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

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

drawn from i.i.d. according to unknown $p(x, y)$.

- Hypothesis class:

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

- 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.

Generalization error

- ERM may fail when $R_{\mathcal{T}^m}(h_m)$ is not a good proxy of $R(h_m)$, because $R_{\mathcal{T}^m}(h)$ is used as a guidance to select h_m.

- We need the generalization error, i.e., the discrepancy between $R(h)$ and $R_{\mathcal{T}^m}(h)$, to become small when the number of examples m grows:

$$\forall\, \varepsilon > 0\colon \ \lim_{m\to\infty} \mathbb{P}\Big(\underbrace{\big|R_{\mathcal{T}^m}(h_m) - R(h_m)\big| \geq \varepsilon}_{\text{high generalization error}} \Big) = 0$$

  where $h_m = A(\mathcal{T}_m)$ is learned by $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$.

Plan for this lecture:

- Conditions on $\mathcal{H}$ which guarantee that the generalization error converges to zero with growing number of examples m.

- Generalization bound for a finite number of examples.

What's wrong with Hoeffding ?

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

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

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 seeing $\mathcal{T}^m$ we have

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

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

$$\forall\, \varepsilon > 0\colon \lim_{m\to\infty} \mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon}_{\substack{\text{high generalization error at least}\\ \text{for one hypothesis}}} \Big) = 0$$

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

- Note that for $h_m = A(\mathcal{T}_m)$ we have

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

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

$$\mathcal{B}(h) = \Big\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \Big| \big| R_{\mathcal{T}^m}(h) - R(h) \big| \geq \varepsilon \Big\}$$

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

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

- ULLN applies for finite hypothesis class

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

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 generalization error 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 and. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality

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

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

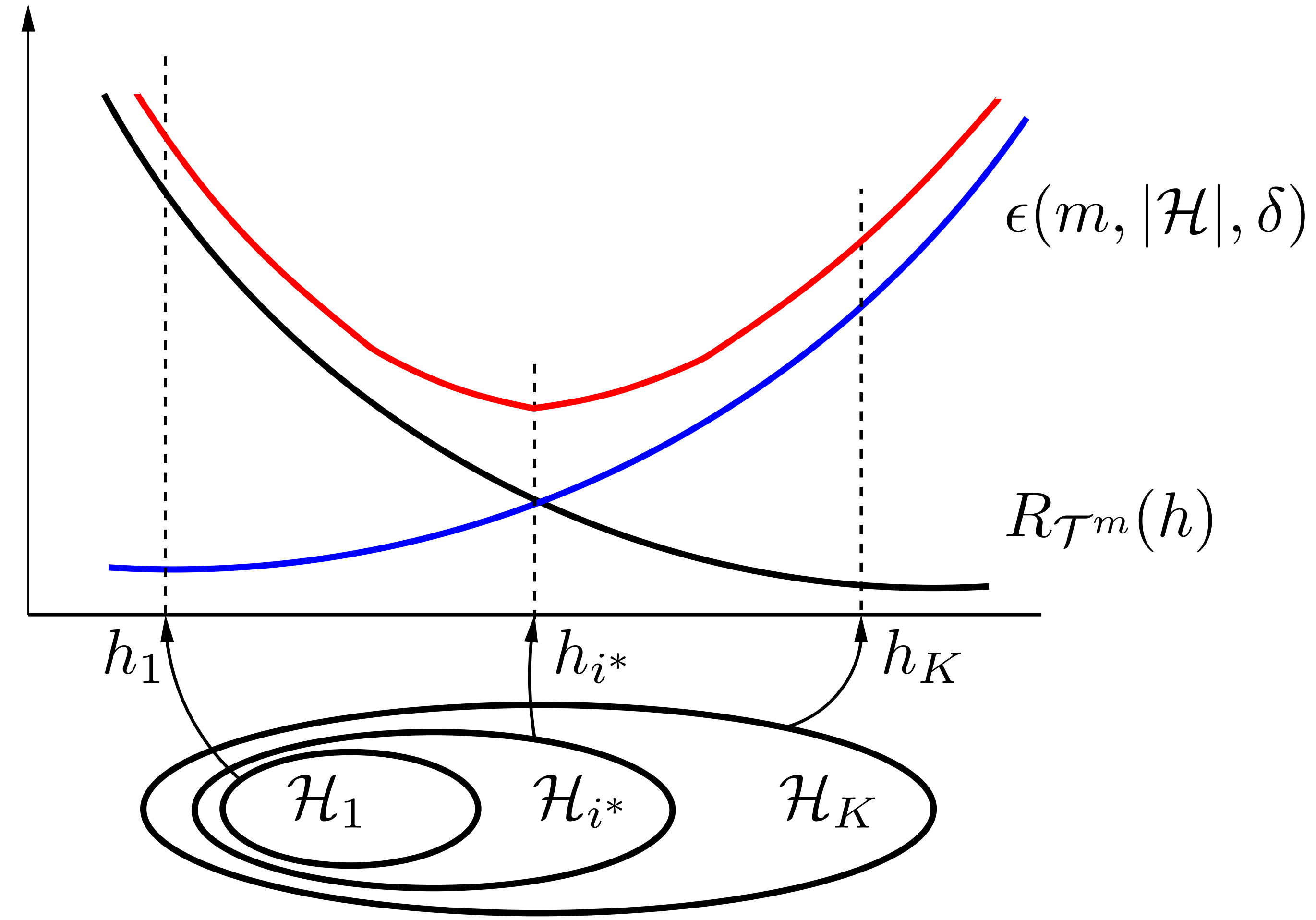
Recommendations that follow from the bound:

- We need to select appropriate trade-off between $|\mathcal{H}|$ and m:
- Little prior knowledge requires a lot of examples.
- Too complex hypothesis class may lead to overfitting.

Structural Risk Minimization

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

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


$\epsilon(m,|\mathcal{H}|,\delta)$
$R_{\mathcal{T}^m}(h)$
h_1
h_{i^*}
h_K
$\mathcal{H}_1$
$\mathcal{H}_{i^*}$
$\mathcal{H}_K$