A.Z
Machine Learning
Nov, 2023

lec 2 PAC for finite H

Abstract

This part introduced Probably Approximate Correct (PAC) theory, and proofed a finite hypothesis set is PAC-learnable.

1 Generalization

In machine learning, we would like to minimize the generalization error , or the true risk

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim D}[L(h(\mathbf{x}), y)],$$

where $L(y, y')$ is a suitable loss function (e.g. zero-one loss)

2 Probably Approximate Correct (PAC)

2.1 Ingredients

- input space X containing all possible inputs X
- set of possible labels $\mathcal{Y}$ (in binary classification $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, +1\}$)
- Concept class $\mathcal{C}$ containing concepts $C : X \mapsto \mathcal{Y}$ (to be learned), concept C gives a label $C(x)$ for each input x
- unknown probability distribution D
- training sample $S = (x_1, C(x_1)), \ldots, (x_m, C(x_m))$ drawn independently from D
- hypothesis class $\mathcal{H}$, in the basic case $\mathcal{H} = \mathcal{C}$ but this assumption can be relaxed

2.2 Goal

The goal in PAC is to learn a hypothesis with a low generalization error

$$R(h) = \mathbb{E}_{x\sim D}\left[L_{0/1}(h(x), C(x))\right] = \Pr_{x\sim D}(h(x) \neq C(x))$$

A class $\mathcal{C}$ is **PAC-learnable**, if there exist an algorithm $\mathcal{A}$ that given a training sample S outputs a hypothesis $h_S \in \mathcal{H}$ that has generalization error satisfying

$$\Pr\left(R\left(h_S\right) \leq \epsilon\right) \geq 1 - \delta$$

For any distribution D, for arbitrary $\epsilon, \delta > 0$ and sample size $m = |S|$ that grows polynomially in $1/\epsilon, 1/\delta$. And for any concept $C \in \mathcal{C}$
In addition, if $\mathcal{A}$ runs in time polynomial in $m, 1/\epsilon$, and $1/\delta$ the class is called **efficiently PAC learnable**

ϵ sets the level of generalization error that is of interest to us, say we are content with predicting incorrectly 10% of the new data points: $\epsilon = 0.1$
$1 - \delta$ sets a level of confidence, if we are content of the training algorithm to fail 5% of the time to provide a good hypothesis: $\delta = 0.05$
Polynomial Growth in $1/\epsilon, 1/\delta$: In simpler terms, if you want to increase the accuracy or the confidence of your learning algorithm, you don't need an exponentially larger sample size; a polynomially larger sample size will suffice. For example, if the required sample size grows as a function like $m = O\left(\frac{1}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right)\right)$, it indicates polynomial growth. As ϵ becomes smaller (more accuracy) or δ becomes smaller (more confidence), m increases, but only polynomially.
The event "low generalization error", $\{R(h_S) \leq \epsilon\}$ is considered as a random variable because we cannot know beforehand which hypothesis $h_S \in \mathcal{H}$ will be selected by the algorithm

Sample complexity bound relying on the size of the hypothesis class

$$\Pr\left(R\left(h_s\right) \leq \epsilon\right) \geq 1 - \delta \text{ if } m \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$$

an equivalent generalization error bound:

$$R(h) \leq \frac{1}{m}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$$

2.3 But...

Imagine we have a Boolean conjunction
There are 2^d possible input vectors, size of the input space is $|X| = 2^d$, where d is the number of variables
We can define a boolean formula that outputs 1 for an arbitrary subset of $S \subset X$ and zero outside that subset:

$$f_S(\mathbf{x}) = (\mathbf{x} = \mathbf{x}_1)\, OR\, (\mathbf{x} = \mathbf{x}_2)\, OR \cdots OR\, \left(\mathbf{x} = \mathbf{x}_{|S|}\right)$$

We can pick the subset in $2^{|X|}$ ways.
Thus we have $|\mathcal{H}| = 2^{2^d}$ different boolean formula (The hypothesis corresponding to a subset S is the rule "output 1 if the input is in S, and output 0 otherwise.")
It means that the hypothesis class is considered not PAC-learnable even for a finite hypothesis set! Well, if this is true, we can drop this class :)
In fact we are not select hypothesis randomly, we can get some 'good' hypothesis after m rounds

2.4 Proof of the PAC bound for finite hypothesis classes

Consider any hypothesis $h \in \mathcal{H}$ with $R(h) > \epsilon$ For h to be consistent $\hat{R}(h) = 0$, all training examples need to miss the region where h is making an error.
The probability of this event is

$$\Pr(\hat{R}(h) = 0 \mid R(h) > \epsilon) \leq (1 - \epsilon)^m$$

where m means the number of repeated trials with success probability ϵ

> $\hat{R}(h) = 0$ means the hypothesis h perfectly classifies the training data without any errors.
> $R(h) > \epsilon$ means the true risk of the hypothesis h over the entire data distribution is greater than some threshold ϵ.
> $\Pr(\hat{R}(h) = 0 \mid R(h) > \epsilon)$ is a measure of the likelihood that a hypothesis appears to be perfect on the training data but actually performs poorly on the overall data distribution. This likehood means overfitting, i.e. a hypothesis performing well on training data but poorly on new data
> $(1 - \epsilon)^m$ means the probability of randomly picking m examples (the size of the training set) from the data distribution and none of these examples falling into the region where the hypothesis h makes errors. Since $R(h) > \epsilon$, the probability of not picking an example where h errs in a single draw is $1 - \epsilon$.
> This formula means as the size of the training set m increases, the likelihood of overfitting decreases exponentially.

But the specific choice of hypothesis h by the learning algorithm is not important for the argument that follows.
By uniform convergence, the empirical risk (error on the training set) converges uniformly to the true risk for all hypotheses in the hypothesis class as the size of the training set increases, We wish to upper bound the probability that there exists at least one hypothesis h in the hypothesis class $\mathcal{H}$ which is consistent with the training data (i.e., $\hat{R}(h) = 0$) but has a high generalization error (i.e., $R(h) > \epsilon$) for a fixed $\epsilon > 0$.

$$\Pr(\exists h \in \mathcal{H} \mid \hat{R}(h) = 0 \wedge R(h) > \epsilon)$$

Using $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ and $\Pr(A \cap C) \leq \Pr(A \mid C)$ and $\Pr\left(\hat{R}(h) = 0 \mid R\left(h_1\right) > \epsilon\right) \leq (1 - \epsilon)^m$, we have:

$$\begin{aligned}
&\Pr(\exists h \in \mathcal{H} \mid \hat{R}(h) = 0 \wedge R(h) > \epsilon) \\
&= \Pr\left(\left\{\hat{R}\left(h_1\right) = 0 \wedge R\left(h_1\right) > \epsilon\right\} \vee \left\{\hat{R}\left(h_2\right) = 0 \wedge R\left(h_2\right) > \epsilon\right\} \vee \cdots\right) \\
&\leq \sum_{h \in \mathcal{H}} \Pr(\hat{R}(h) = 0 \wedge R(h) > \epsilon) \\
&\leq \sum_{h \in \mathcal{H}} \Pr(\hat{R}(h) = 0 \mid R(h) > \epsilon) \\
&\leq |\mathcal{H}|(1 - \epsilon)^m
\end{aligned}$$

So we have established

$$\Pr(\exists h \in \mathcal{H} \mid \hat{R}(h) = 0 \wedge R(h) > \epsilon) \leq |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}| \exp(-m\epsilon)$$

Solve m to obtain the bound:

$$\begin{aligned} \delta &= |\mathcal{H}| \exp(-m\epsilon) \\ \log \delta &= \log |\mathcal{H}| - m\epsilon \\ m &= \frac{1}{\epsilon}(\log(|\mathcal{H}|) + \log(1/\delta)) \end{aligned}$$

So far we have assumed that there is a consistent hypothesis $h \in \mathcal{H}$ that achieves zero empirival risk on training sample.
In practical, this is often not the case. However as long as the empirical rish $\hat{R}(h)$ is small, a low generalization error can still be achieved.
Theorem (generalization error bound): Let $\mathcal{H}$ be a finite hypothesis set. Then for any $\delta > 0$ with probability at least $1 - \delta$ we have for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log(|\mathcal{H}|) + \log(2/\delta)}{2m}}$$

Means the true risk $R(h)$ has a slower convergence w.r.t. number of examples.