Advanced Machine Learning

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

University of Bucharest, 2nd semester, 2020-2021

Administrative

- seminar 2 class this week, 5 exercises

- seminar 2 also next next week (Thursday + Friday)

PAC vs. Agnostic PAC learning

	PAC	Agnostic PAC
Distribution	$\mathcal{D}$ over $\mathcal{X}$	$\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
Truth	$f \in \mathcal{H}$	not in class or doesn't exist
Risk	$L_{\mathcal{D},f}(h) =$ $\mathcal{D}(\{x : h(x) \neq f(x)\})$	$L_{\mathcal{D}}(h) =$ $\mathcal{D}(\{(x, y) : h(x) \neq y\})$
Training set	$(x_1, \ldots, x_m) \sim \mathcal{D}^m$ $\forall i,\ y_i = f(x_i)$	$((x_1, y_1), \ldots, (x_m, y_m)) \sim \mathcal{D}^m$
Goal	$L_{\mathcal{D},f}(A(S)) \leq \epsilon$	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

The Bayes optimal predictor

- given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the best label prediction function we can achieve is the Bayes rule:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases} \Leftrightarrow \mathcal{D}((x,1)|x) \geq ½$$

- for any probability distribution $\mathcal{D}$, the Bayes predictor $f_{\mathcal{D}}$ is optimal, in the sense that no other classifier g: $\mathcal{X} \rightarrow \{0,1\}$has a lower error, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$ (seminar exercise)

- we don't know the probability distribution $\mathcal{D}$ that produces the data (x, y), we only see a sample S generated by $\mathcal{D}$

- so, we cannot utilize the Bayes optimal predictor $f_{\mathcal{D}}$

Loss functions

- let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- given hypothesis $h \in \mathcal{H}$ and an example z = $(x,y) \in \mathcal{Z}$, how good is h on (x,y)?
- loss function $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow R_+$
  - measures the error that model h does it on the instance $z = (x,y)$
  - the true risk (generalization error) of model h is: $L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \underset{z \sim \mathcal{D}}{\mathbb{E}}[\ell(h, z)]$
- example of other loss functions:

Squared loss: $\ell(h, (x, y)) = (h(x) - y)^2$
Absolute-value loss: $\ell(h, (x, y)) = |h(x) - y|$
Cost-sensitive loss: $\ell(h, (x, y)) = C_{h(x),y}$ where C is some $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix

Today’s lecture: Overview

- The general PAC learning definition (agnostic PAC)
- Uniform convergence
- The No-Free-Lunch theorem

The general PAC learning problem

- we wish to Probably Approximately solve:

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \quad \text{where} \quad L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(h, z)]$$

- learner knows $\mathcal{H}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and loss function ℓ
- learner receives accuracy parameter ε and confidence parameter δ
- learner can decide on training set size m based on ε, δ
- learner doesn't know $\mathcal{D}$ but can sample S from $\mathcal{D}^m$
- using S the learner outputs some hypothesis $A(S) = h_S$
- we want that with probability at least 1 - δ over the choice of S, the following would hold:

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

Formal definition

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

- for every $\varepsilon > 0$ (*accuracy* → *"approximately correct"*)
- for every $\delta > 0$ (*confidence* → *"probably"*)
- for every distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

when we run the learning algorithm A on a training set S, consisting of $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from $\mathcal{D}$ the algorithm A returns a hypothesis A(S) from $\mathcal{H}$ such that, with probability at least $1-\delta$ (over the choice of examples) it holds that:

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

- if the realizability assumption holds, agnostic PAC = PAC
- in 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 class $\mathcal{H}$.

Agnostic PAC learnability of a class $\mathcal{H}$

A hypothesis class $\mathcal{H}$ is called ***agnostic PAC learnable*** if:

There exists a learning algorithm A with the property that given enough samples $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ drawn i.i.d. from $\mathcal{D}$, with probability $1 - \delta$ it will return a hypothesis $h_S = A(S)$ from $\mathcal{H}$ that has an error smaller than ε wrt the best achievable error by a predictor from the class $\mathcal{H}$:

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

$$\underset{S\sim D^m}{P}(L_D(h_S) \leq \min_{h\in H} L_D(h) + \varepsilon) \geq 1-\delta \Leftrightarrow \underset{S\sim D^m}{P}(L_D(h_S) > \min_{h\in H} L_D(h) + \varepsilon) < \delta$$

Agnostic PAC learnability of a class $\mathcal{H}$

A hypothesis class $\mathcal{H}$ is called ***agnostic PAC learnable*** if:

I can find a hypothesis h from $\mathcal{H}$ based on the learning algorithm A with

- whatever accuracy $\varepsilon > 0$ wrt the best achievable error by a predictor in $\mathcal{H}$ I want
- whatever confidence $\delta > 0$ I want
- whatever the distribution $\mathcal{D}$ is

given that I provide to A enough samples m ≥ $m_{\mathcal{H}}(\varepsilon, \delta)$ drawn from $\mathcal{D}$ such that:

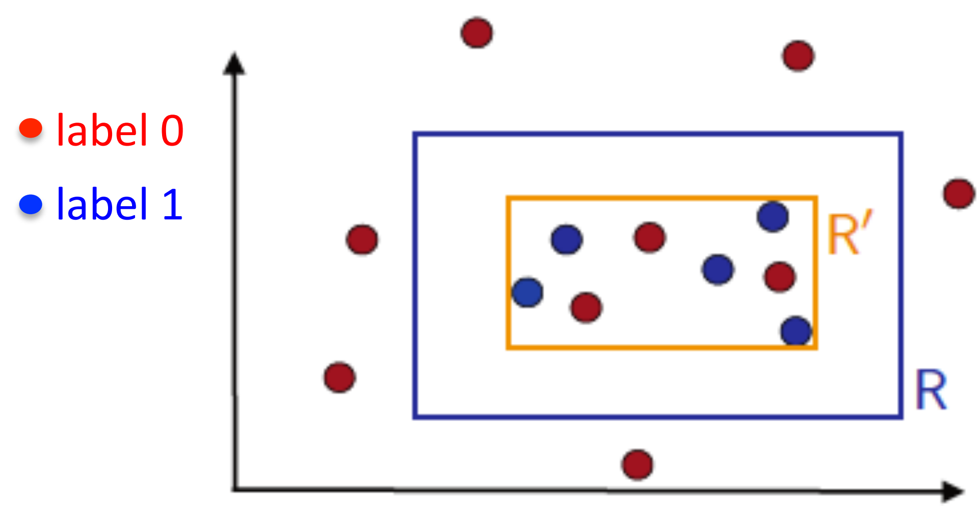
$$\underset{S \sim D^m}{P}\left(L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon\right) \geq 1 - \delta$$

Learning in the presence of noise - rectangles

- $\mathcal{X} = R^2$ points in the plane
- $\mathcal{H}$ = set of all axis-aligned rectangle lying in R^2
- each concept $h \in \mathcal{H}$ is an indicator function of a rectangle
- the learning problem consists of determining with small error a target axis-aligned rectangle using the labeled training sample
- the training points received by the learner are subject to noise:
  - points negatively labeled are unaffected by noise
  - the label of a positive training points is randomly flipped to negative with probability $0 < \eta < ½$ (η is unknown)

$\mathcal{H}$ is agnostic PAC learnable

$\min_h L_{\mathcal{D}}(h) = \eta \times \mathcal{D}(R)$



A note of Caution

The fact that $\mathcal{H}$ is agnostically PAC learnable using the ERM paradigm doesn't mean that the result is any good.

It only means that you can be reasonable sure the ERM paradigm gives you a result that is close to the optimal result.

If the optimal result is bad (because, for example, the hypothesis class $\mathcal{H}$ fits the data really badly) the ERM paradigm will also give you a bad result.

PAC doesn't tell you that your hypothesis class $\mathcal{H}$ fits the data well, it only tells you that, if it fits well, the ERM paradigm will probably give you a reasonable good hypothesis.

Beyond the general PAC learning definition

- the definition of the general PAC learning tells us:
  - when we consider we can learn something
- the definition of the general PAC learning doesn’t tell us:
  - what we can learn
  - how we learn
- discover what can be general PAC-learned and how

Uniform Convergence

Sufficient learning condition for agnostic PAC learnability

- given $\mathcal{H}$, the $\text{ERM}_{\mathcal{H}}$ learning paradigm works as follows:
  - based on a received training sample S of examples draw i.i.d from an unknown distribution $\mathcal{D}$ over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\text{ERM}_{\mathcal{H}}$ evaluates the risk (error) of each h in $\mathcal{H}$ on S and outputs a member $h_S = \text{ERM}_{\mathcal{H}}(S)$ that minimizes the empirical error $L_S(h_S)$;
  - we want that h_S will generalize wrt true data probability distribution $\mathcal{D}$, i.e $L_{\mathcal{D}}(h_S)$ is small;
  - it suffices to ensure that the empirical risks of all h in $\mathcal{H}$ are good approximations of their true risk

- we need that *uniformly* over all hypothesis h in the hypothesis class $\mathcal{H}$, the empirical risk based on S will be close to true risk for all possible probability distributions $\mathcal{D}$ over the domain $\mathcal{Z}$

ε - Representative

- how well you can learn a hypothesis depends on the quality of that sample:
  - you can't learn anything from a bad sample
  - a bad sample will make a bad hypothesis to look good and a good one to look bad

- when is a sample good?
  - a sample is good if the estimated quality (the loss) of a hypothesis on that sample is very close to its true error

Definition (ε – representative sample)

A sample S is called ε – representative wrt domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$ if:

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

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(h, z)] \qquad L_S(h) = \frac{1}{m} \sum_{z \in S} l(h, z)$$

ε – Representative Samples are Good

Lemma

Let S be a sample that is $\varepsilon/2$ – representative wrt domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$. Then any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ i.e any $h_S \in \mathrm{argmin}_h\, L_S(h)$ satisfies:

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

Proof

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \varepsilon/2 \leq \min_h\, L_S(h) + \varepsilon/2 \leq \min_h\, L_{\mathcal{D}}(h) + \varepsilon/2 + \varepsilon/2$$

S is $\varepsilon/2$ – representative sample

Uniform convergence

If ε-representative samples allows us to learn as good as possible, we can agnostically PAC learn if we can guarantee that we will almost always get (with probability 1 – δ) ε-representative sample.

Definition (*uniform convergence*)

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* wrt a domain $\mathcal{Z}$, loss function ℓ if:

- there exists a function $m_H^{UC} : (0,1)^2 \to \mathrm{N}$
- such that for all $(\varepsilon, \delta) \in (0,1)^2$
- and for any probability distribution $\mathcal{D}$ over $\mathcal{Z}$

if S is a sample of $m \geq m_H^{UC}(\varepsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least 1 – δ, S is ε-representative.

The term *uniform* refers to having a fixed sample size that works for all members of $\mathcal{H}$ and over all possible probability distributions $\mathcal{D}$ over the domain $\mathcal{Z}$

A tool to prove PAC learnability

- uniform converges serves as a tool to prove that we can PAC learn a hypothesis class $\mathcal{H}$

Corollary

If hypothesis class $\mathcal{H}$ has the uniform convergence property with function m_H^{UC} then $\mathcal{H}$ is agnostically PAC learnable with the sample complexity:

$$m_H(\varepsilon,\delta) \leq m_H^{UC}(\varepsilon/2,\delta)$$

Moreover, the $\text{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

Finite classes are agnostic PAC learnable

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

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

Moreover, the class $\mathcal{H}$ is agnostically PAC learnable using the ERM paradigm with sample complexity:

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

Proof - Finite classes are agnostic PAC learnable

- uniform converges serves as a tool to prove that we can PAC learn a hypothesis class $\mathcal{H}$

- to prove that finite hypothesis classes have the uniform convergence property, we need to:
  - for fixed ε and δ
  - find a sample size *m*
  - such that for any distribution $\mathcal{D}$ over $\mathcal{Z}$
  - and a sample $S = (z_1, z_2, \ldots, z_m)$ of examples i.i.d from $\mathcal{D}$
  - with probability at least 1- δ
  - it holds that for all h $\in \mathcal{H}$ $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$

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

$$\Updownarrow$$

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

Proof - union bound

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\},$$

Use the union bound to obtain:

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

For a sufficiently large m, each summand of the right-hand side of this inequality is small enough.

Show that for any fixed hypothesis h (which is chosen in advance prior to the sampling of the training set), the gap between the true and empirical risks, $|L_S(h) - L_{\mathcal{D}}(h)|$, is likely to be small.

Proof - Hoeffding's inequality

Lemma (Hoeffding's Inequality). *Let $\theta_1,\ldots,\theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i]=\mu$ and $\mathbb{P}[a\le\theta_i\le b]=1$. Then, for any $\epsilon>0$*

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

Apply in our case by setting:

$$\theta_i = l(h,z_i) \quad L_S(h)=\frac{1}{m}\sum_{z\in S} l(h,z)=\frac{1}{m}\sum_{i}^{m}\theta_i \quad L_D(h)=\mu \qquad \text{a = 0, b = 1}$$

Then, we have:

$$\mathcal{D}^m(\{S:|L_S(h)-L_{\mathcal{D}}(h)|>\epsilon\})=\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i-\mu\right|>\epsilon\right]\le 2\exp\left(-2m\epsilon^2\right)$$

Proof - final step

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} 2\exp\left(-2m\epsilon^2\right)$$
$$= 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)$$

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

Then, we have:

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

Beyond the result

By going from realizability to agnostic, we go:

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

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

The denominator goes from ε to ε^2, which means that for the same of accuracy the minimal sample size grows by a factor of 1/ε.

The No-Free-Lunch theorem

Prior knowledge

Empirical Risk Minimization (ERM) = learning paradigm that returns a predictor h that minimizes $L_S(h)$, S –training sequence of examples sampled i.i.d. from an unknown distribution $\mathcal{D}$ over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

- ERM might overfit if we are not careful

To guard against overfitting we introduced some prior knowledge (inductive bias)

- hypothesis class = $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$
- revised ERM rule: apply the ERM learning paradigm over $\mathcal{H}$
- for the training sample S, the $\mathrm{ERM}_{\mathcal{H}}$ learner chooses a predictor $h \in \mathcal{H}$ with the lowest possible error over S:

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

Universal learner?

Do we need prior knowledge ($\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$) for the success of learning?
Consider $\mathcal{H}$ the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$, $\mathcal{H} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$. This class represents lack of prior knowledge: every member of it is a good candidate.

Maybe there exists some kind of universal learner = a learner who has no prior knowledge about a certain task and is ready to be challenged by any task. A specific task is defined by an unknown distirbution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where the goal of the learner is to find a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ whose risk $L_{\mathcal{D}}(h)$ is small enough.

Does there exists a learning algorithm A and a training set size *m* such that for every distribution $\mathcal{D}$, if A receives *m* i.i.d. samples from $\mathcal{D}$ it will output with high confidence a predictor *h* that has a small error?

The No-Free-Lunch theorem states that no such universal learner exists. For binary classification prediction tasks ($\mathcal{Y} = \{0,1\}$) for every learner there exists a distribution on which it fails (the learner will output a hypothesis with large generalization error)

The No-Free-Lunch theorem

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification with respect to the 0−1 loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size.

Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. there exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$.
2. with probability of at least 1/7 over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.

- *In other words, for every learning algorithm A there are cases for which this algorithm will fail whereas there is another learner (e.g. a trivial successful learner in this case would be an ERM learner with the hypothesis class $\mathcal{H} = \{f\}$, or more generally, ERM with respect to any finite hypothesis class that contains f and whose size satisfies the equation $m \geq 8\log(7|H|/6)$ that solves the task. It simply means that an adversary can use the fact that A has no clue what happens on the other half of the domain. We cannot learn perfectly without the proper background knowledge.*