Machine Learning Theory 2023
Lecture 2

Tim van Erven

Download these slides from elo.mastermath.nl!

- Review
- (Agnostic) PAC learning
- Agnostic PAC-learnability for finite classes
- Uniform convergence
- No-Free-Lunch Theorem (without proof)

Formal Setup Review

$$S = \begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_m \\ \boldsymbol{X}_m \end{pmatrix} \quad \sim \quad \mathcal{D}$$

Risk: $L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h, \boldsymbol{X}, Y)]$ for $(\boldsymbol{X}, Y) \sim \mathcal{D}$

Empirical Risk: $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, \boldsymbol{X}_i, Y_i)$ for $(\boldsymbol{X}_i, Y_i)$ in S

Classification (0/1-loss counts mistakes):

$$\ell(h, \boldsymbol{X}, Y) = \mathbf{1}\{h(\boldsymbol{X}) \neq Y\} = \begin{cases} 0 & \text{if } h(\boldsymbol{X}) = Y \\ 1 & \text{if } h(\boldsymbol{X}) \neq Y \end{cases}$$

Regression (Squared Error):

$$\ell(h, \boldsymbol{X}, Y) = (Y - h(\boldsymbol{X}))^2$$

No Overfitting for (Multiclass) Classification

Realizability assumption: Exists perfect predictor $h^* \in \mathcal{H}$, i.e. $\Pr(h^*(\boldsymbol{X}) = Y) = 1$.

Theorem (First Example of PAC-Learning)

Assume $\mathcal{H}$ is **finite**, **realizability** *holds. Choose any $\delta \in (0,1)$, $\epsilon > 0$. Then, for all $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$, ERM over $\mathcal{H}$ guarantees*

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta.$$

NB Lower bound on m does not depend on $\mathcal{D}$ or on h^*!

PAC learning: probably approximately correct

(Agnostic) PAC Learning

- PAC learning (always for binary classification)
- Agnostic PAC learning for binary classification
- Agnostic PAC learning in general
- Improper Agnostic PAC learning in general

Definition: PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Sample complexity:
The function $m_{\mathcal{H}}$ such that $m_{\mathcal{H}}(\epsilon, \delta)$ is smallest possible for all ϵ, δ

No Overfitting for (Multiclass) Classification

Theorem (First Example of PAC-Learning)

Assume $\mathcal{H}$ is **finite**, **realizability** *holds. Choose any $\delta \in (0, 1)$, $\epsilon > 0$. Then, for all $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$, ERM over $\mathcal{H}$ guarantees*

$$L_{\mathcal{D}}(h_S) \leq \epsilon$$

with probability at least $1 - \delta$.

For binary classification this is equivalent to:

Theorem

Every **finite** *hypothesis class $\mathcal{H}$ is PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

and learning algorithm ERM.

Definition: PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ for which **realizability** holds w.r.t. $\mathcal{H}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Agnostic PAC Learning (Binary Classification)

A hypothesis class $\mathcal{H}$ is **Agnostic PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ ~~for which **realizability** holds w.r.t. $\mathcal{H}$~~
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

Definition: Agnostic PAC Learning ~~(Binary Classification)~~ (In General)

A hypothesis class $\mathcal{H}$ is **Agnostic** **PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$ ~~for which **realizability** holds w.r.t. $\mathcal{H}$~~
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}(\epsilon, \delta).$$

Definition: Agnostic PAC Learning (In General)

A hypothesis class $\mathcal{H}$ is **Agnostic PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs $h_S \in \mathcal{H}$

such that for all

- distributions $\mathcal{D}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

Definition: Improper Agnostic PAC Learning (In General)

A hypothesis class $\mathcal{H}$ is **Improperly** **Agnostic PAC-learnable** if there exist

- a **function** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$
- and **learning algorithm** that outputs h_S ~~$\in \mathcal{H}$~~

such that for all

- distributions $\mathcal{D}$
- and all $\epsilon, \delta \in (0,1)$

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

Agnostic PAC-Learnability for Finite Classes via Uniform Convergence

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$*. Then every* **finite** *hypothesis class* $\mathcal{H}$ *is* **agnostically** *PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$*. Then every* **finite** *hypothesis class* $\mathcal{H}$ *is* **agnostically** *PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

- Worse dependence on ϵ compared to $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$ for PAC-learnability

Agnostic PAC-Learnability for Finite Classes

Theorem (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$. *Then every* **finite** *hypothesis class* $\mathcal{H}$ *is* **agnostically** *PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

- Worse dependence on ϵ compared to $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$ for PAC-learnability
- Losses with different range $[a, b]$ can be reduced to $[0, 1]$ range by subtracting a and dividing by $(b - a)$.

Technical Tool: Uniform Convergence

A hypothesis class $\mathcal{H}$ has the **uniform convergence** property if there exists

- a **function** $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$

such that for all

- distributions $\mathcal{D}$
- and all $\epsilon, \delta \in (0,1)$

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon \qquad \text{with probability} \ge 1 - \delta,$$

$$\text{whenever } m \ge m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta).$$

Uniform Convergence $\rightarrow$ Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

Lemma

If $\mathcal{H}$ has the **uniform convergence property**, *then it is* **agnostic PAC-learnable** *with*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

Uniform Convergence $\rightarrow$ Agnostic PAC-Learnability

Uniform convergence implies agnostic PAC-learnability:

Lemma

If $\mathcal{H}$ has the **uniform convergence property**, *then it is* **agnostic PAC-learnable** *with*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$

and learning algorithm ERM.

- We will prove uniform convergence for finite $\mathcal{H}$ and loss range $[0, 1]$
- Then the desired agnostic PAC-learnability follows

Proof (Handwritten)

To show, for h_S ERM hypothesis:

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}^{\mathrm{UC}}\left(\tfrac{\epsilon}{2}, \delta\right).$$

Assuming uniform convergence, applied for $\epsilon/2$:

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \tfrac{\epsilon}{2} \qquad \text{with probability} \geq 1 - \delta,$$

$$\text{whenever } m \geq m_{\mathcal{H}}^{\mathrm{UC}}\left(\tfrac{\epsilon}{2}, \delta\right).$$

Proof: On the event that $|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2}$ for all $h \in \mathcal{H}$, we have for all $h' \in \mathcal{H}$

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h') + \frac{\epsilon}{2} \leq L_D(h') + \epsilon.$$

Then take the infimum over h'.

Uniform Convergence for Finite Classes

Lemma (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$. *Then every* **finite** *hypothesis class* $\mathcal{H}$ *has the* **uniform convergence property** *with*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

To show:

$$\Pr\Big(\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \Big) \geq 1 - \delta$$

$$\text{whenever} \qquad m \geq \tfrac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

Proof (Handwritten)

$$\Pr\big(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon\big) \overset{?}{\ge} 1-\delta$$

$$\Pr\big(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\big) \overset{?}{\le} \delta$$

$$\Pr\big(\text{exists } h\in\mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\big) \overset{?}{\le} \delta$$

Part I (union bound):

$$\Pr\big(\text{exists } h\in\mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\big) \le \sum_{h\in\mathcal{H}} \Pr\big(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\big)$$

Part II (Hoeffding's inequality): Let $Z_i = \ell(h, \boldsymbol{X}_i, Y_i) \in [0,1]$.

$$\Pr\big(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\big) = \Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z]\right| > \epsilon\right) \overset{\text{Hoeffding}}{\le} 2e^{-2m\epsilon^2}$$

Proof Continued (Handwritten)

Part I+II:

$$\Pr\left(\text{exists } h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right) \leq \sum_{h \in \mathcal{H}} \Pr\left(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right)$$
$$\leq |\mathcal{H}| 2e^{-2m\epsilon^2} \overset{?}{\leq} \delta$$

Yes, for $m \geq \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2}$

Putting Everything Together

Theorem (Bounded Loss, Finite Class)

Suppose $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Then every **finite** *hypothesis class $\mathcal{H}$ has the uniform convergence property with*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\ln(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil,$$

and is therefore **agnostically** *PAC-learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\tfrac{\epsilon}{2}, \delta\right) \leq \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

and learning algorithm ERM.

No-Free-Lunch Theorem

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

Let A be **any learning algorithm** *for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

Let A be **any learning algorithm** *for binary classification. If $m \le |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \ge 1/8\right) \ge 1/7$ *for $S \sim \mathcal{D}^m$.*

Interpretation:

- $\mathcal{H}_{\text{all}}$ = all functions from $\mathcal{X}$ to $\{-1, +1\}$
- $m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$ for any $\epsilon < 1/8$, $\delta < 1/7$

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

Let A be **any learning algorithm** *for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

Interpretation:

- $\mathcal{H}_{\text{all}}$ = all functions from $\mathcal{X}$ to $\{-1, +1\}$
- $m_{\mathcal{H}_{\text{all}}}(\epsilon, \delta) > |\mathcal{X}|/2$ for any $\epsilon < 1/8$, $\delta < 1/7$

Corollary

Suppose $|\mathcal{X}| = \infty$. Then $\mathcal{H}_{all}$ is not PAC-learnable.

No-Free-Lunch Theorem (Binary Classification)

Is there a learner that works on all learning tasks? No!

Theorem (No-Free-Lunch)

Let A be **any learning algorithm** *for binary classification. If $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ such that*

1. *There exists a perfect predictor f with $L_{\mathcal{D}}(f) = 0$.*
2. $\Pr\left(L_{\mathcal{D}}(A(S)) \geq 1/8\right) \geq 1/7$ *for $S \sim \mathcal{D}^m$.*

Proof Intuition:

- Suppose $\mathcal{D}$ is uniform on $2m$ points in $\mathcal{X}$, and $Y = f(X)$ for some unknown function f.
- From S we only know $f(X)$ for m observed points.
- Without any assumptions about f, learner cannot do better than random guessing on m unobserved points.