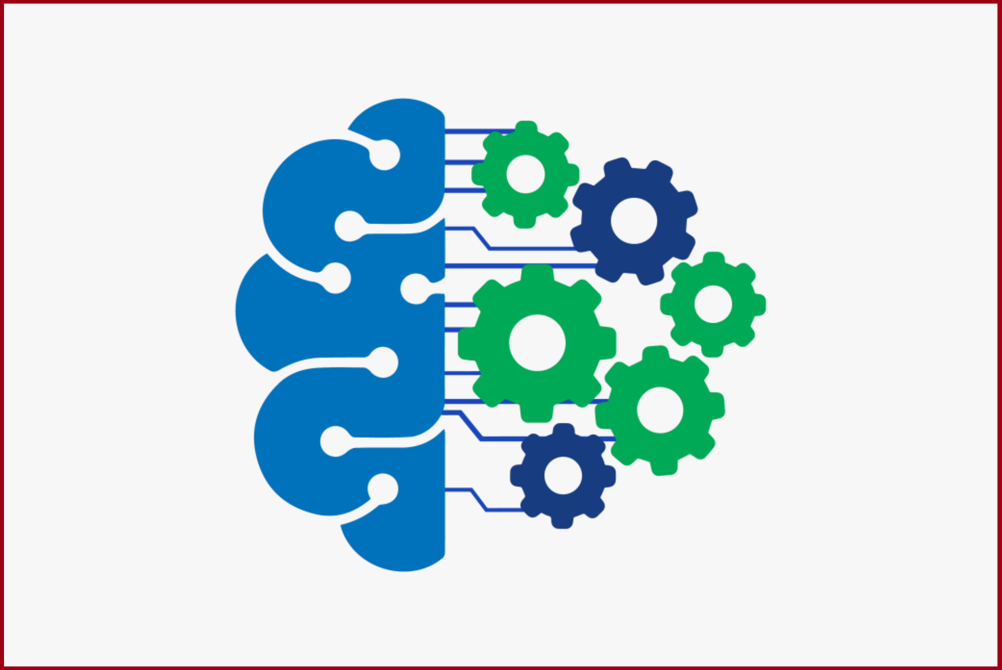
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA



Learning from Uniform Convergence

Machine Learning 2023-24
UML Book Chapter 4
Slides P. Zanuttigh (some material from F. Vandin slides)

Empirical and True Risk

Consider a generic learning algorithm:

1. Receive a training set S
2. Evaluate the error of each possible $h \in \mathcal{H}$ on S and select the one with lowest empirical error h^*

➢ Is $h^* \in \mathcal{H}$ minimizing the empirical error on S also minimizing the true error on *D* ?

It suffices to ensure that the empirical error of ***all*** *$h \in \mathcal{H}$ is a good approximation of their true error (i.e., $L_S(h)$ similar to $L_D(h), \forall h$)*

Notice: ***sufficient*** *not necessary condition*

ε-Representative Set

Idea: focus on when the empirical risks (errors) of all members of $\mathcal{H}$ are good approximations of their true risk

Definition (ε-representative)

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

$$\forall h \in \mathcal{H}: \ |L_S(h) - L_D(h)| \leq \epsilon$$

Theorem:

Assume that training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, distribution *D*). Then, any output of $ERM_{\mathcal{H}}(S)$ (i.e., any $\mathrm{h_s} \in arg\min_{h\in\mathcal{H}} L_S(h)$) satisfies:

$$L_D(h_S) \leq \min_{h\in\mathcal{H}} L_D(h) + \epsilon$$

Consequence: if with probability at least 1-δ, a random training set S is ε-representative then the ERM rule is an agnostic PAC learner

Demonstration

Proof of the theorem:

1. ε/2−representative : $\forall h \in \mathcal{H}\colon\ |L_S(h_s) - L_D(h_s)| \leq \frac{\epsilon}{2} \rightarrow L_D(h_s) \leq L_S(h_s) + \frac{\epsilon}{2}$
2. h_s ERM predictor: $\forall h \in \mathcal{H}\colon\ L_S(h_s) \leq L_S(h) \rightarrow L_D(h_s) \leq L_S(h) + \frac{\epsilon}{2}$
3. ε/2−representative : $\forall h \in \mathcal{H}\colon\ |L_S(h) - L_D(h)| \leq \frac{\epsilon}{2} \rightarrow L_S(h) \leq L_D(h) + \frac{\epsilon}{2}$

Combine together:

$$L_D(h_s) \leq L_S(h_s) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq \left(L_D(h) + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}$$

$$\Downarrow$$

$$L_D(h_s) \leq L_D(h) + \epsilon$$

Uniform Convergence

Same m for all h and all D

Property of hypothesis class

Definition (uniform convergence):

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

Uniform Convergence and PAC Learnability

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then:

1. The class is agnostically PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$
2. The $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$

- Demonstration follows from the previous theorem and the definition of uniform convergence
- *Recall theorem*: if with probability at least 1-δ, a random training set S is ε-representative then the ERM rule is an agnostic PAC learner
- Notice: the theorem requires an $\frac{\epsilon}{2} - representative$ set to achieve an accuracy of ϵ

Finite Classes are Agnostic PAC Learnable

Proposition:

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain and let $\ell: \mathcal{H} x Z \rightarrow [0,1]$ be a loss function. Then:

- $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

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

Need to be $\frac{\epsilon}{2} - representative$

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

Proof not part of the course, basic idea: first prove that uniform convergence holds for a finite hypothesis class, then use previous result on uniform convergence and PAC learnability

Discretization Trick

Note: In many real world applications we consider hypothesis classes determined by a set of parameters in $\mathbb{R}$

- Assume an hypothesis class determined by *d* real number parameters
- In principle the hypothesis class is of infinite size, but…
- … in practice we use a finite representation for numbers:
  - e.g., real numbers represented with 64 bits double precision variables
- For *d* 64-bits parameters, $|\mathcal{H}| = 2^{64d}$ -> $|\mathcal{H}|$ is large but finite
- Sample complexity bounded by $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta) \leq \frac{2\log(2\frac{2^{64d}}{\delta})}{\epsilon^2}$
  - Check book recalling that $log = \log_e \neq \log_2$!
- Issue: *the bound depends on the chosen number representation*

Demonstration (1)

1. Uniform convergence (UC): with probability $\geq 1-\delta$, S is $\epsilon - representative$:
$$D^m(\{S: \forall h \in \mathcal{H}, |L_s(h) - L_D(h)| \leq \epsilon\}) \geq 1-\delta$$

2. Rewrite focusing on the probability of not having UC:
$$P_{bad} = D^m(\{S: \exists h \in \mathcal{H}, |L_s(h) - L_D(h)| > \epsilon\}) \leq \delta$$
error in the book!

3. Rewrite set $\{S: \exists h \in \mathcal{H}, |L_s(h) - L_D(h)| > \epsilon\}$ as the union over *h*:
$$\{S: \exists h \in \mathcal{H}, |L_s(h) - L_D(h)| > \epsilon\} = \bigcup_{h \in \mathcal{H}} \{S: |L_s(h) - L_D(h)| > \epsilon\}$$

4. Apply union bound:
$$D^m(\{S: \exists h \in \mathcal{H}, |L_s(h) - L_D(h)| > \epsilon\}) = D^m\left(\bigcup_{h \in \mathcal{H}} \{S: |L_s(h) - L_D(h)| > \epsilon\}\right) \leq \sum_{h \in \mathcal{H}} D^m(\{S: |L_s(h) - L_D(h)| > \epsilon\})$$

Demonstration not part of the course

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Demonstration (2)

Consider:

$$D^m(\{S\colon \exists h \in \mathcal{H}, |L_s(h) - L_D(h)| > \epsilon\}) = D^m\left(\bigcup_{h\in\mathcal{H}} \{S\colon |L_s(h) - L_D(h)| > \epsilon\}\right) \le \sum_{h\in\mathcal{H}} D^m(\{S\colon |L_s(h) - L_D(h)| > \epsilon\})$$

- Next step: demonstrate that for any fixed hypotheses *h* the difference $|L_s(h) - L_D(h)|$ is likely to be small
- Notice that $L_D(h)$ is the expectation and $L_s(h)$ the average value: the random variable should not deviate too much from its expectation
- *INTUITIVE IDEA from* ***law of large numbers****: if m is large the average converges to the expectation*

Demonstration not part of the course

Demonstration (3)

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 $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\varepsilon > 0$

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

- Apply Hoeffding inequality to our case (assuming [0,1] interval):

$$D^m(\{S: |L_s(h) - L_D(h)| > \epsilon\}) = P\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

- Apply to the sum over h:

$$\sum_{h\in\mathcal{H}} D^m(\{S: |L_s(h) - L_D(h)| > \epsilon\}) \leq |\mathcal{H}|2e^{-2m\epsilon^2}$$

- Finally: we already demonstrated that purple part is smaller or equal than red, need to find m for which it is smaller or equal than δ :
  - force red part to be smaller than δ $\Rightarrow m \geq \log(\frac{2|\mathcal{H}|}{\delta})/2\epsilon^2$

Demonstration not part of the course

Demonstration (4)

For $m \geq \log(\frac{2|\mathcal{H}|}{\delta})/2\epsilon^2$ it holds that $\sum_{h\in\mathcal{H}} D^m(\{S: |L_s(h) - L_D(h)| > \epsilon\}) \leq \delta$
(demonstrated in previous slide)

Consequence: any finite hypothesis class has uniform convergence property with sample complexity $m_{\mathcal{H}}^{UC} \leq \left\lceil \log(\frac{2|\mathcal{H}|}{\delta})/2\epsilon^2 \right\rceil$

From theorem* (uniform convergence implies PAC learnable): $\mathcal{H}$ is PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2},\delta) \leq \left\lceil 2\log(\frac{2|\mathcal{H}|}{\delta})/\epsilon^2 \right\rceil$

(*) recall:

Proposition

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 $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2,\delta)$. Furthermore, in that case the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

Demonstration not part of the course