PAC Learnability

Spring 2025

Outline

- The learner's input:
  - Domain set (Instances Space): An arbitrary set $\mathcal{X}$.
  - Domain point (Instance) : $x \in \mathcal{X}$.
  - Label set: $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{+1, -1\}$.
  - Training set: $S = \{(x_i, y_i)\}_{i=1}^{m}$, where every $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.
- The learner's output: $h : \mathcal{X} \to \mathcal{Y}$.
- A simple data-generation model: we assume that each pair in the training set S is generated by
  - first sampling a point x_i according to a fixed but unknown distribution $\mathcal{D}$ on $\mathcal{X}$,
  - and then labeling it by the ''correct'' labeling function f, that is, $y_i = f(x_i)$.

- Generalization error: a measure of success.

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}(h(x) \neq f(x)) = \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

- Training error: $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(h(x_i) \neq y_i)$
- Hypothesis class $\mathcal{H}$: A set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$.
- The $\mathrm{ERM}_{\mathcal{H}}$ Learner: for a given class $\mathcal{H}$, and a training set S, the $\mathrm{ERM}_{\mathcal{H}}$ learner uses the ERM rule to choose a predictor $h \in \mathcal{H}$, with the lowest possible error over S. Formally,

  $$\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h).$$

  We also use h_S to denote a result of applying $\mathrm{ERM}_{\mathcal{H}}$ to S, that is,

  $$h_S \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h).$$

Definition (The Realizability Assumption)

There exists $h^\star \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^\star) = 0$.

- This assumption implies that with probability 1, we have
  - $L_S(h^\star) = 0$.
  - $L_S(h_S) = 0$ for every ERM hypothesis h_S.

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0, 1)$ and $\epsilon > 0$ and let m be an integer that satisfies

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Then, for any labeling function, f, and for any distribution $\mathcal{D}$, for which the realizability assumption holds, with probability at least $1 - \delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, h_S, it holds that

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

- Notes: for a sufficiently large m, the $\mathrm{ERM}_{\mathcal{H}}$ rule over a finite hypothesis class will be Probably (with confidence $1 - \delta$) Approximately (up to an error of ϵ) Correct.

Proof. Let $\mathcal{H}_B$ be the set of "bad" hypotheses, that is,

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}.$$

Let $S|_x = \{x_1, \cdots, x_m\}$ be the instances of the training set. Then we upper bound the probability

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}).$$

In addition, let $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$. Note that

$$\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}.$$

Hence

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\bigcup_{h\in\mathcal{H}_B} \{S|_x : L_S(h) = 0\})$$
$$\leq \sum_{h\in\mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$$

Since the instances are sampled i.i.d., we get that

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) = \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}).$$

Note for every $h \in \mathcal{H}_B$,

$$\mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon, \text{ and}$$

$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \le (1-\epsilon)^m \le e^{-m\epsilon}$.
Therefore,

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \le |\mathcal{H}_B| e^{-m\epsilon} \le |\mathcal{H}| e^{-m\epsilon}.$$

Let

$$|\mathcal{H}| e^{-m\epsilon} \le \delta,$$

then

$$m \ge \frac{\log(|\mathcal{H}|/\delta)}{\epsilon},$$

and

$$1 - \mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \ge 1 - \delta. \ \Box$$

Outline

Definition (PAC Learnability)

A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm with the following property: For every $\delta, \epsilon \in (0,1)$, for every distribution over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \rightarrow \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}$, $\mathcal{D}$, f, then when running the algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples genenated by $\mathcal{D}$ and labeled by f, the algorithms returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples),

$$L_{(\mathcal{D},f)}(h) \leq \epsilon.$$

Probably Approximately Correct Learnability

- Approximately Correct: the accuracy parameter ϵ determines how far the output classifier can be from the optimal one.
- Probably: the confidence parameter δ indicates how likely the classifier is to meet that accuracy requirement.

- Sample complexity: How many samples are required to guarantee a probably approximately correct solution.
  - If $\mathcal{H}$ is PAC learnable, there are many functions m_H that satisfy the requirements given the definition of PAC learnability.
  - The sample complexity of learning $\mathcal{H}$ is defined as minimal function, in the sense that for any ϵ, δ, $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal integer that satisfies the requirements g of PAC learning with accuracy ϵ and confidence δ.

Corollary

Every finite hypothesis class is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil.$$

- Q: Does the finiteness determine the PAC learnability of a hypothesis class?
- A: No.

Outline

Empirical Risk Minimization

Probably Approximately Correct Learning

Agnostic PAC learnability

To waive the realizability assumption

- Recall that the realizability assumption requires that there exists $h^\star \in \mathcal{H}$ s.t. $L_{\mathcal{D},f}(h^\star) = 0$.
- For practical learning tasks, the realizability assumption may be too strong.
- From PAC learning to Agnostic PAC learning: releasing the realizability assumption.

A More Realistic Model for the Data-Generating Distribution

- From the deterministic case of a fixed but unknown distribution over $\mathcal{X}$ and a correct labeling function f to the stochastic case.
- Let $\mathcal{D}$ be a probability distribution over $\mathcal{X} \times \mathcal{Y}$.
- Two parts of such a distribution:
  - a marginal distribution $\mathcal{D}_x$ over unlabelled domain points.
  - a conditional probability $\mathcal{D}((x, y)|x)$ over labels for each point.

Generalization Error Revised:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = \mathcal{D}(\{(x, y) : h(x) \neq y\}).$$

- The Goal: to find some hypothesis, $h : \mathcal{X} \to \mathcal{Y}$, that (probably approximately) minimizes the generalization error, $L_{\mathcal{D}}(h)$.
- The Bayes Optimal Predictor: Given any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the best label predicting function from $\mathcal{X}$ to $\{0,1\}$ will be

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}.$$

- It is easy to verify that for every distribution $\mathcal{D}$,

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$$

for every classifier $g : \mathcal{X} \to \{0,1\}$.

- $\mathcal{D}$ is a fixed but unknown distribution.
- We cannot utilize the optimal predictor $f_{\mathcal{D}}$.
- Instead, we require that the learning algorithm will find a predictor whose error is not much larger than the best possible error of a predictor in some given benchmark hypothesis class.

Definition (Agnostic PAC Learnability)

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\delta, \epsilon \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, then when running the algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples genenated by $\mathcal{D}$, the algorithms returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

- Agnostic PAC learning generalizes the definition of PAC learning.
  - If the realizability assumption holds, agnostic PAC learning provides the same guarantee as PAC learning.
- When the realizability assumption does not hold, no learner can guarantee an arbitrarily small error.
- Under the definition agnostic PAC learning, a learner can still declare success if its error is not much larger than the best error achievable by a predictor from the hypothesis class $\mathcal{H}$.

- Generalized loss functions :
  - Given any set $\mathcal{H}$ and some domain Z, let ℓ be any function from $\mathcal{H} \times Z$ to the set of nonnegative real numbers, $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$.
  - We call such functions loss functions.
  - For prediction tasks, $Z = \mathcal{X} \times \mathcal{Y}$.
    - 0-1 loss:

$$\ell_{0-1}(h, (x, y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases} .$$

    - Square loss: $\ell_{sq}(h, (x, y)) = (h(x) - y)^2$.

- Risk function: the expected loss of a classifier $h \in \mathcal{H}$ with respect to A distribution $\mathcal{D}$ over the domain set Z:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)].$$

- Empirical Risk: the expected loss of a classifier $h \in \mathcal{H}$ over a given a sample $S = (z_1, z_2, \cdots, z_m) \in Z^m$:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i).$$

Definition (Agnostic PAC Learnability for General Loss Functions)

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\delta, \epsilon \in (0, 1)$, and for every distribution $\mathcal{D}$ over Z, then when running the algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples genenated by $\mathcal{D}$, the algorithms returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

Outline

Empirical Risk Minimization

Probably Approximately Correct Learning

Agnostic PAC learnability

Uniform Convergence

Definition (ϵ-representative sample)

A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

Lemma

Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l, and distribution $\mathcal{D}$). Then any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely, any $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Lemma

Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l, and distribution $\mathcal{D}$). Then any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely, any $h_S \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Proof.
For every $h \in \mathcal{H}$,

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\epsilon}{2} && (S \text{ is } \epsilon - \text{representative.}) \\ &\leq L_S(h) + \frac{\epsilon}{2} && (h_S \text{ is an ERM predictor.}) \\ &\leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} && (S \text{ is } \epsilon - \text{representative.}) \\ &= L_{\mathcal{D}}(h) + \epsilon. \end{aligned}$$

□

Definition (Uniform Convergence)

We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property*(w.r.t. domain Z and loss function l) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability at least $1 - \delta$, S is ϵ-representative.

Corollary

If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$. Furthermore, in that case, the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

Corollary

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

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

Theorem (Hoeffding's Inequality)

Let $\theta_1, \cdots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \epsilon \right] \leq 2\exp(-2m\epsilon^2/(b-a)^2).$$

Proof.

Fix some $\epsilon, \delta \in (0,1)$. We need to find a sample size m that guarantees that for any $\mathcal{D}$, with probability of at least $1-\delta$ of the choice of $S = (z_1, \cdots, z_m)$ sampled i.i.d. from $\mathcal{D}$ we have that for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon$. That is,

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon\}) \ge 1 - \delta.$$

Equivalently, we need to show that

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

Notice that

$$\begin{aligned} & \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\ \le & \sum_{h \in \mathcal{H}} \mathcal{D}^m(S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \end{aligned}$$

□

Applying Hoeffding's inequality, then we obtain that

$$\mathcal{D}^m(S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \leq 2\exp(-2m\epsilon^2).$$

Hence

$$\begin{aligned} & \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\ \leq & \sum_{h \in \mathcal{H}} \mathcal{D}^m(S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \\ \leq & 2|\mathcal{H}| \exp(-2m\epsilon^2). \end{aligned}$$

Finally, if we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2},$$

then

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta. \square$$