Aalto University
School of Science

CS-E4070 — Computational learning theory

Slide set 03 : agnostic PAC learning and uniform convergence

Cigdem Aslay and Aris Gionis

Aalto University

spring 2019

reading material

- SS&BD, chapters 3, 4, and 5

what we have seen so far

- $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}$ is sampled from $\mathcal{D}$, and $y = c(\mathbf{x})$ labeled by the target concept $c : X \to Y$ that we want to learn
- the learner observes sample set S and outputs hypothesis $h : X \to Y$ for predicting the label of unseen data points drawn from $\mathcal{D}$.
- the error of the learner is defined as the probability that the learner does not predict the correct label on a random data point sampled from $\mathcal{D}$

$$error_{\mathcal{D}}(h) = \mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x}) \neq c(\mathbf{x})]$$

what we have assumed so far

- learning task: learning from examples with binary labels
- example generation: the sample data are drawn from $\mathcal{D}$ and labeled according to a target concept $c \in \mathcal{C}$
- realizability assumption: there exists a hypothesis $h^* \in \mathcal{H}$ such that $error_{\mathcal{D}}(h^*) = 0$
- concept class $\mathcal{C}$ is finite or can efficiently be discretized

relaxing the realizability assumption

- realizability assumption: there exists a hypothesis $h^* \in \mathcal{H}$ such that $error_{\mathcal{D}}(h^*) = 0$ (with probability 1)
  - requires that labels are fully determined by the features we measure on input elements
  - e.g., papayas with same color and softness will have the same taste
- in many practical problems this assumption does not hold
- so how do we remove the realizability assumption?

relaxing the realizability assumption

- sampling process under realizability assumption for an example $(\mathbf{x}, y) \in S$:
  - $\mathbf{x}$ is sampled from $\mathcal{D}$
  - $y = c(\mathbf{x})$ labeled by the target concept $c : X \to Y$
- unrealizable setting: modify the sampling process to allow for noise
- replace the target concept labeling with a data-labels generating distribution
  - define the sampling distribution $\mathcal{D}$ to be a joint distribution over $X \times Y$

relaxing the realizability assumption

- we can view $\mathcal{D}$ $(\mathbf{x}, y)$ as product of two distributions
  - the marginal distribution $\mathcal{D}_{\mathbf{x}}$ over unlabeled data $\mathbf{x}$
  - the conditional distribution $\mathcal{D}_{y|\mathbf{x}} = \mathcal{D}((\mathbf{x}, y) \mid \mathbf{x})$ over labels for each data $\mathbf{x}$
- the conditional distribution $\mathcal{D}((\mathbf{x}, y) \mid \mathbf{x})$ over labels introduces noise
  - the same example can have different labels in different draws
- generalization error can be redefined as

$$error_{\mathcal{D}}(h) = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}\,[\,h(\mathbf{x}) \neq y\,] = \mathcal{D}(\,\{(\mathbf{x}, y) \mid h(\mathbf{x}) \neq y\}\,)$$

optimal Bayes hypothesis

- given any probability distribution $\mathcal{D}$ over $X \times \{0, 1\}$, the best hypothesis we can hope for is $b : X \to Y$, s.t.

$$b(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{Pr}_{\mathcal{D}_{y|\mathbf{x}}}[y = 1 \mid \mathbf{x}] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

- for any other hypothesis h and for any distribution $\mathcal{D}$

$$error_{\mathcal{D}}(b) \leq error_{\mathcal{D}}(h)$$

- learner does not have access to distribution $\mathcal{D}$, so we cannot find the optimal Bayes hypothesis
- but learner has access to sample set S drawn from $\mathcal{D}$

agnostic PAC learning

- extension of PAC learning to unrealizable setting
- learner is agnostic to the data-labels distribution
  - no assumption on $\mathcal{D}$
  - no learner can guarantee an arbitrarily small error
- in contrast to PAC learning, the learner is not required to achieve a small error in absolute terms, but relative to the minimum possible error achievable by the hypothesis class

agnostic PAC learning

- learner can declare success if the generalization error is not much larger than the smallest error achievable by a hypothesis from $\mathcal{H}$
- approximately correct criterion: we want to find an h such that

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

- if the realizability assumption holds, agnostic PAC learning provides the same guarantees as in PAC learning

agnostic PAC learnability

- **definition** (agnostic PAC learning):

  a hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

  for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $X \times Y$, when running A on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, A returns a hypothesis h that satisfies

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

  with probability at least $1 - \delta$ (over the choice of examples).

scope of learning problems

- so far we have focused on examples with binary labels
- formalization can be generalized to other types of learning from examples
- regression: find a linear function that best predicts a baby's birth from ultrasound measures of his head circumference, abdominal circumference, and femur length

  X: possible values of ultrasound measurements, set of triplets in $\mathbb{R}^3$

  Y: possible values of weight at birth, $\mathbb{R}$

scope of learning problems

- given $\mathcal{H}$ and domain $X \times Y$, a loss function $\ell : \mathcal{H} \times (X \times Y) \to \mathbb{R}_+$ quantifies how good h is on $(\mathbf{x}, y)$
- $error_{\mathcal{D}}(h)$ is the expected loss of hypothesis h with respect to distribution $\mathcal{D}$ over $X \times Y$

$$error_{\mathcal{D}}(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h, (\mathbf{x}, y))\right]$$

- $error_S(h)$ is the empirical loss over a given sample S

$$error_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, (\mathbf{x}, y))$$

example loss functions

- 0-1 loss:

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

- square loss:

$$\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$$

- absolute value loss:

$$\ell(h, (\mathbf{x}, y)) = |h(\mathbf{x}) - y|$$

learnability for general loss functions

- **definition** (agnostic PAC learning):

  a hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a domain $X \times Y$ and a loss function $\ell : \mathcal{H} \times (X \times Y) \to \mathbb{R}_+$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

  for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $X \times Y$, when running A on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, A returns a hypothesis h such that with probability at least $1 - \delta$

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

  where $error_{\mathcal{D}}(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$

learning via uniform convergence

- the definition(s) of (agnostic) PAC learning states when we can learn something
- it does not provide much information about what and how we can learn
- how well we can learn a hypothesis from a sample depends on the quality of that sample
- a sample has good quality when the estimated error of any hypothesis on the sample is close to its true error

learning via uniform convergence

- remember the empirical risk minimization rule $ERM_{\mathcal{H}}(S)$
  - given a sample set S of m examples, return the hypothesis h_S from finite $\mathcal{H}$ such that

$$h_S = \arg\min_{h \in \mathcal{H}} error_S(h)$$

- under the realizability assumption we have

  $error_S(h_S) = 0$, and

  $\mathbf{Pr}\left[error_{\mathcal{D}}(h_S) \leq \epsilon\right] \geq 1 - \delta$ when $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$

- what about the unrealizable setting?

learning via uniform convergence

- recall that $error_{\mathcal{D}}(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$
- if we can ensure that empirical risks of all members of $\mathcal{H}$ are good approximations of their true error, $ERM_{\mathcal{H}}(S)$ can return a hypothesis h that has error close to minimum possible error
- in other words, we want to obtain, uniformly over all members of $\mathcal{H}$, an empirical risk that is close to its expectation

learning via uniform convergence

- ϵ-representative sample:

  a sample set S is ϵ-representative with respect to a domain $X \times Y$, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$ if

$$\forall h \in \mathcal{H}, |error_{\mathcal{D}}(h) - error_S(h)| \leq \epsilon$$

learning via uniform convergence

- lemma: assume that a sample set S is $\epsilon/2$-representative, then any output h_S of $ERM_{\mathcal{H}}(S)$ satisfies

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

- proof: for every $h \in \mathcal{H}$ we have

$$\begin{aligned} error_{\mathcal{D}}(h_S) &\leq error_S(h_S) + \frac{\epsilon}{2} \\ &\leq error_S(h) + \frac{\epsilon}{2} \\ &\leq error_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &\leq error_{\mathcal{D}}(h) + \epsilon \end{aligned}$$

learning via uniform convergence

- to ensure that $ERM_{\mathcal{H}}(S)$ is an agnostic PAC learner, it is sufficient to have an ϵ-representative sample with probability at least $1 - \delta$ (over the randomness of S)
- uniform convergence formalizes this sufficiency condition

learning via uniform convergence

- uniform convergence: a hypothesis class $\mathcal{H}$ has the uniform convergence property with respect to domain $X \times Y$ 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 distribution $\mathcal{D}$ over $X \times Y$, a sample S of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ i.i.d. examples drawn from $\mathcal{D}$ is ϵ-representative with probability at least $1 - \delta$.

- the term uniform refers to the fact that the (minimal) sample complexity $m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ is the same for all hypothesis in $\mathcal{H}$ and all probability distributions $\mathcal{D}$.

learning via uniform convergence

- to prove that we can agnostic PAC learn a hypothesis class, just prove that it has the uniform convergence property
- corollary: if $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC} : (0, 1)^2 \to \mathbb{N}$, then $\mathcal{H}$ is agnostic PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.

  Furthermore, in that case, $ERM_{\mathcal{H}}(S)$ is a successful agnostic PAC learner for $\mathcal{H}$.

finite classes are agnostic PAC learnable

- theorem: let $\mathcal{H}$ be a finite hypothesis class and let $\ell : \mathcal{H} \times (X \times Y) \to [a, b]$ be a bounded loss function. Then $\mathcal{H}$ is agnostic PAC learnable using $ERM_{\mathcal{H}}(S)$ with sample complexity

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

Hoeffding's inequality

- let $\theta_1, \cdots, \theta_m$ be a sequence of i.i.d. random variables and assume that $\forall i, \mathbf{E}[\theta_i] = \mu$ and $\mathbf{Pr}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon \geq 0$,

$$\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

finite classes are agnostic PAC learnable

- proof: it suffices to show that $\mathcal{H}$ has the uniform convergence property with

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2(b-a)^2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

- so we need to find $m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$ for fixed ϵ and δ such that for any distribution $\mathcal{D}$, an i.i.d. sample S of $m \geq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$

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

finite classes are agnostic PAC learnable

- proof cont'd: from union bound, we have

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\})$$
$$\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\})$$

- so if we can prove that for a large enough m each

$$\mathcal{D}^m(\{S : |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\})$$

is small enough, result follows.

finite classes are agnostic PAC learnable

- proof cont'd: we know that

$$error_{\mathcal{D}}(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$$

- using Hoeffding's inequality we have

$$\mathcal{D}^m(\{S : |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\}) \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

- which implies

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\}) \leq 2|\mathcal{H}|e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

- so if $m \geq \frac{2(b-a)^2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2}$, then the RHS is at most δ as required

finite classes are agnostic PAC learnable

- proof cont'd: we know that

$$error_{\mathcal{D}}(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$$

- using Hoeffding's inequality we have

$$\mathcal{D}^m(\{S : |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\}) \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

- which implies

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |error_{\mathcal{D}}(h) - error_S(h)| > \epsilon\}) \leq 2|\mathcal{H}|e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

- so if $m \geq \frac{2(b-a)^2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2}$, then the RHS is at most δ as required

discussion of sample complexity

- we started with realizability assumption and 0-1 loss and obtained
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$
- by relaxing the realizability assumption and assuming general loss functions, we ended up with
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2(b-a)^2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$
- for the same level of accuracy, sample complexity grows by a factor of $1/\epsilon$
- contribution of a general loss function is smaller ($[a, b]$ can often be normalized to $[0, 1]$)

the discretization trick

- allows to get a good estimate of practical sample complexity of infinite hypothesis classes
- consider the class of signum functions: $X = \mathbb{R}$ and $Y = \{+1, -1\}$.
- let $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$ where $h_\theta = sign(\mathbf{x} - \theta)$
- each h_θ is parametrized by one parameter, $\theta \in \mathbb{R}$ and outputs -1 for instances smaller than θ

the discretization trick

- $\mathcal{H}$ is infinite but in practice we only need 64 bits to maintain a real number using floating point representation
- so $\mathcal{H}$ is parametrized by set of scalars represented using a 64 bits floating point number
- there are at most 2^{64} such numbers hence actual size of $\mathcal{H}$ is at most 2^{64}
- so sample complexity of $\mathcal{H}$ is bounded by

$$\frac{128 + 2\log(2/\delta)}{\epsilon^2}$$

- practical estimate but dependent on machine-specific representation of $\mathbb{R}$

we have seen that

- finite classes are PAC learnable with sample complexity

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

- finite classes are agnostic PAC learnable with sample complexity

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

- discretization trick can allow to obtain a practical estimate of the sample complexity for infinite classes

  e.g., class of signum functions

application : no-free-lunch theorem

SS&BD, chapter 5

- we can show that there is no universal learner
  - some form of prior knowledge is necessary
  - we should know something about $\mathcal{D}$ and/or $\mathcal{C}$

- **theorem** (no-free-lunch) : let A be a learner over X.
  Then there exists a distribution $\mathcal{D}$ over $X \times \{0, 1\}$ such that
  1. there exists concept $c : X \to \{0, 1\}$ with $error_{\mathcal{D}}(c) = 0$
  2. with probability at least $1/7$ over $S \sim \mathcal{D}^m$ we have that $error_{\mathcal{D}}(A(S)) \geq 1/8$

- **corollary** : let $\mathcal{C}$ be the set of all mappings from an infinite domain X to $\{0, 1\}$. Then, $\mathcal{C}$ is not PAC learnable.

no-free-lunch theorem

- no-free-lunch theorem: without restricting the hypothesis class, for any learning algorithm, an adversary can construct a distribution for which the learning algorithm will perform poorly, while there is another algorithm that will succeed in the same distribution
- corollary: let $\mathcal{C}$ be the set of all mappings from an infinite domain X to $\{0, 1\}$. Then, $\mathcal{C}$ is not PAC learnable.
- so an infinite class with rich representation cannot be (agnostic) PAC learned
- so how do we learn an infinite hypothesis class $\mathcal{H}$?

learning threshold functions

- lemma: let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ be the set of threshold functions over the real line where $\forall h_a \in \mathcal{H}$

$$h_a : \mathbb{R} \to \{0, 1\}, h_a(\mathbf{x}) = \mathbb{I}[\mathbf{x} \leq a]$$

- $\mathcal{H}$ is PAC learnable using the ERM rule with sample complexity

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

learning threshold functions

- $\mathcal{H}$ is of infinite size
- we want to get close to the true threshold value
  we just need to prove that for any $\mathcal{D}$, ERM rule will probably get us close
- we know that all values to the left are classified as negative, all values to the right are classified as positive

proof (sketch)

- let a^* be the true value and define $a_1, a_2 \in \mathbb{R}$ such that

$$\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [\mathbf{x} \in (a_1, a^*)] = \mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [\mathbf{x} \in (a^*, a_2)] = \epsilon$$

- we want to prove that we most likely get an example from this interval
- given a sample S,
  - let $b_1 = max\{\mathbf{x} : (\mathbf{x}, 1) \in S\}$,
  - let $b_2 = min\{\mathbf{x} : (\mathbf{x}, 0) \in S\}$, and
  - let b_S denote the threshold of ERM hypothesis h_S which implies $b_S \in (b_1, b_2)$

proof (sketch)

- a sufficient condition for $error_{\mathcal{D}}(h_S) \leq \epsilon$ is to have $b_1 \geq a_1$ and $b_2 \leq a_2$

$$\mathbf{Pr}_{S \sim \mathcal{D}^m}[error_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbf{Pr}_{S \sim \mathcal{D}^m}[b_1 < a_1] + \mathbf{Pr}_{S \sim \mathcal{D}^m}[b_2 > a_2]$$

- the event $b_1 < a_1$ happens iff there exists no $\mathbf{x} \in S$ such that $\mathbf{x} \in (a_1, a^*)$

$$\mathbf{Pr}_{S \sim \mathcal{D}^m}[b_1 < a_1] = (1 - \epsilon)^m \leq e^{-\epsilon m} \leq \delta/2.$$

free lunch vs threshold functions

- so finiteness of $\mathcal{H}$ is a sufficient condition for PAC learnability, but not a necessary condition
- does learnability of threshold functions contradict the no-free-lunch theorem?

free lunch vs threshold functions

- the class of threshold functions is so simple that an adversary has no room to create an adversarial distribution
- if two threshold functions agree on a large enough sample, their respective thresholds will be close to each other
- there is no way you can force them to behave differently on unseen examples
- so a necessary condition for PAC learnability is that $\mathcal{H}$ should not be too expressive?

how expressive $\mathcal{H}$ should be?

- consider binary classification: $h : X \to \{0, 1\}$
- expressiveness of $\mathcal{H}$ is a measure of how many functions it can express
- from the corollary of no-free-lunch theorem, we should consider not only functions on X but also functions on (finite) subsets of X

the Vapnik-Chervonenkis dimension theory

- developed during 1960 – 1990 by Vladimir Vapnik and Alexey Chervonenkis
- provides a combinatorial measure to quantify the bias of the hypothesis class
- main idea: do not measure the size of the hypothesis class but the number of distinct instances that can be completely discriminated using $\mathcal{H}$