Statistical Machine Learning (BE4M33SSU)  
Lecture 2: Empirical Risk  

Czech Technical University in Prague  
V. Franc
Prediction problem: the definition

- $\mathcal{X}$ a set of input **observations/features**
- $\mathcal{Y}$ a finite set of **hidden states**
- $h: \mathcal{X} \rightarrow \mathcal{Y}$ a **prediction strategy**
- $(x, y) \in \mathcal{X} \times \mathcal{Y}$ samples **randomly drawn** from r.v. with p.d.f. $p(x, y)$
- $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ a **loss function**

**Task** is to find a strategy with the minimal **expected risk**

$$ R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) \, p(x, y) \, dx = \mathbb{E}_{(x,y) \sim p}(\ell(y, h(x))) $$
Example of a prediction problem

- The statistical model:

  - $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases}$

  - $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu y)^2}$, $y \in \mathcal{Y}$. 

![Graph showing prediction problem](image)
Solving the prediction problem from examples

- **Assumption**: we have an access to examples

  \[ \{(x^1, y^1), (x^2, y^2), \ldots \} \]

  drawn from i.i.d. r.v. distributed according to unknown \( p(x, y) \).

- **1) Testing**: estimate \( R(h) \) of a given \( h : \mathcal{X} \rightarrow \mathcal{Y} \) using test set

  \[ S^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l \} \]

  drawn i.i.d. from \( p(x, y) \).

- **2) Learning**: find \( h : \mathcal{X} \rightarrow \mathcal{Y} \) with small \( R(h) \) using training set

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

  drawn i.i.d. from \( p(x, y) \).
Testing: estimation of the expected risk

- Given a predictor \( h : \mathcal{X} \rightarrow \mathcal{Y} \) and a test set \( S^l \) draw i.i.d. from distribution \( p(x, y) \), compute the empirical risk

\[
R_{S^l}(h) = \frac{1}{l} \left( \ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) \right) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))
\]

and use it as an estimate of \( R(h) = \mathbb{E}_{(x, y) \sim p}(\ell(y, h(x))) \).

- The empirical risk \( R_{S^l}(h) \) is a random variable.

- We will show how to compute an interval such that

\[
R(h) \in (R_{S^l}(h) - \varepsilon, R_{S^l}(h) + \varepsilon) \text{ with probability (confidence) } \gamma \in (0, 1)
\]

- We will show relation between \( \varepsilon, l \) and \( \gamma \).
Law of large numbers

- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

- Example: The expected value of a single roll of a fair die is

\[ \mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^{6} z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5 \]

\[ \hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z_i \]

\[ z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad z^l = 2 \]
Theorem 1. Let $\{z^1, \ldots, z^l\} \in [a, b]^l$ be a sample from i.i.d. r.v. with expected value $\mu$. Let $\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i$. Then for any $\varepsilon > 0$ it holds that

$$\mathbb{P}\left( |\hat{\mu} - \mu| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}}$$

Example (rolling a die): $\mu = 3.5$, $z_i \in [1, 6]$, $\varepsilon = 0.5$. 
Confidence intervals

\((l, \gamma) \rightarrow \varepsilon\)

- Let \(\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i\) be the sample average computed from \(\{z^1, \ldots, z^l\} \in [a, b]^l\) sampled from r.v. with expected value \(\mu\).

- Find \(\varepsilon\) such that \(\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)\) with probability at least \(\gamma\).

Using the Hoeffding inequality we can write

\[
\mathbb{P}(|\hat{\mu} - \mu| < \varepsilon) = 1 - \mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \geq 1 - 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}} = \gamma
\]

and solving the last equation for \(\varepsilon\) yields

\[
\varepsilon = |b - a| \sqrt{\frac{\log(2) - \log(1 - \gamma)}{2l}}
\]
Confidence intervals

$(\varepsilon, \gamma) \rightarrow l$

Let $\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i$ be the sample average computed from 
\( \{z^1, \ldots, z^l\} \in [a, b]^l \) sampled from r.v. with expected value $\mu$.

Given a fixed $\varepsilon > 0$ and $\gamma \in (0, 1)$, what is the minimal number of examples $l$ such that $\mu \in (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$ with probability $\gamma$ at least?

Starting from

\[
\mathbb{P}\left( |\hat{\mu} - \mu| < \varepsilon \right) = 1 - \mathbb{P}\left( |\hat{\mu} - \mu| \geq \varepsilon \right) \geq 1 - 2e^{-\frac{2l \varepsilon^2}{(b-a)^2}} = \gamma
\]

and solving for $l$ yields

\[
l = \frac{\log(2) - \log(1 - \gamma)}{2 \varepsilon^2} (b - a)^2
\]
Testing: estimation of the expected risk

- Given \( h: \mathcal{X} \rightarrow \mathcal{Y} \) estimate the expected risk \( R(h) = \mathbb{E}_{(x,y) \sim p}(\ell(y, h(x))) \) by the empirical risk \( R_{S^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i)) \) using the test set \( S^l \).

- The incurred losses \( z^i = \ell(y^i, h(x^i)) \in [\ell_{\text{min}}, \ell_{\text{max}}], i \in \{1, \ldots, l\} \), are realizations of i.i.d. r.v. with the expected value \( \mu = R(h) \).

- According to the Hoeffding inequality, for any \( \varepsilon > 0 \) the probability of seeing a “bad test set” can be bound by

\[
\mathbb{P}\left( \left| R_{S^l}(h) - R(h) \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l \varepsilon^2}{(\ell_{\text{min}} - \ell_{\text{max}})^2}}
\]
Testing: recipe for constructing confidence intervals

- Given $h : \mathcal{X} \rightarrow \mathcal{Y}$ estimate the expected risk $R(h) = \mathbb{E}_{(x,y) \sim p}(\ell(y, h(x)))$ by the empirical risk $R_{Sl}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$ using the test set $S^l = \{(x^1, y^1), \ldots, (x^l, y^l)\}$.

- **Confidence interval:**

  $$R(h) \in \left( R_{Sl}(h) - \varepsilon, R_{Sl}(h) + \varepsilon \right) \quad \text{with probability} \quad \gamma \in (0, 1)$$

- **Interval width:** For fixed $l$ and $\gamma \in (0, 1)$ compute

  $$\varepsilon = (\ell_{\max} - \ell_{\min}) \sqrt{\frac{\log(2) - \log(1 - \gamma)}{2l}}.$$ 

- **Number of examples:** For fixed $\varepsilon$ and $\gamma \in (0, 1)$ compute

  $$l = \frac{\log(2) - \log(1 - \gamma)}{2\varepsilon^2} (\ell_{\max} - \ell_{\min})^2.$$
Example: confidence intervals

The width of $R(h) \in (R_{Sl}(h) - \varepsilon, R_{Sl}(h) + \varepsilon)$ is for $\ell(y, y') = [y \neq y']$ given by

$$\varepsilon = \sqrt{\frac{\log(2) - \log(1 - \gamma)}{2I}}$$

for $\gamma = 0.95$.

<table>
<thead>
<tr>
<th>$l$</th>
<th>100</th>
<th>1,000</th>
<th>10,000</th>
<th>18,445</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>0.135</td>
<td>0.043</td>
<td>0.014</td>
<td>0.01</td>
</tr>
</tbody>
</table>