Machine Learning

Learning Model

Fabio Vandin

October 13th, 2023

previous PDF

Empirical Risk Minimization

Learner outputs $h_S : \mathcal{X} \rightarrow \mathcal{Y}$.

Goal: find h_S which minimizes the generalization error $L_{\mathcal{D},f}(h)$

$L_{\mathcal{D},f}(h)$ is unknown!

What about considering the error on the training data, that is, reporting in output h_S that minimizes the error on training data?

Training error: $L_S(h) \overset{def}{=} \frac{|\{i : h(x_i) \neq y_i, 1 \leq i \leq m\}|}{m}$

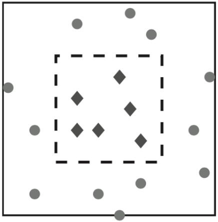
Note: the *training error* is also called *empirical error* or *empirical risk*

Empirical Risk Minimization (ERM): produce in output h minimizing $L_S(h)$

previous PDF

What can go wrong with ERM?

Consider our simplified movie ratings prediction problem. Assume data is given by:



Assume $\mathcal{D}$ and f are such that:

- instance x is taken uniformly at random in the square ($\mathcal{D}$)
- label is 1 if x inside the inner square, 0 otherwise (f)
- area inner square $= 1$, area larger square $= 2$

Consider classifier given by

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in \{1, \dots, m\} : x_i = x \\ 0 & \text{otherwise} \end{cases}$$

previous PDF

Is it a good predictor?

$L_S(h_S) = 0$ but $L_{\mathcal{D},f}(h_S) = 1/2$

Good results on training data but poor generalization error
$\Rightarrow$ **overfitting**

When does ERM lead to good performances in terms of generalization error?

Hypothesis Class and ERM

Apply ERM over a **restricted set** of hypotheses $\mathcal{H}$ = *hypothesis class*

- each $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$

$\mathrm{ERM}_{\mathcal{H}}$ learner:

$$\mathrm{ERM}_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

model picked by ERM procedure considering only models from $\mathcal{H}$

Hypothesis Class and ERM

Apply ERM over a **restricted set** of hypotheses $\mathcal{H}$ = *hypothesis class*

- each $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$

$\mathrm{ERM}_{\mathcal{H}}$ learner:

$$\mathrm{ERM}_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

Which hypothesis classes $\mathcal{H}$ do not lead to overfitting?

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

movies example: $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$, $\mathcal{Y} = \{-1, 1\}$

$$\mathcal{H} = \left\{ h_{a,b}(\vec{x}) : h_{a,b}(\vec{x}) = \text{sign}\left(\boxed{a} x_1 + \boxed{b} x_2\right), \ a, b \in \mathbb{R} \right\}$$

$$|\mathcal{H}| = +\infty$$

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

training set

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

Assumptions

- **Realizability:** there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

Assumptions

- **Realizability:** there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$
- **i.i.d.:** examples in the training set are independently and identically distributed (i.i.d) according to $\mathcal{D}$, that is $S \sim \mathcal{D}^m$

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

Assumptions

- **Realizability:** there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$
- **i.i.d.:** examples in the training set are independently and identically distributed (i.i.d) according to $\mathcal{D}$, that is $S \sim \mathcal{D}^m$

Observation: realizability assumption implies that $L_S(h^*) = 0$

$$\Rightarrow L_S(h_S) = 0$$

because the training set is generated by the distribution D, so if the generalization error is 0 so is the training error

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

Assumptions

- **Realizability:** there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$
- **i.i.d.:** examples in the training set are independently and identically distributed (i.i.d) according to $\mathcal{D}$, that is $S \sim \mathcal{D}^m$

Observation: realizability assumption implies that $L_S(h^*) = 0$

Can we *learn* (i.e., find using ERM) h^*?

(Simplified) PAC learning

Probably Approximately Correct (PAC) learning

Since the training data comes from $\mathcal{D}$:

- we can only be **approximately** correct
- we can only be **probably** correct

(Simplified) PAC learning

Probably Approximately Correct (PAC) learning

Since the training data comes from $\mathcal{D}$:

- we can only be **approximately** correct
- we can only be **probably** correct

Parameters:

- *accuracy parameter* ε: we are satisfied with a *good* h_S: $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ (ε small)
- *confidence parameter* δ: want h_S to be a *good* hypothesis with probability $\geq 1 - \delta$ (δ small)

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$, $\varepsilon \in (0,1)$, and $m \in \mathbb{N}$ such that

we don't know f,
we don't know $\mathcal{D}$

$$|S| = m \ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$

Then for any f and any $\mathcal{D}$ for which the realizability assumption holds, with probability $\ge 1 - \delta$ we have that for every ERM hypothesis h_S it holds that

$$L_{\mathcal{D},f}(h_S) \le \varepsilon.$$

the hypothesis class H needs to be "powerful" enough

Note: log = natural logarithm

With finite hypotheses classes ($\mathcal{H}$), I can "almost always" find a "good hypothesis" if I have "enough data"

with prob $\geq 1 - \delta$

$L_{\mathcal{D},f}(h_s) \leq \varepsilon$

$m \geq \frac{1}{\varepsilon} \log(|\mathcal{H}|/\delta)$

Proof (see book as well, Corollary 2.3)

Let $S|_x = \{x_1, x_2, \dots, x_m\}$ be the instances in the training set S. We want to bound (i.e., an upper bound) to:

$\mathcal{D}^m(\{S_{|x} : L_{\mathcal{D},f}(h_S) > \varepsilon\})$. Let $\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \varepsilon\}$ (BAD HYPOTHESES)

and $M \doteq \{S_{|x} : \exists\, h \in \mathcal{H}_B, L_S(h) = 0\}$ (MISLEADING SAMPLES)

Since we have the realizability assumption: $L_S(h_S) = 0$

$\Rightarrow L_{\mathcal{D},f}(h_S) > \varepsilon$ only if some $h \in \mathcal{H}_B$ has $L_S(h) = 0$.

That is, our training data must be in the set M:

$$\{S_{|x} : L_{\mathcal{D},f}(h_S) > \varepsilon\} \subseteq M.$$

datasets in which the hs has big gen. error \subseteq datasets in which some h has big gen. error

Note that: $M = \bigcup_{h \in \mathcal{H}_B} \{S_{|x} : L_S(h) = 0\}$.

Therefore $\mathcal{D}^m(\{S_{|x} : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \le \mathcal{D}^m(M) = \mathcal{D}^m\left(\bigcup_{h \in \mathcal{H}_B} \{S_{|x} : L_S(h) = 0\}\right)$

UNION BOUND $\le \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S_{|x} : L_S(h) = 0\})$ (★)

Now let's fix $h \in \mathcal{H}_B$: $L_S(h) = 0 \iff \forall i = 1, \dots, m : h(x_i) = f(x_i)$

Therefore : $\mathcal{D}^m(\{S_{|x} : L_S(h) = 0\}) = \mathcal{D}^m(\{S_{|x} : \forall i = 1, \dots, m; h(x_i) = f(x_i)\})$

because $x_1, \dots, x_m$ are i.i.d. from $\mathcal{D}$

$$\boxed{=} \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) \qquad (\star\star)$$

Consider some i, $1 \le i \le m$: $\mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$

$$= 1 - \underbrace{\mathcal{D}(\{x_i : h(x_i) \neq f(x_i)\})}_{L_{\mathcal{D},f}(h) = \Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)]}$$

since $h \in \mathcal{H}_B$

$$= 1 - L_{\mathcal{D},f}(h)$$

$$\boxed{\le} 1 - \varepsilon \boxed{\le} e^{-\varepsilon}$$

Taylor expansion : $e^x = \sum_{n=0}^{+\infty} \left(\frac{x^n}{n!}\right) \Rightarrow e^{-x} \ge 1 - x$

Combining this with $(\star\star)$: $\mathcal{D}^m(\{S_{|x} : L_S(h) = 0\}) \le \prod_{i=1}^{m} e^{-\varepsilon} = e^{-m\varepsilon}$

Combining the above with $(\star)$: $\mathcal{D}^m(\{S_{|x} : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \le \sum_{h \in \mathcal{H}_B} e^{-m\varepsilon} =$

$= |\mathcal{H}_B| \cdot e^{-m\varepsilon} \le \underbrace{|\mathcal{H}| \cdot e^{-m\varepsilon}}$. Now, given the choice of m $\left(\ge \frac{1}{\varepsilon} \log(|\mathcal{H}|/\delta)\right)$

we have $\le |\mathcal{H}| \cdot e^{-\varepsilon (\log(|\mathcal{H}|/\delta)) \cdot \frac{1}{\varepsilon}} = |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta$ $\square$

PAC Learning

Definition (PAC learnability)

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

$m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$: *sample complexity* of learning $\mathcal{H}$.

- $m_{\mathcal{H}}$ is the minimal integer that satisfies the requirements.

Corollary

Every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$.

What is the algorithm to find the good hypothesis hs? ERM!

the smallest integer so \le and \ceil