The Probably Approximately Correct (PAC) Learning

Prof. Dan A. Simovici

UMB

1. The Agnostic PAC Learning

2. The Scope of Learning Problems

A learning algorithm $\mathcal{A}$ starts with a hypothesis class $\mathcal{H}$ and a sample S, under certain conditions, it returns a hypothesis h that has a small true error. A specific definition is given next.

We assume that the data set $\mathcal{X}$ is equipped with a probabilty distribution $\mathcal{D}$.

What is the PAC Model?

Definition

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ and a learning algorithm $\mathcal{A}$ such that for every $\epsilon, \delta \in (0,1)$, every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \longrightarrow \{0,1\}$, if realizability assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the algorithm $\mathcal{A}$ on a sample S that consists of $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ generated by $\mathcal{D}$ and labeled by f, $\mathcal{A}$ returns a hypothesis h such that, with probability at least $1 - \delta$ (over the choice of examples), we have for the true error $L_{(\mathcal{D},f)}(h)$:

$$P(L_{(\mathcal{D},f)}(h) \leqslant \epsilon) \geqslant 1 - \delta.$$

ϵ is the accuracy parameter and δ is the confidence parameter

Approximation Parameters

- the accuracy parameter ϵ determines how far the output classifier can be from the optimal one, and

- the confidence parameter δ indicates how likely is the classifier is to meet that accuracy requirement.

What is Agnostic PAC Learning?

- The realizability assumption, the existence of a hypothesis $h^* \in \mathcal{H}$ such that $P_{x \sim \mathcal{D}}(h^*(x) = f(x)) = 1$ is not practical in many cases.

- Agnostic learning replaces the realizability assumption and the targeted labeling function f, with a distribution $\mathcal{D}$ defined on pairs (data, labels), that is, with a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.

- When the probability distribution $\mathcal{D}$ was defined on $\mathcal{X}$, the generalization error of a hypothesis was defined as:

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

- For agnostic learning $\mathcal{D}$ is defined over $\mathcal{X} \times \mathcal{Y}$, so we redefine the generalization error as:

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

  We seek a predictor for which $L_{\mathcal{D}}(h)$ is minimal.

- The definition of the empirical risk remains the same:

$$L_S(h) = \frac{|\{i \mid h(x_i) \neq y_i \text{ for } 1 \leqslant i \leqslant m\}|}{m}.$$

The Bayes Classifier and Its Optimality

Let $\mathcal{D}$ be any probability distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$.
Let X be a random variable ranging over $\mathcal{X}$ and Y be a random variable ranging over $\mathcal{Y} = \{0, 1\}$.

The Bayes predictor is the function $f_{\mathcal{D}}$ defined as

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } P(Y = 1 | X = x) \geqslant \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Theorem

Given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ the best label predicting function $f : \mathcal{X} \longrightarrow \{0,1\}$ is the Bayes predictor $f_{\mathcal{D}}$.

In other words, we need to prove that for hypothesis $g : \mathcal{X} \longrightarrow \{0,1\}$ we have $L_{\mathcal{D}}(f_{\mathcal{D}}) \leqslant L_{\mathcal{D}}(g)$.

Proof

Let X be a random variable ranging over $\mathcal{X}$, Y be a random variable ranging over $\mathcal{Y} = \{0, 1\}$, and let α_x be the probability of a having a label 1 given x, that is, $\alpha_x = P(Y = 1|X = x)$.

With this notation the Bayes predictor is

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \alpha_x \geqslant \frac{1}{2} \\ 0 & \text{if } \alpha_x < \frac{1}{2}. \end{cases}$$

Proof (cont'd)

We have:

$$
\begin{aligned}
L_{\mathcal{D}}(f_{\mathcal{D}}) &= P(f_{\mathcal{D}}(X) \neq y | X = x) \\
&= P(f_{\mathcal{D}}(x) = 1 | X = x) P(Y = 0 | X = x) \\
&\quad + P(f_{\mathcal{D}}(x) = 0 | X = x) P(Y = 1 | X = x) \\
&= P\left(\alpha_x \geqslant \frac{1}{2}\right) P(Y = 0 | X = x) \\
&\quad + P\left(\alpha_x < \frac{1}{2}\right) P(Y = 1 | X = x)
\end{aligned}
$$

Note: when we write $P(f_{\mathcal{D}}(X) \neq y | X = x)$ we mean the probability that a pair (x, y) is such that $f_{\mathcal{D}}(X) \neq y$ assuming that $X = x$. Similar conventions apply to all probabilities listed above.

Proof (cont'd)

If $\alpha_x \geqslant \frac{1}{2}$, then

$$\min\{\alpha_x, 1-\alpha_x\} = 1-\alpha_x, P\left(\alpha_x \geqslant \frac{1}{2}\right) = 1, P\left(\alpha_x < \frac{1}{2}\right) = 0,$$

and

$$\begin{aligned} & P\left(\alpha_x \geqslant \frac{1}{2}\right)(1-\alpha_x) + P\left(\alpha_x < \frac{1}{2}\right)\alpha_x \\ = \; & 1-\alpha_x = \min\{1-\alpha_x, \alpha_x\}. \end{aligned}$$

If $\alpha_x < \frac{1}{2}$, then $\min\{\alpha_x, 1-\alpha_x\} = \alpha_x$, $P\left(\alpha_x \geqslant \frac{1}{2}\right) = 0$, $P\left(\alpha_x < \frac{1}{2}\right) = 1$ and

$$\begin{aligned} & P\left(\alpha_x \geqslant \frac{1}{2}\right)(1-\alpha_x) + P\left(\alpha_x < \frac{1}{2}\right)\alpha_x \\ = \; & \alpha_x = \min\{1-\alpha_x, \alpha_x\}. \end{aligned}$$

Proof (cont'd)

Let g be any other classifier. We have:

$$
\begin{aligned}
P(g(X) \neq Y | X = x) &= P(g(X) = 0 | X = x) P(Y = 1 | X = x) \\
&\quad + P(g(X) = 1 | X = x) P(Y = 0 | X = x) \\
&= P(g(X) = 0 | X = x) \alpha_x \\
&\quad + P(g(X) = 1 | X = x)(1 - \alpha_x) \\
&\geqslant P(g(X) = 0 | X = x) \min\{\alpha_x, 1 - \alpha_x\} \\
&\quad + P(g(X) = 1 | X = x) \min\{\alpha_x, 1 - \alpha_x\} \\
&\geqslant (P(g(X) = 0 | X = x) + P(g(X) = 1 | X = x)) \\
&\quad \cdot \min\{\alpha_x, 1 - \alpha_x\} \\
&= \min\{\alpha_x, 1 - \alpha_x\} = P(f_{\mathcal{D}}(X) \neq y | X = x).
\end{aligned}
$$

Agnostic PAC-Learnability

Definition

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

- $\epsilon, \delta \in (0,1)$ and
- for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

when running $\mathcal{A}$ on $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ iid examples generated by $\mathcal{D}$, $\mathcal{A}$ returns a hypothesis h such that with probability at least $1-\delta$ (over the choice of the m training examples) we have

$$L_{\mathcal{D}}(h) \leqslant \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

- If the realizability assumption holds, agnostic PAC learning provides the same guarantees as PAC learning.
- When the realizability assumption does not hold, no learner can guarantee an arbitrary small error.
- A learner $\mathcal{A}$ can declare success if the error is not much larger than the smallest error achievable by a hypothesis from $\mathcal{H}$.

Multiclass Classification

Example

Let $\mathcal{X}$ be a set of document features, and $\mathcal{Y}$ a set of topics (sports, politics, health, etc.).

By document features we mean counts of certain key words, size, or origin of the document.

The loss function will be the probability of the event that occurs when the predictor suggest a wrong label.

Regression

Example

In *regression* we seek to find a functional relationship h between the $\mathcal{X}$ and $\mathcal{Y}$ components of the data.

For example, to predict the weight of a baby at birth $\mathcal{X}$ can be a set of triplets in $\mathbb{R}^3$

(head circumference, abdominal circumference, femur length)

and $\mathcal{Y}$ is is the weight at birth. We seek h that will minimize the loss $L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}}(h(x) - y)^2$.

Generalized Loss Functions

Definition

Given a set of hypotheses $\mathcal{H}$, a domain Z, a loss function is a function $\ell : \mathcal{H} \times Z \longrightarrow \mathbb{R}_+$.

For prediction problems we have $Z = \mathcal{X} \times \mathcal{Y}$.

Definition

The risk function is the expected loss of the classifier $h \in \mathcal{H}$ with respect to a probability distribution $\mathcal{D}$ over Z, namely

$$L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}(\ell(h, z)).$$

The empirical risk is the expected loss over the sample $S = (z_1, \ldots, s_m) \in Z^m$ as

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i).$$

0-1 Loss

The random variable z ranges over $\mathcal{X} \times \mathcal{Y}$ and the loss function is

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

This is used in binary or multiclass classification problems.

For the 0/1 loss the definition of $L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}(\ell(h,z))$ coincides with the previous definition in the agnostic PAC, $L_{\mathcal{D}}(h) = \mathcal{D}(\{(x,y) \mid h(x) \neq y\})$.

Square Loss

The random variable z ranges over $\mathcal{X} \times \mathcal{Y}$ and the loss function is

$$\ell_{sq}(h, (x, y)) = (h(x) - y)^2.$$

Agnostic PAC Learnability for General Loss Functions

Definition

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \longrightarrow \mathbb{R}_+$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ and a learning algorithm $\mathcal{A}$ with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over Z, when running $\mathcal{A}$ on $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ iid examples generated by $\mathcal{D}$, $\mathcal{A}$ returns a hypothesis h such that with probability at least $1 - \delta$ (over the choice of the m training examples) we have

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

where $L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}(\ell(h, z))$.