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

Czech Technical University in Prague
V. Franc

Learning

- The goal: Find a strategy $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing $R(h)$ using the training set of examples

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

drawn from i.i.d. rv. with unknown $p(x, y)$.

- Hypothesis class (space):

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$$

- Learning algorithm: a function

$$A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$

which returns a strategy $h_m = A(\mathcal{T}^m)$ for a training set $\mathcal{T}^m$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples $\mathcal{T}^m$,

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM based algorithm returns h_m such that

$$h_m \in \operatorname*{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h) \tag{1}$$

- Depending on the choince of $\mathcal{H}$ and ℓ and algorithm solving (1) we get individual instances e.g. Support Vector Machines, Linear Regression, Logistic Regression, Neural Networks learned by back-propagation, AdaBoost, Gradient Boosted Trees, ...

Example of ERM failure

- Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$.

- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

- Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns memorizing strategy

$$h_m(x) = \begin{cases} y^j & \text{if } \; x = x^j \; \text{ for some } \; j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

- The empirical risk is $R_{\mathcal{T}^m}(h_m) = 0$ with probability 1 for any m.

- The expected risk is $R(h_m) = 0.8$ for any m.

Wrap up of the previous lecture

- We use the empirical risk $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(y^i))$ as a proxy of the true risk $R(h) = \mathbb{E}_{x,y\sim p}[\ell(y, h(x))]$.

- In case of evaluation, h is fixed and due to the law of large numbers, $R_{\mathcal{S}^l}(h)$ gets close to $R(h)$ if we have enough examples:

$$\mathbb{P}\Big(\big|R_{\mathcal{S}^l}(h) - R(h)\big| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(\ell_{max}-\ell_{min})^2}}$$

  We say that $R_{\mathcal{S}^l}(h)$ converges in probability to $R(h)$, i.e.

$$\forall\,\varepsilon > 0\colon\; \lim_{l\to\infty} \mathbb{P}\Big(\big|R_{\mathcal{S}^l}(h) - R(h)\big| \geq \varepsilon\Big) = 0$$

- In case of learning, $h_m = A(\mathcal{T}_m)$ is learned from $\mathcal{T}^m$ then $R_{\mathcal{T}^m}(h)$ does not have to get close to $R(h)$ even if we have enough examples:

$$\forall\,\varepsilon > 0\colon\; \lim_{m\to\infty} \mathbb{P}\Big(\big|R_{\mathcal{T}^m}(h_m) - R(h_m)\big| \geq \varepsilon\Big) \neq 0$$

Why law of large numbers does not apply for learning?

- Hoeffding inequality $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$, $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} z^i$, requires $\{z^1, \ldots, z^m\}$ to be sample from i.i.d. rv. with expeted value μ.

- $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ is drawn from i.i.d. rv. with $p(x, y)$.

Evaluation:

- h fixed independently on $\mathcal{T}^m$, $z^i = \ell(y^i, h(x^i))$ and $\{z^1, \ldots, z^m\}$ is i.i.d.
- Therefore $\forall\, \varepsilon > 0$: $\lim_{m\to\infty} \mathbb{P}(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) = 0$

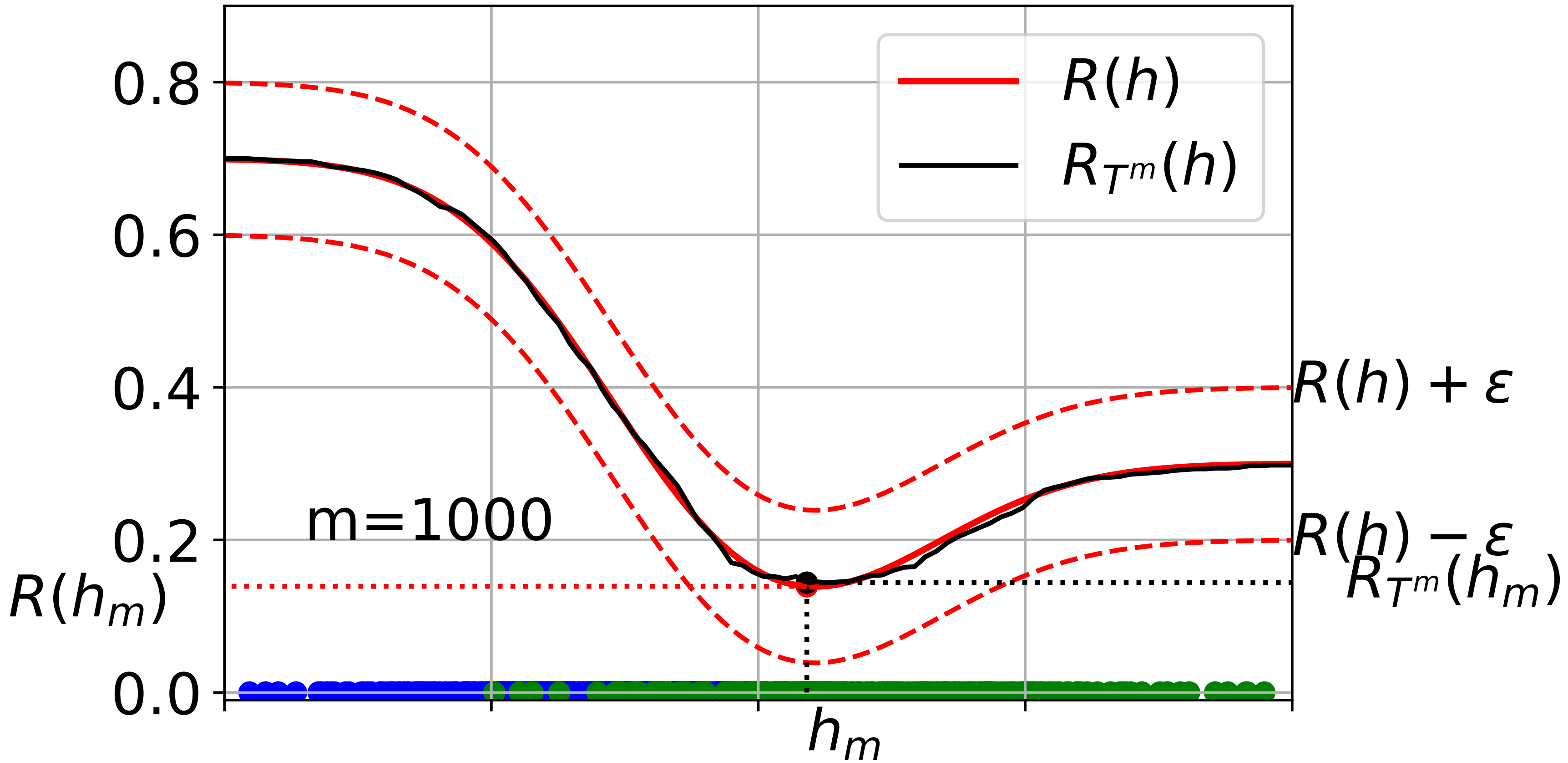
Learning:

- $h_m = A(\mathcal{T}^m)$, $z^i = \ell(y^i, h_m(x^i))$ and thus $\{z^1, \ldots, z^m\}$ is not i.i.d.
- No guarantee that $\forall\, \varepsilon > 0$: $\lim_{m\to\infty} \mathbb{P}(|R_{\mathcal{T}^m}(h_m) - R(h_m)| \geq \varepsilon) = 0$
- The task for the rest of the lecture is to show how to fix it.

To fix the problem we need uniform law of large numbers

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

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



Uniform Law of Large Numbers

- Law of Large Numbers: for any $p(x, y)$ generating $\mathcal{T}^m$, and $h \in \mathcal{H}$ fixed without using $\mathcal{T}^m$ we have

$$\forall \varepsilon > 0\colon \lim_{m \to \infty} \mathbb{P}\Big(\underbrace{\big| R(h) - R_{\mathcal{T}^m}(h) \big| \geq \varepsilon}_{\text{empirical risk fails for h}} \Big) = 0$$

- Uniform Law of Large Numbers: if for any $p(x, y)$ generating $\mathcal{T}^m$ it holds that

$$\forall \varepsilon > 0\colon \lim_{m \to \infty} \mathbb{P}\Big(\underbrace{\sup_{h \in \mathcal{H}} \big| R(h) - R_{\mathcal{T}^m}(h) \big| \geq \varepsilon}_{\text{empirical risk fails for some } h \in \mathcal{H}} \Big) = 0$$

  we say that ULLN applies for $\mathcal{H}$.

- Alternatively we say: the empirical risk converges uniformly to the true risk, or that the hypothesis class $\mathcal{H}$ has the uniform convergence property.

ULLN applies for finite hypothesis class

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- Define the set of all "bad" training sets for a strategy $h \in \mathcal{H}$ as

$$\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \,\middle|\, \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \right\}$$

- Hoeffding inequality generalized for finite hypothesis class $\mathcal{H}$:

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

- ULLN applies for finite hypothesis class

$$\forall\, \varepsilon > 0\colon \; \lim_{m \to \infty} \mathbb{P}\Big(\max_{h \in \mathcal{H}} \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \Big) = 0$$

Generalization bound for finite hypothesis class

- Hoeffding inequality generalized for a finite hypothesis class $\mathcal{H}$:

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

- Find an upper bound ε on the discrepancy between $R_{\mathcal{T}^m}(h)$ and $R(h)$ which holds uniformly for all $h \in \mathcal{H}$ with probability $1-\delta$ at least:

$$\begin{aligned}\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| < \varepsilon \Big) &= 1 - \mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \\ &\geq 1 - 2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = 1-\delta\end{aligned}$$

and solving the last equality for ε yields

$$\varepsilon = (b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$$

Generalization bound for finite hypothesis class

Theorem: Let $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be draw from i.i.d. rv. with p.d.f. $p(x, y)$ and let $\mathcal{H}$ be a finite hypothesis class. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality

$$R(h) \leq \underbrace{R_{\mathcal{T}^m}(h)}_{\text{empirical risk}} + \underbrace{(b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}}_{\text{complexity term}}$$

holds for all $h \in \mathcal{H}$ simultaneously and any loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [a, b]$.

- Recommendations that follow from the generalization bound:
  1. Minimize the empirical risk.
  2. Use as much training examples as possible.
  3. Limit the size of the hypothesis space $|\mathcal{H}|$:

  Note that 1) and 3) are conflicting recommendations.

- The generalization bound holds for any learning algorithm not just ERM.

Structural Risk Minimization

- Learn $h\colon \mathcal{X} \to \mathcal{Y}$ by minimizing the generalization bound

$$R(h) \le R_{\mathcal{T}^m}(h) + \underbrace{(b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}}_{\epsilon(m,|\mathcal{H}|,\delta)}$$