Statistical Machine Learning (BE4M33SSU)
Lecture 4: Empirical Risk Minimization II

Czech Technical University in Prague
V. Franc

Linear classifier minimizing 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 $h\colon \mathcal{X} \to \mathcal{Y}$

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

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

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

- Algorithmic issues (next lecture): in general, there is no known algorithm solving the task (1) in time polynomial in m.
- The uniform bound on the generalization error (this lecture):

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq B(m,\mathcal{H},\varepsilon)$$

Vapnik-Chervonenkis (VC) dimension

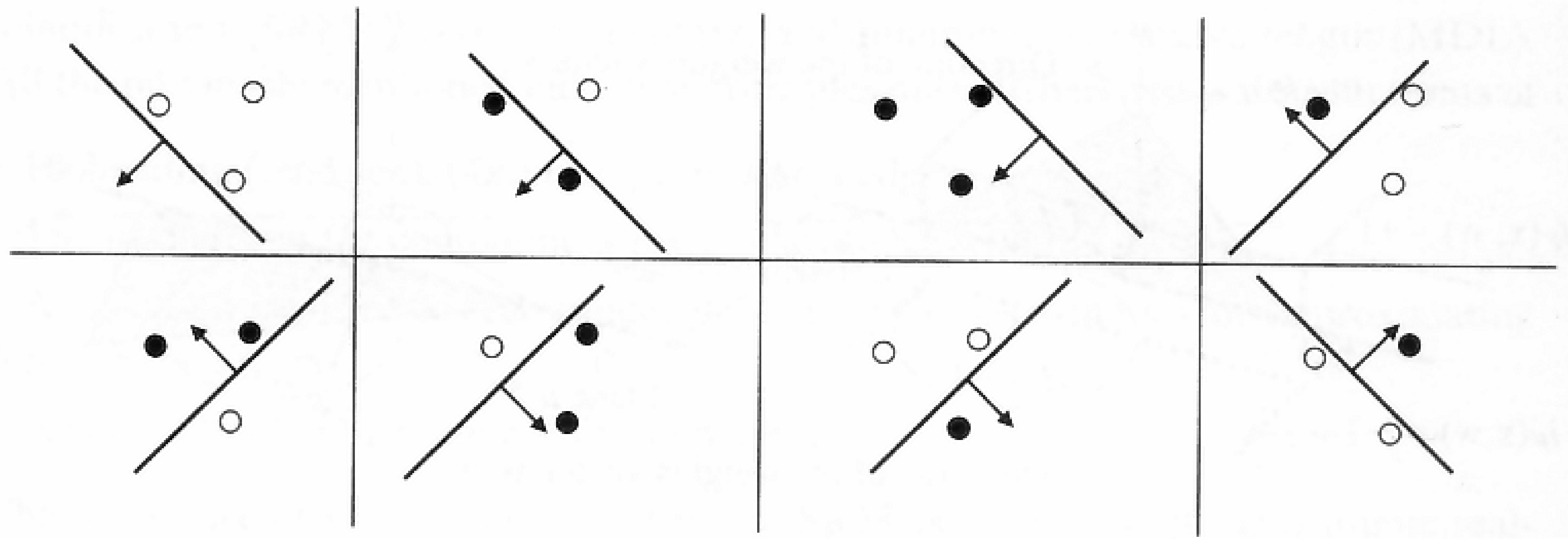
Definition: Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $\{x^1,\ldots,x^m\} \in \mathcal{X}^m$ be a set of m input observations. The set $\{x^1,\ldots,x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $\boldsymbol{y} \in \{+1,-1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i)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

Definition: 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 $\mathcal{H}$.

VC dimension of class of two-class linear classifiers

Theorem: The VC-dimension of the hypothesis class of all two-class 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 $n+1$.

Example for $n=2$-dimensional feature class



ULLN for two class predictors and $0/1$-loss

Theorem: Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis class 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

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

Corollary: Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$. Then ULLN applies.

Summary: uniform bounds on the generalization error

- We learned how to bound the generalization error uniformly for:
  - Finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$:

$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \leq 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

  - Two-class classifiers $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ a finite VC-dimensions d:

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

  In both cases the bound goes to zero, i.e., ULLN applies.

- Does ERM algorithm $h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$ finds strategy with the minimal risk $R(h)$?

Statistically consistent learning algorithm

- $h_{\mathcal{H}} \in \mathrm{Argmin}_{h \in \mathcal{H}} R(h)$ the best strategy in $\mathcal{H}$ has the risk $R(h_{\mathcal{H}})$
- $h_m = A(\mathcal{T}_m)$ strategy learned from $\mathcal{T}_m$ with has risk $R(h_m)$
- $R(h_m) - R(h_{\mathcal{H}})$ is the estimation error
- The statistically consistent algorithm can make the estimation error arbitrarily small if it has enough examples.

Definition: The algorithm $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ is statistically consistent in $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if for any $p(x, y)$ it holds that

$$\forall\, \varepsilon > 0\colon \lim_{m \to \infty} \mathbb{P}\Bigg(\underbrace{R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon}_{\text{high estimation error}} \Bigg) = 0$$

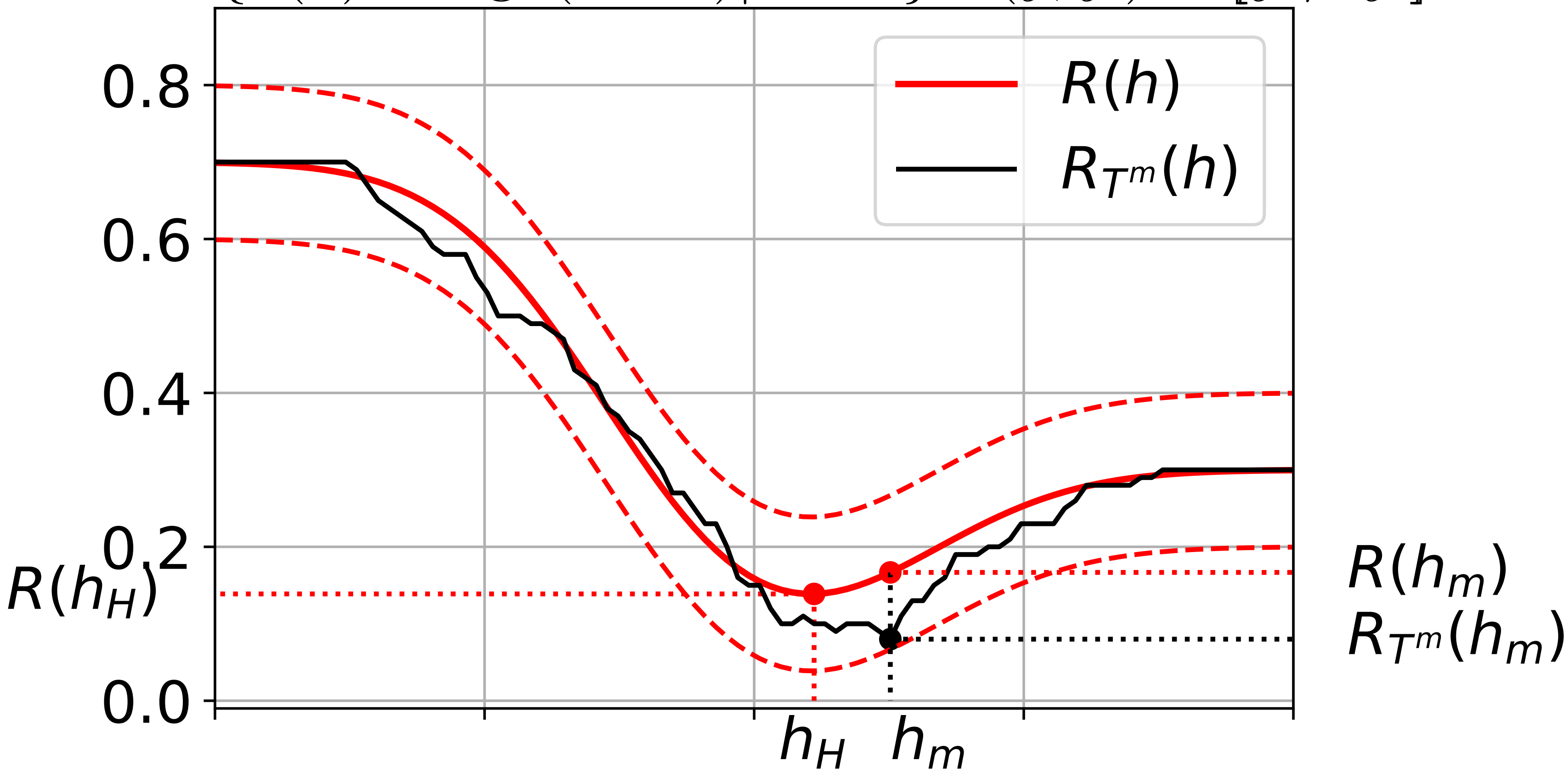
where $h_m = A(\mathcal{T}^m)$ is learned by A for $\mathcal{T}^m$ generated from $p(x, y)$.

Example: generalization error and estimation error

$$\mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big|}_{\text{highest generalization error}} \geq \varepsilon \Big) \leq B(m,\mathcal{H},\varepsilon)$$

$$\mathbb{P}\Big(\underbrace{R(h_m) - R(h_{\mathcal{H}})}_{\text{estimation error}} \geq \varepsilon \Big) \leq \mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \tfrac{\varepsilon}{2}}_{\text{highest generalizaton error}} \Big)$$

$$\mathcal{H} = \{h(x) = \text{sign}(x-\theta) | \theta \in \mathbb{R}\},\ \ell(y,y') = [\![y \neq y']\!]$$



Theorem: ULLN implies consistency of ERM

For fixed $\mathcal{T}^m$ and $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ we have:

$$\begin{aligned} R(h_m) - R(h_{\mathcal{H}}) &= \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}})\Big) \\ &\le \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big) \\ &\le 2 \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \end{aligned}$$

Therefore $\varepsilon \le R(h_m) - R(h_{\mathcal{H}})$ implies $\frac{\varepsilon}{2} \le \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big|$ and

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \ge \varepsilon \Big) \le \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \ge \frac{\varepsilon}{2} \Big)$$

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).

Finite sample bound on the estimation error

- Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the hypothesis class with the best predictor

$$h_{\mathcal{H}} \in \operatorname*{Argmin}_{h \in \mathcal{H}} R(h)$$

- We learn h_m from $\mathcal{T}^m \sim p(x, y)$ with the ERM algorithm

$$h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$$

- Assume we have for $\mathcal{H}$ the uniform bound on the generalization error

$$\mathbb{P}\Big(\sup_{h \in \mathcal{H}} \big| R_{\mathcal{T}^m}(h) - R(h) \big| \geq \varepsilon \Big) \leq B(m, \mathcal{H}, \varepsilon)$$

- Then, for any $\varepsilon > 0$ the inequality

$$R(h_m) \leq R(h_{\mathcal{H}}) + \varepsilon$$

holds with the probability $1 - B(m, \mathcal{H}, \varepsilon\,/2)$ at least.

Excess error = Estimation error + Approximation errors

The characters of the play:

- $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$ best attainable true risk
- $R(h_\mathcal{H})$ best risk in $\mathcal{H}$ where $h_\mathcal{H} \in \mathrm{Argmin}_{h \in \mathcal{H}} R(h)$
- $R(h_m)$ risk of $h_m = A(\mathcal{T}_m)$ learned from $\mathcal{T}^m$

Excess error: the quantity we want to minimize

$$\underbrace{\Big(R(h_m) - R^*\Big)}_{\text{excess error}} = \underbrace{\Big(R(h_m) - R(h_\mathcal{H})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h_\mathcal{H}) - R^*\Big)}_{\text{approximation error}}$$

Questions:

- What causes individual errors ?
- How do the errors depend on $\mathcal{H}$ and m?