Statistical Machine Learning (BE4M33SSU)
Lecture 4: Probably Approximately Correct Learning

Czech Technical University in Prague
V. Franc

Error decomposition

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

Error decomposition:

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

- The approximation error: depends on $\mathcal{H}$ chosen prior to learning.
- The estimation error: depends on $\mathcal{H}$, data $\mathcal{T}$ and the algorithm A.

Probably Approximately Correct (PAC) learning

Successful PAC learning algorithm

- Given a hypothesis space $\mathcal{H}$ and the loss ℓ, the algorithm with high probability learns a predictor that has low estimation error.
- The following can be arbitrary: desired estimation error $\varepsilon > 0$, probability of failure $\delta \in (0, 1)$, and data distribution $p(x, y)$.

Definition. Algorithm is a successful PAC learner for hypothesis space $\mathcal{H}$ w.r.t. loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ if there exists a function (called sample complexity) $m_{\mathcal{H}}^{\text{pac}}\colon \mathbb{R}_{>0} \times (0, 1) \to \mathbb{N}$ such that: For every $\varepsilon > 0$, $\delta \in (0, 1)$, and every distribution $p(x, y)$, when running the algorithm on $m \geq m_{\mathcal{H}}^{\text{pac}}(\varepsilon, \delta)$ examples $\mathcal{T}^m$ i.i.d. drawn from $p(x, y)$, then the algorithm returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \leq \varepsilon\Big) \geq 1 - \delta\,.$$

ULLN implies that ERM is successful PAC learner

ULLN applies for $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$: there exists $m^{\rm ul}_{\mathcal{H}} \colon \mathbb{R}_{>0} \times (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0, \delta \in (0,1)$, every distribution $p(x,y)$ and every $m \geq m^{\rm ul}_{\mathcal{H}}(\varepsilon,\delta)$ it holds that

$$\mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon}_{\text{ER can fail}}\Big) \leq \delta\,.$$

Successful PAC learner for $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$: there exists $m^{\rm pac}_{\mathcal{H}} \colon \mathbb{R}_{>0} \times (0,1) \to \mathbb{N}$ such that when running the algorithm on $m \geq m^{\rm pac}_{\mathcal{H}}(\varepsilon,\delta)$ examples $\mathcal{T}^m \sim p^m$ then it returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}\Big(\underbrace{R(h_m) - R(h_{\mathcal{H}}) \leq \varepsilon}_{\text{low estimation error}}\Big) \geq 1 - \delta\,.$$

Theorem: If ULLN applies for $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ with a function $m^{\rm ul}_{\mathcal{H}}$ then ERM is a successful PAC learner for $\mathcal{H}$ with the sample complexity $m^{\rm pac}_{\mathcal{H}}(\varepsilon,\delta) = m^{\rm ul}_{\mathcal{H}}(\frac{\varepsilon}{2},\delta)$.

ULLN implies that ERM is successful PAC learner: proof (1)

ULLN: $m \geq m^{\text{ul}}_{\mathcal{H}}(\varepsilon,\delta) \quad \Rightarrow \quad \mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| > \varepsilon}_{\text{ER can fail}}\Big) \leq \delta$

$$\underbrace{R(h_m) - R(h_{\mathcal{H}})}_{\text{estimation error}} \leq 2\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big|$$

$$R(h_m) - R(h_{\mathcal{H}}) > \bar{\varepsilon} \quad \Rightarrow \quad \sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| > \tfrac{\bar{\varepsilon}}{2}$$

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) > \bar{\varepsilon}\Big) \leq \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| > \tfrac{\bar{\varepsilon}}{2}\Big)$$

$$m \geq m^{\text{ul}}_{\mathcal{H}}(\tfrac{\bar{\varepsilon}}{2},\delta) \quad \Rightarrow \quad \mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) > \bar{\varepsilon}\Big) \leq \delta$$

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \leq \bar{\varepsilon}\Big) = 1 - \mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) > \bar{\varepsilon}\Big) \geq 1 - \delta$$

Successful PAC: $m \geq m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon},\delta) \quad \Rightarrow \quad \mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \leq \bar{\varepsilon}\Big) \leq 1 - \delta$

where $m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon},\delta) = m^{\text{ul}}_{\mathcal{H}}(\tfrac{\bar{\varepsilon}}{2},\delta)$

ULLN implies that ERM is successful PAC learner: proof (2)

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

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

ERM is successful PAC learner for finite hypothesis space

- We showed that for finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$ it holds

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

and hence ULLN applies with $m^{\text{ul}}_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log 2|H| - \log \delta}{2\,\varepsilon^2}(\ell_{\max} - \ell_{\min})^2$.

- Therefore ERM is successful PAC learner for $\mathcal{H}$ with sample complexity

$$m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon}, \delta) = 2\frac{\log 2|H| - \log \delta}{\bar{\varepsilon}^2}(\ell_{\max} - \ell_{\min})^2\,,$$

that is, when running ERM on $\mathcal{T}^m$ with $m \geq m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon}, \delta)$ then it returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \leq \bar{\varepsilon} \Big) \geq 1 - \delta\,.$$

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 $h \colon \mathcal{X} \to \mathcal{Y}$, parametrized by a vector $\boldsymbol{w} \in \mathbb{R}^n$,

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

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

- ERM for $\mathcal{H} = \{h(x;\boldsymbol{w},b) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x)\rangle + b) \mid (\boldsymbol{w},b) \in \mathbb{R}^{n+1}\}$ leads to

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

  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.
- Does ULLN applies for the class of two-class linear classifiers? If yes then ERM is PAC successful learner.

Vapnik-Chervonenkis (VC) dimension

- VC dimension is a concept to measure complexity of an infinite hypothesis space $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$.

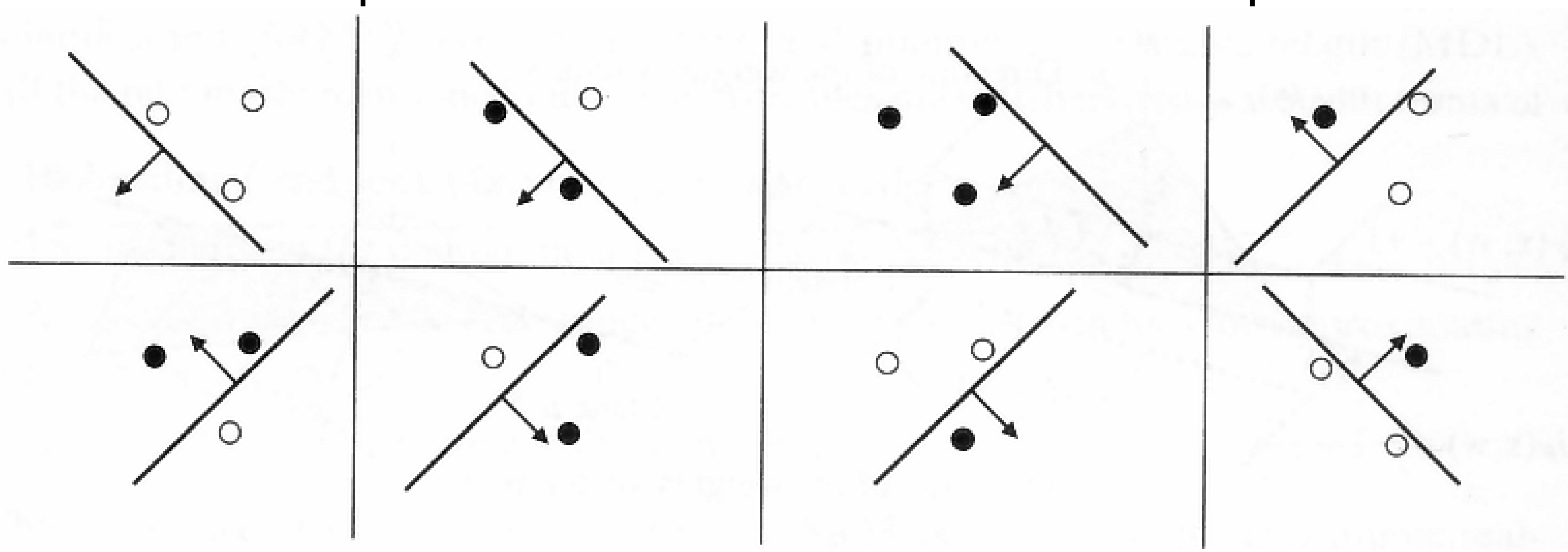
Definition: Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $\{x^1,\dots,x^m\} \in \mathcal{X}^m$ be a set of m input observations. The set $\{x^1,\dots,x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $\boldsymbol{y} \in \{+1,-1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i)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

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 space



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

Theorem: Let $\mathcal{H} \subset \{+1,-1\}^{\mathcal{X}}$ 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

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

Corollary: Let $\mathcal{H} \subset \{+1,-1\}^{\mathcal{X}}$ be a hypothesis class with a finite VC dimension $d < \infty$. Then, ULLN applies for $\mathcal{H}$ and there exists a constant C such that

$$m^{\text{pac}}_{\mathcal{H}}(\varepsilon, \delta) \leq C \frac{d - \log \delta}{\varepsilon^2}$$

that is, ERM is PAC successful learner.

Remark: Recall that in case of finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$ and $0/1$-loss we have the sample complexity $m^{\text{pac}}_{\mathcal{H}}(\varepsilon, \delta) = 2 \frac{\log 2|\mathcal{H}| - \log \delta}{\varepsilon^2}$.

Summary

- Error decomposition: Generalization error $=$ estimation error $+$ approximation error $+$ Bayes risk.
- Probably Approximately Correct (PAC) learning.
- ULLN implies that ERM is successful PAC learner.
- VC dimension: hypothesis space complexity of two-class classifier.
- VC dimension of linear hypothesis space.
- Finite VC dimension implies that ERM is a successful PAC learner.