Statistical Machine Learning (BE4M33SSU)
Lecture 2: Empirical Risk

Czech Technical University in Prague
V. Franc
Definition of the prediction problem

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

Task: find a strategy with the minimal true risk (expected loss)

\[
R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) \ p(x, y) \ dx = \mathbb{E}_{(x, y) \sim p} \left( \ell(y, h(x)) \right)
\]

Bayes predictor $h^*$ attains the minimal risk $R(h^*) = \inf_{h \in \mathcal{Y}^X} R(h)$
Example of a prediction problem

- The statistical model is known:

  - \( \mathcal{X} = \mathbb{R} \), \( \mathcal{Y} = \{+1, -1\} \), \( \ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases} \)
  
  - \( p(x, y) = p(y) \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2} \), \( y \in \mathcal{Y} \).
Machine Learning: solving the prediction problem based on 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) Evaluation:** estimate true risk \( R(h) \) of given \( h: \mathcal{X} \to \mathcal{Y} \) 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 \( h: \mathcal{X} \to \mathcal{Y} \) 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 \( p(x, y) \).
Evaluation: 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

$$R_{S^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$

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

- $R_{S^l}(h)$ is a random number with the variance depending on $l$.

- We construct a confidence interval such that

$$R(h) \in (R_{S^l}(h) - \varepsilon, R_{S^l}(h) + \varepsilon)$$

with probability (confidence) $\gamma \in (0, 1)$

where $\varepsilon$ is a deviation.
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

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

$$
\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i
$$

$$
\begin{align*}
z^1 &= 3 \\
z^2 &= 1 \\
z^3 &= 5 \\
z^l &= 2
\end{align*}
$$

[Graph showing rolling a die: 5 experiments]
Rolling a die: 100 experiments

Hoeffding inequality

\[
\frac{\#\left( |\hat{\mu} - \mu| \geq \varepsilon \right)}{\# \text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\left( |\hat{\mu} - \mu| \geq \varepsilon \right) \leq 2e^{-\frac{2\varepsilon^2}{(b-a)^2}}
\]

\( a = 1, \ b = 6 \)
Hoeffding inequality

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

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

**Properties:**

- **Conservative:** the bound may not be tight.
- **General:** the bound holds for any distribution.
- **Cheap:** The bound is simple and easy to compute.
Confidence intervals

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

- Let \(\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} z^i\) be the sample mean 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

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

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 mean 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

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

and solving for \(l\) yields

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

- Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ estimate the true 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 $z^i = \ell(y^i, h(x^i)) \in [\ell_{\min}, \ell_{\max}]$, $i \in \{1, \ldots, l\}$, 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

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

- **Remark:** For any $p(x, y)$ and $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [\ell_{\min}, \ell_{\max}]$, the empirical risk $R_{S^l}(h)$ convergences in probability to the true risk $R(h)$:

$$
\forall \varepsilon > 0: \lim_{l \rightarrow \infty} \mathbb{P}\left(\left|R_{S^l}(h) - R(h)\right| \geq \varepsilon\right) = 0
$$
Evaluation: recipe for constructing confidence intervals

- Given $h : \mathcal{X} \to \mathcal{Y}$ estimate the true 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 $S^l$.

- Confidence interval:

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

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

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

- For fixed $\varepsilon$ and $\gamma \in (0, 1)$ compute number of test examples

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

- The width of $R(h) \in (R_{S_l}(h) - \varepsilon, R_{S_l}(h) + \varepsilon)$ is for $\ell(y, y') = [y \neq y']$
given by $\varepsilon = \sqrt{\frac{\log(2) - \log(1 - \gamma)}{2l}}$

![Graph showing confidence intervals for classification error](image)

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>

- Example: $l = 10,000$, $R_{S_l}(h) = 0.162$, then classification error is $16.2 \pm 1.4\%$ with confidence 95%.