ 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:
    $(\langle \boldsymbol{\phi}(x), \boldsymbol{w}\rangle - \langle \boldsymbol{\phi}(x'), \boldsymbol{w}\rangle)^2 \leq \|\boldsymbol{\phi}(x) - \boldsymbol{\phi}(x')\|^2 \|\boldsymbol{w}\|^2$

Example: Primal SVM problem

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





SVM as Quadratic Program

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

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

where $C > 0$ is the regularization constant.

- It can be re-formulated as a convex *quadratic program*

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

subject to

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

From Primal SVM to Dual SVM problem

- Lagrangian of the primal SVM problem:

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \underbrace{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i}_{\text{original objective}}$$

$$\underbrace{- \sum_{i=1}^{m}\alpha_i(y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) - 1 + \xi_i) - \sum_{i=1}^{m}\mu_i\xi_i}_{\text{constraint violation penalty}}$$

- Strong duality:

$$\underbrace{\min_{\substack{\boldsymbol{w}\in\mathbb{R}^n\\ b\in\mathbb{R}\\ \boldsymbol{\xi}\in\mathbb{R}^m}} \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^m_+\\ \boldsymbol{\mu}\in\mathbb{R}^m_+}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}_{\text{primal problem}} = \underbrace{\max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^m_+\\ \boldsymbol{\mu}\in\mathbb{R}^m_+}} \min_{\substack{\boldsymbol{w}\in\mathbb{R}^n\\ b\in\mathbb{R}\\ \boldsymbol{\xi}\in\mathbb{R}^m}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}_{\text{dual problem}}$$

Dual SVM problem

- The dual SVM formulation is a convex quadratic program

$$\boldsymbol{\alpha}^* = \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^m} \Bigg(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^i y^j \langle \boldsymbol{\phi}(x^i), \boldsymbol{\phi}(x^j) \rangle \Bigg)$$

$$\text{s.t.} \qquad \sum_{i=1}^m \alpha_i \, y^i = 0 \,, \quad 0 \le \alpha_i \le C \,, \quad i \in \{1, \dots, m\}$$

- The primal variables $(\boldsymbol{w}, b)$ are obtained from the dual variables $\boldsymbol{\alpha}$ by

$$\boldsymbol{w} \;=\; \sum_{i=1}^m y^i \, \boldsymbol{\phi}(x^i) \, \alpha_i = \sum_{i \in \mathcal{I}_{\mathrm{SV}}} y^i \, \boldsymbol{\phi}(x^i) \, \alpha_i$$

$$b \;=\; y^i - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i) \rangle \,, \; \forall i \in \mathcal{I}_{\mathrm{sv}}^{\mathrm{b}} = \{ j \mid 0 < \alpha_j < C \}$$

- $\boldsymbol{\alpha}$ is sparse; $\boldsymbol{w}$ is lin. combination of Support Vectors $\mathcal{I}_{\mathrm{sv}} = \{ j \mid \alpha_j > 0 \}$

Example: SVM classifier

$$f(x) = \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b = \langle \underbrace{\sum_{i=1}^{m} y^i \alpha_i \, \boldsymbol{\phi}(x^i)}_{\boldsymbol{w}}, \boldsymbol{\phi}(x) \rangle + b$$