Data Science

Marc Tommasi

September 19, 2023

Outline

Outline

Formalization

- Supervised learning
- Data are described by some attributes. Data lie in a representation space $x \in \mathcal{X}$
- The target is $y \in \mathcal{Y}$
- A sample $S \subseteq \mathcal{X} \times \mathcal{Y}$
- Output: a prediction rule, a model is a function $f : \mathcal{X} \to \mathcal{Y}$

Vocabulary

- representation space: features, attributes, (voire champs)
- échantillons ou jeu de données, dataset
- instances, records, examples, records, enregistrements
- target, label, or class

Basic assumptions

- The data is generated from a fixed, but unknown data distribution $\mathcal{D}$.
- There exists a target function f that labels the data.
- The true error, the generalization error, the true risk of an hypothesis h is:

$$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x \mid h(x) \neq f(x)\}).$$

- $\mathcal{D}$ is unknown the learner cannot compute $L_{\mathcal{D},f}(h)$

ERM : Empirical Risk Minimization Principle

- The learner can compute the empirical risk or empirical error

$$L_S(h) = \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m},$$

where $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

ERM principle

The learner outputs an hypothesis h_S that minimizes the empirical risk

$$h_S = \operatorname*{argmin}_h L_S(h)$$

First limits of the ERM principle

What is the empirical error of this rule?

$$h(x) = \begin{cases} y_i \text{ if there exists } i \text{ s.t. } x = x_i \\ 0 \text{ otherwise} \end{cases}$$

First limits of the ERM principle

What is the empirical error of this rule?

$$h(x) = \begin{cases} y_i \text{ if there exists } i \text{ s.t. } x = x_i \\ 0 \text{ otherwise} \end{cases}$$

- Learning by heart minimizes the empirical error!
- Note that some classes of function can mimic learning by heart (the class of all polynomials)

ERM with a bias

- Bias: we fix a priori an hypothesis class $\mathcal{H}$

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

- This can be a background knowledge
- Example: rectangles or lines
- Trade-off:
  - a large and expressive class allows to model more complex tasks but can be subject to overfitting
  - a small and poorly expressive class limits overfitting but increases the risk of having bad performances.

Question

Do we learn when we follow the ERM principle?

Confidence and Approximation

Confidence

- Note that $L_S(h_S)$ is a random variable. The stochasticity comes from the fact that S is drawn from $\mathcal{D}$.
- Let δ be the probability to sample a "bad" sample. Then $1 - \delta$ is the confidence.

Approximation

- The computation of an hypothesis h is error-prone.
- We can tolerate some error up to some threshold ϵ: $L_{\mathcal{D},f}(h_S) \leq \epsilon$
- Bad hypothesis corresponds to $L_{\mathcal{D},f}(h_S) > \epsilon$

The case of finite hypothesis classes

Theorem (Finite classes are learnable under the realizability assumption)

If $\mathcal{H}$ is finite, for any $\epsilon > 0$ and $\delta \in [0, 1]$, for any target function f, for any distribution $\mathcal{D}$ such that there exists $h \in \mathcal{H}, L_{\mathcal{D},f}(h) = 0$ (Realizability assumption), then with probability $1 - \delta$ over the choice of an i.i.d. sample S of size larger than $\frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$, the true error of and ERM hypothesis h_S is lower than ϵ

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

- $\mathcal{H}$ is finite, take $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$ examples in S, then
$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \delta.$$

- Let us compute $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\})$ to bound the probability to draw a sample to leads to a failure of the learning process. . .

A bad sample...

- The ERM hypothesis h_S is wrong if $L_{\mathcal{D},f}(h_S) > \epsilon$ and $L_S(h_S) = 0$ (realizability assumption).
- Let

$$M = \bigcup_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \{S' \mid L_{S'}(h) = 0\},$$

- if S belongs to M then

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M).$$

- Hence applying the union bound ($\mathcal{D}(\bigcup_i A_i) \leq \sum_i \mathcal{D}(A_i)$)

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \sum_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}).$$

A bound for the right hand side

- $L_{S'}(h) = 0$ when for any x_i in S' we have $h(x_i) = f(x_i)$. Samples are i.i.d. hence

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) = \prod_i^m \mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\}).$$

- But $L_{\mathcal{D},f}(h) = \mathcal{D}(\{x \mid f(x) \neq h(x)\})$ hence

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

- If h has an error larger than ϵ then

$$\mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\}) \leq 1 - \epsilon,$$

and

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) \leq (1 - \epsilon)^m.$$

End of the proof!

- We combine

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \sum_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\})$$

and

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) \leq (1-\epsilon)^m.$$

If we denote by $\mathcal{H}_B = \{h \text{ s.t. } L_{\mathcal{D},f}(h) > \epsilon\}$ the set of bad hypothesis

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq |\mathcal{H}_B|(1-\epsilon)^m \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}.$$

ERM and error estimation

- The final objective is to minimize the true error, but the learner cannot compute it.
- She needs new data (different from the learning data) to estimate the true error.
- An important assumption is that new data will be generated by the same distribution $\mathcal{D}$.
- A practical approach: split the data into 2 sets:
  - a training set to learn a model
  - a test set to estimate the true error