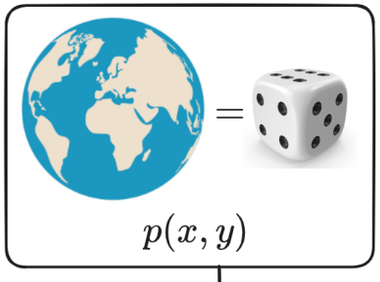
Machine Learning Fundamentals - LS2026
Empirical Risk Minimization

Czech Technical University in Prague
V. Franc

- What is a learning algorithm?
- How training data and the prior knowledge influence the performance on unseen data?
- Do we need to use a prior knowledge at all if we have planty of data?
- What is the performance on training data saying about performance on unseen data?

Empirical Risk Minimization

ERM: a principle to construct algorithms learning predictors from data.



$p(x, y)$

Hypothesis class

$\mathcal{H} = \{h(x;\theta) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$

Loss function

$$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if} \quad y = \hat{y} \\ 1 & \text{if} \quad y \neq \hat{y} \end{cases}$$

Training set

i.i.d. data

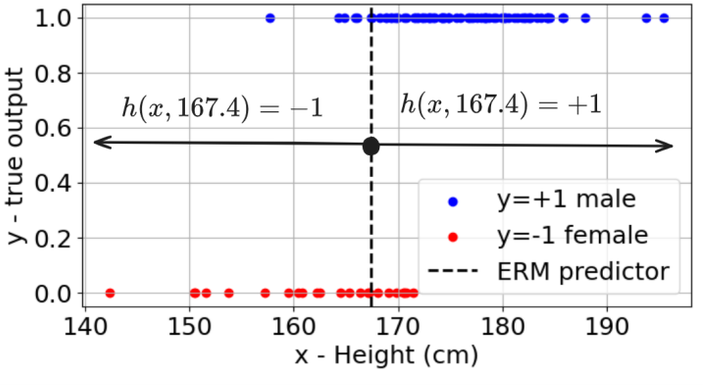
$T_m = ((x_1, y_1), \ldots, (x_m, y_m))$

x	y
179.5	+1
161.5	+1
176.9	+1
157.7	-1
184.0	+1
167.4	+1
175.1	+1

Learning algorithm based on ERM

Find a predictor from $\mathcal{H}$ which minimizes the empirical error:

$$\hat{R}(T_m, h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i;\theta))$$



Predictor

$h_m(x, 167.4) = \text{sign}(x - 167.4)$

Instances: Linear regression, Logistic Regression, Neural Networks learn by back-propagation, Gradient Boosted Trees, ...

Empirical Risk Minimization

- **Goal of learning:** Given a training set $T_m = ((x_1, y_1), \ldots, (x_m, y_m)) \sim p^m$, learn a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ with small true risk $R(p, h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x))]$.

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

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

- **Learning algorithm:** is a function

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

- **Emprirical risk** evaluated on T_m (a.k.a training error):

$$\hat{R}(T_m, h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i))$$

- **ERM based learning algorithm**:

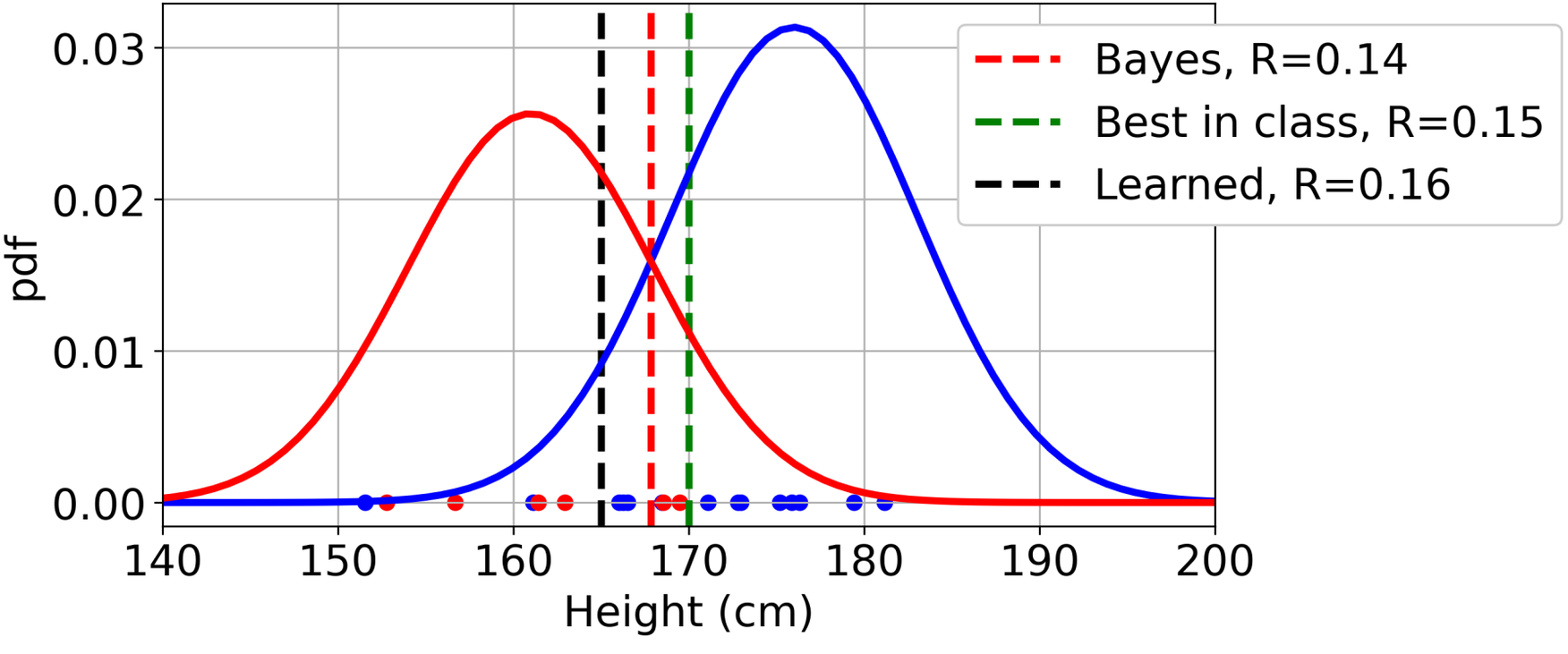
$$h_m = A(T_m) = \operatorname*{Argmin}_{h\in\mathcal{H}} \hat{R}(T_m, h)$$

Error decomposition

Errors:

1. Best (Bayes) attainable risk $R(p, h_*)$, where $h_*(x) = \arg\min_{h \in \mathcal{Y}^{\mathcal{X}}} R(p, h)$
2. Best risk in the class $R(p, h_{\mathcal{H}})$, where $h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(p, h)$
3. Risk of the learned predictor $R(p, h_m)$, where $h_m = A(T_m)$

$$\mathcal{H} = \{h(x, \theta) = \text{sign}(x - \theta) \mid \theta \in \{140, 145, \ldots, 200\}\}$$



Error decomposition

Errors:

1. Best (Bayes) attainable risk $R(p, h_*)$, where $h_*(x) = \arg\min_{h \in \mathcal{Y}^{\mathcal{X}}} R(p, h)$
2. Best risk in the class $R(p, h_{\mathcal{H}})$, where $h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(p, h)$
3. Risk of the learned predictor $R(p, h_m)$, where $h_m = A(T_m)$

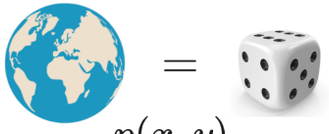
Error decomposition:

$$\underbrace{R(p, h_m)}_{\text{learned predictor risk}} = \underbrace{\Big(R(p, h_m) - R(p, h_{\mathcal{H}})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(p, h_{\mathcal{H}}) - R(p, h_*)\Big)}_{\text{approximation error}} + \underbrace{R(p, h_*)}_{\text{Bayes risk}}$$

- The approximation error: depends on $\mathcal{H}$ which represents our prior knowledge.
- The estimation error: depends on $\mathcal{H}$ and the training data T_m.
- Best (Bayes) attainable risk: irreducible error.

Big Picture of ERM based learning

Empirical Risk Minimization



Hypothesis class

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

Loss function

$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Training set

i.i.d. data

$T_m = ((x_1, y_1), \ldots, (x_m, y_m)) \sim p^m$

ERM Learning Algorithm

Learning algorithm $A \colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ finding a predictor in $\mathcal{H}$ which minimizes the empirical error:

$\hat{R}(T_m, h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i; \theta))$

Predictor

$\hat{y} = h_m(x)$

Error decomposition

Predictor Error = Estimation Error + Approximation Error + Bayes Error

Uniform Law of Large Numbers

Empirical risk of all member in $\mathcal{H}$ is a good proxy of the true risk.

ULLN holds for $\mathcal{H} \iff \mathcal{H}$ is PAC learnable $\iff$ ERM is succesful PAC learner

VC dimension

0/1-loss + binary classifier

$\mathrm{VCdim} \colon \{-1, +1\}^{\mathcal{X}} \to \mathbb{N}$

$\mathrm{VCdim}(\mathcal{H}) < \infty \iff$ ULLN holds for $\mathcal{H}$

PAC learning

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists an lagorithm that, with high probability, can make the estimation error arbitrarily small given sufficiently many examples

Hypotheisis class represents our prior knowledge

"Too complex" hypothesis class

E.g. Memorizer

$\mathcal{H}$ is not PAC learner

$\ldots$

Finite hypothesis class

$\mathcal{H} = \{h_1, h_2, \ldots, h_{\mathcal{H}}\}$

ULLN holds for $\mathcal{H}$

ERM Fails When Hypothesis Class is Excessively Rich

- **Setup:** $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ uniform on $\mathcal{X}$, with $p(y = +1) = 0.8$. The Bayes optimal predictor $h_*(x) = +1$ with true risk $R(p, h_*) = 0.2$.

- "Memorizer" learning algorithm: for training set $T_m = ((x_1, y_1), \ldots, (x_m, y_m))$, define

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

- Let $\mathcal{H}$ contain all "memorizing" predictors and the Bayes optimal predictor h_*. Thus:
  - The best-in-class predictor is the Bayes predictor, $h_{\mathcal{H}} = h_*$.

- **What we can say about the memorizer learning algorithm?**
  - Implements ERM over $\mathcal{H}$: $\hat{R}(T_m, h_m) = 0$
  - The learned predictor's risk: $R(p, h_m) = 0.8$
  - The estimation error: $R(p, h_m) - R(p, h_{\mathcal{H}}) = 0.8 - 0.2 = 0.6$

- **Conclusion:**
  - If the hypothesis class is excessively rich (weak prior knowledge), empirical risk is not a good proxy of the true risk regardless of the sample size m.
  - Although $\mathcal{H}$ contains the Bayes predictor (zero approximation error), ERM selects an overfitting hypothesis.

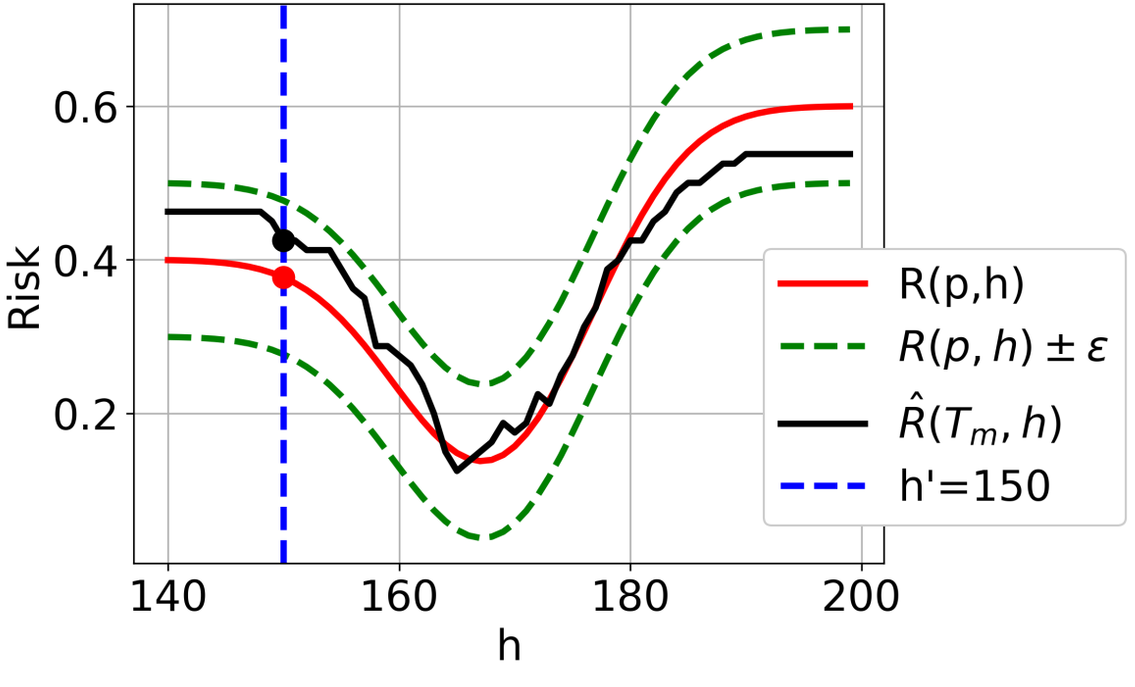
Uniform Law of Large Numbers

- **LLN**: $m \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right) \quad \Rightarrow \quad \mathbb{P}\Big(\underbrace{\big|R(p,h) - \hat{R}(T_m,h)\big| > \varepsilon}_{\text{generalization gap of } h \text{ is high}} \Big) \leq \delta$

  LLN applies for any $h\colon \mathcal{X} \to \mathcal{Y}$ fixed prior to observating the data T_m

- **ULLN:** $m \geq m^{\mathcal{H}}_{\text{ul}}(\varepsilon,\delta) \quad \Rightarrow \quad \mathbb{P}\Big(\underbrace{\max_{h\in\mathcal{H}} \big|R(p,h) - \hat{R}(T_m,h)\big| > \varepsilon}_{\text{exists } h\in\mathcal{H} \text{ with a hight generalization gap}} \Big) \leq \delta$

  ULLN applies only for some $\mathcal{H}$, e.g., when $\mathcal{H}$ is finite $m^{\mathcal{H}}_{\text{ul}}(\varepsilon,\delta) = \frac{1}{2\varepsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$



Example:

$\mathcal{H} = \{h(x;\theta) = \text{sign}(x-\theta) \mid \theta \in \{140, 141, \ldots, 200\}\}$

$\varepsilon = 0.1,\ \delta = 0.05,\ |\mathcal{H}| = 60$

$m^{\mathcal{H}}_{\text{ul}}(0.1, 0.05) = 389.2$

Uniform Law of Large Numbers

- Assume a finite hypothesis class $\mathcal{H} = \{h^i \colon \mathcal{X} \to \mathcal{Y} \mid i \in \{1, 2, \ldots, H\}\}$.
- By combining Hoeffding's inequality with the union bound, we obtain the uniform deviation bound:

$$\mathbb{P}\Big(\underbrace{\max_{h \in \mathcal{H}} \big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon}_{\text{exits } h \in \mathcal{H} \text{ with a high generalization gap}} \Big) \overset{(1)}{=} \mathbb{P}\left(\begin{array}{ll} \big|R(p,h^1) - \hat{R}(T_m,h^1)\big| \geq \varepsilon & \text{or} \\ \big|R(p,h^2) - \hat{R}(T_m,h^2)\big| \geq \varepsilon & \text{or} \\ \vdots & \\ \big|R(p,h^H) - \hat{R}(T_m,h^H)\big| \geq \varepsilon & \end{array} \right)$$

$$\overset{(2)}{\leq} \sum_{h \in \mathcal{H}} \mathbb{P}\Big(\big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon \Big)$$

$$\overset{(3)}{\leq} 2\,|\mathcal{H}|\, e^{-2m\varepsilon^2}$$

1. $a \geq \varepsilon$ or $b \geq \varepsilon \iff \max\{a,b\} \geq \varepsilon$
2. Union bound: $\mathbb{P}\Big(A_1 \text{ or } A_2 \text{ or} \cdots \text{or } A_n\Big) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$
3. Hoeffding's inequality: $\mathbb{P}\Big(\big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon \Big) \leq 2\, e^{-2\,m\,\varepsilon^2}$

Uniform Law of Large Numbers

- Assume a finite hypothesis class $\mathcal{H} = \{h^i \colon \mathcal{X} \to \mathcal{Y} \mid i \in \{1, 2, \ldots, H\}\}$.
- By combining Hoeffding's inequality with the union bound, we obtain the uniform deviation bound:

$$\mathbb{P}\Big(\underbrace{\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big| \geq \varepsilon}_{\text{exits } h \in \mathcal{H} \text{ with a high generalization gap}} \Big) \leq 2\,|\mathcal{H}|\, e^{-2m\varepsilon^2}$$

- Setting $2\,|\mathcal{H}|\, e^{-2m\varepsilon^2} = \delta$ and solving for m, we show that the ULLN applies for the finite hypotheses class:

$$m_{\text{ul}}^{\mathcal{H}}(\varepsilon, \delta) = \frac{1}{2\varepsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right) \qquad \Rightarrow \qquad \mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R(h) - \hat{R}(T_m, h) \big| \geq \varepsilon \Big) \leq \delta$$

- We will see (next lecture) that if ULLN applies for given $\mathcal{H}$, then the ERM based lagorithm is guaranteed to successed.

Generalization bound for finite hypothesis class

- Assume a finite hypothesis class $\mathcal{H} = \{h^i \colon \mathcal{X} \to \mathcal{Y} \mid i \in \{1, 2, \ldots, H\}\}$.
- By combining Hoeffding's inequality with the union bound, we obtain the uniform deviation bound:

$$\mathbb{P}\Big(\underbrace{\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big| \geq \varepsilon}_{\text{exits } h \in \mathcal{H} \text{ with a high generalization gap}} \Big) \leq 2\,|\mathcal{H}|\, e^{-2m\varepsilon^2}$$

- Solving $2\,|\mathcal{H}|\, e^{-2m\varepsilon^2} = \delta$ for ε, and pluging the result back to the uniform deviation bound yields, that for any $\delta \in (0, 1)$ we have:

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} \big| \hat{R}(T_m, h) - R(p,h) \big| \geq \sqrt{\frac{\ln(2|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)}{2m}} \right) \leq \delta.$$

- Equivalently, with probability at least $1 - \delta$, we have simultaneously for all $h \in \mathcal{H}$:

$$|R(p,h) - \hat{R}(T_m, h)| < \sqrt{\frac{\ln(2|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)}{2m}}$$

Generalization bound for finite hypothesis class

Theorem: Let $T_m = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$ be draw from i.i.d. rv. with p.d.f. $p(x, y)$ and let $\mathcal{H}$ be a finite hypothesis space. Let $R(p, h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x)]$ be the true error and $\hat{R}(T_m, h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(x_i))$ the training error under the $0/1$ loss $\ell(y, y') = [\![y \neq y']\!]$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality

$$\underbrace{R(p,h)}_{\text{true error}} \leq \underbrace{\hat{R}(T_m,h)}_{\text{training error}} + \underbrace{\sqrt{\frac{\ln 2|\mathcal{H}| + \ln\frac{1}{\delta}}{2m}}}_{\text{complexity term}}$$

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

- 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. Use prior knowledge to limit the size of the hypothesis space $|\mathcal{H}|$.

Summary

- The empricial risk minimization: optimize performance on the trainin examples.
- The error decomposition: predictor's risk = estimation error + approximation error + Bayes error.
- ERM fails when the hypothesis class is excessively rich.
- If Uniform Law of Large Numbers applies for a hypothesis class $\mathcal{H}$, then the empirical risk for all members of $\mathcal{H}$ is a good proxy of the true risk.
- ULLN aplies for a finite hypothesis class.
- Generalization bound for a fininte hypothesis class.