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

Czech Technical University in Prague
V. Franc

Learning

- **Goal:** Given a training set $\mathcal{T}^m \sim p^m$, find a strategy $h\colon \mathcal{X} \to \mathcal{Y}$ with minimizing the generalization error $R(h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x))]$.

- Hypothesis class (space): fixed before learning based on prior knowledge

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

- Learning algorithm: a function

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

returns a strategy $h_m = A(\mathcal{T}^m)$ from $\mathcal{H}$ based on a training set $\mathcal{T}^m$

Empirical Risk Minimization learning

- The generalization error $R(h)$ is approximated by the empirical risk $R_{\mathcal{T}^m}(h)$ computed on the training examples $\mathcal{T}^m \sim p^m$:

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

- The ERM based learning algorithm returns h_m such that

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

- 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, ...

ERM can fail due to overfitting

- 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 $R^* = 0.2$.
- Learning algorithm "lookup table": given training set $\mathcal{T}^m$ it returns

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

  - Implements ERM principle as $\mathbb{P}(R_{\mathcal{T}^m}(h_m) = 0) = 1$.
  - Fails to find a good solution as $\mathbb{P}(R(h_m) = 0.8) = 1$, $\forall m \in \mathbb{N}$.
- Overfitting: the case when $h_m = A(\mathcal{T}^m)$ and the training error $R_{\mathcal{T}^k}(h_m)$ is low while the generalization error $R(h_m)$ is high.
- **Problem:** under which conditions the overfitting can be eliminated?

Why the 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 independent random variables with the expected 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.

- We can apply Hoeffding $\mathbb{P}(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) \leq 2e^{-\frac{2m\,\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}}$

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.

- We cannot apply Hoeffding to bound $\mathbb{P}(|R_{\mathcal{T}^m}(h_m) - R(h_m)| \geq \varepsilon)$

The overfitting can be eliminited in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have anough examples:

$$\mathbb{P}\Big(\underbrace{\big|R(h_m) - R_{\mathcal{T}^m}(h_m)\big| \geq \varepsilon}_{\text{ER fails}}\Big) \stackrel{(1)}{\leq} \mathbb{P}\Big(\max_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big)$$

$$\stackrel{(2)}{\leq} \sum_{h\in\mathcal{H}} \mathbb{P}\Big(\big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big)$$

$$\stackrel{(3)}{\leq} 2\,|\mathcal{H}|\, e^{-\frac{2m\,\varepsilon^2}{(\ell_{\max}-\ell_{\min})^2}}$$

1. $\mathbb{P}($ ER fails for $h_m \in \mathcal{H})$ is replaced by $\mathbb{P}($ ER can fail for some $h \in \mathcal{H})$.
2. Union bound.
3. Hoeffding inequality.

Uniform Law of Large Numbers

We have shown for that for the finite hypothesis space, $\mathcal{H} = \{h_1, \ldots, h_K\}$, the Law of Large Numbers holds simultaneously (uniformly) for every $h \in \mathcal{H}$:

$$\underbrace{\mathbb{P}\Big(\max_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big)}_{\text{ER can fail}} \leq \underbrace{2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(\ell_{\max}-\ell_{\min})^2}}}_{\text{converges to } 0 \text{ for } m\to\infty}$$

Definition: We say that Uniform Law of Large Numbers applies for hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ if there exists a function $m_{\mathcal{H}} \colon \mathbb{R}_{>0} \times (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0, \delta \in (0,1)$, every distribution $p(x,y)$ and every $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ the following inequality holds

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

The next lecture:

- If ULLN applies then ERM learning is guaranteed to succeed.
- VC dimension as a tool to recognize that ULLN applies for given $\mathcal{H}$.

Generalization bound for finite hypothesis class

Theorem: Let $\mathcal{T}^m = \big((x^1, y^1), \ldots, (x^m, y^m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$ 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

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

holds for all $h \in \mathcal{H}$ simultaneously.

- To decreases the complexity term: increase m or decrease $|\mathcal{H}|$.
- The generalization bound holds for any learning algorithm not just ERM.
- Recommendations for learning:
  1. Minimize the empirical risk.
  2. Use as much training examples m as you can.
  3. Limit the size of the hypothesis space $|\mathcal{H}|$, i.e. use prior knowledge.

Generalization bound for finite hypothesis class: the proof

- We have shown that ULLN holds for finite hypothesis class $\mathcal{H}$:

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

- Prob. $R_{\mathcal{T}^m}(h)$ is a good proxy of $R(h)$ for all $h \in \mathcal{H}$ simultaneously:

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

- Solving the last equality for ε yields $\varepsilon = L\sqrt{\frac{\log 2|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$ so that:

$$\mathbb{P}\left(|R_{\mathcal{T}^m}(h) - R(h)| < L\sqrt{\frac{\log 2|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}\,,\ \forall h \in \mathcal{H} \right) \geq 1 - \delta$$

Summary

- Learning algorithm: the definition.
- Empirical Risk Minimization.
- Unrestricted hypothesis space: the ERM can overfit regardless the number of training examples.
- Finite hypothesis space: the chance of overfitting can be always eliminated.
- Uniform Law of Large Numbers.
- Generalization bound for finite hypothesis space.