A.Z
Machine Learning
Nov, 2023

lec 3 PAC for infinite H

Abstract

This part introduced Probably Approximate Correct (PAC) theory, and proofed a infinite hypothesis set is PAC-learnable. And introduced a new evaluation for model: Rademacher complexity

1 Recall

1.1 PAC learnability

A class C is PAC-learnable, if there exist an algorithm $\mathcal{A}$ that given a training sample S outputs a hypothesis h_S 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 at polynomially in $1/\epsilon, 1/\delta$

1.2 PAC learning of a hypothesis class

Sample complexity bound relying on the size of the hypothesis class (Mohri et al, 2018): $\Pr\left(R\left(h_s\right) \leq \epsilon\right) \geq 1-\delta$ 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 VC dimension

Intuitively , VC-dim is a measure to the capacity of a hypothesis class to adapt to different concepts
The underlying concept in VC-dim is shattering
$\mathcal{H}$ is said to shatter S if for any possible partition of S into positive S_+ and negative subset S_-, we can find a hypothesis for which $h(x) = 1$ if and only if $x \in S_+$
Point in general position: In a n-dimensional feature space a set of m points $(m > n)$ is in general position if and only if no subset of $(n+1)$ points lie on $(n-1)$ dimensional hyperplane

How to show that $VC\dim(\mathcal{H}) = d$ for a hypothesis class
1. There **exists** a set of inputs of size d that can be shattered by hypothesis in $\mathcal{H}$, i.e. $VCdim(\mathcal{H} \geq d)$
2. There doesnot exist any set of inputs of size d+1 that can be shattered i.e. $VCdim(\mathcal{H} < d+1)$

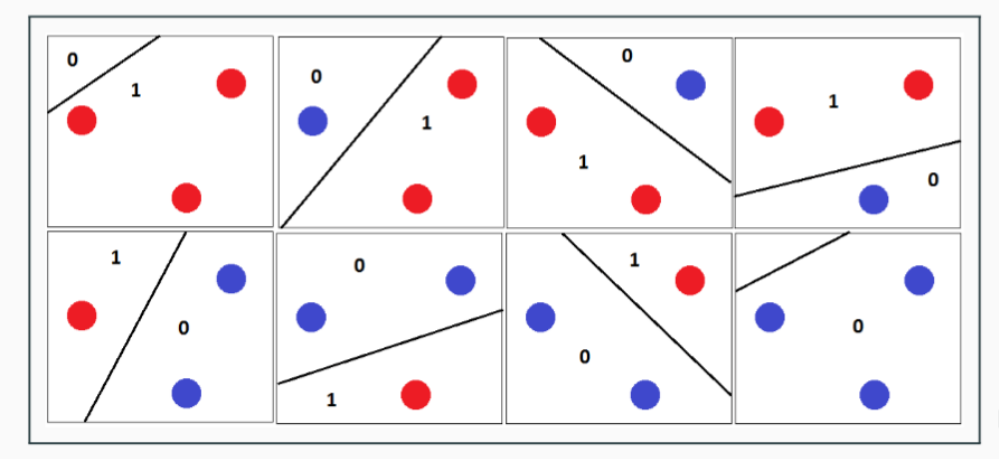

Figure 1: the VC-dim of a line in 2-D is greater than 3

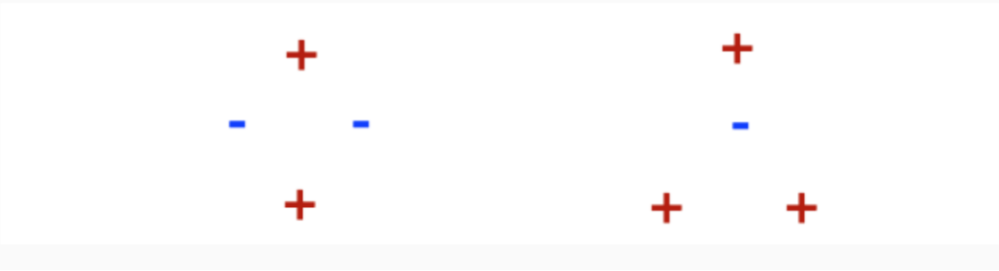

Figure 2: but less than 4

2.1 Growth function

$$\Pi_{\mathcal{H}}(m) = \max_{\{x_1,\ldots,x_m\} \subset X} |\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}|$$

The growth function gives the maximum number of unique labelings the hypothesis class $\mathcal{H}$ can provide for an arbitrary set of input points
The maximum of the growth function is 2^m for a set of m examples Vapnik-Chervonenkis dimension is then

$$\text{VCdim}(\mathcal{H}) = \max_m \{m \mid \Pi_{\mathcal{H}}(m) = 2^m\}$$

Any finite hypothesis class has VC dimension $VC\dim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$

Consider a set of m examples $S = \{x_1, \ldots, x_m\}$ This set can be labeled 2^m different ways, by choosing the labels $y_i \in \{0, 1\}$ independently
Each hypothesis in $h \in \mathcal{H}$ fixes one labeling, a length- m binary vector $\mathbf{y}(h, S) = (h(x_1), \ldots, h(x_m))$
All hypotheses in $\mathcal{H}$ together can provide at most $|\mathcal{H}|$ different labelings in total (different vectors $\mathbf{y}(h, S), h \in \mathcal{H}$)
If $|\mathcal{H}| < 2^m$ we cannot shatter $S \Longrightarrow$ we cannot shatter a set of size $m > \log_2 |\mathcal{H}|$

Example:
Consider a hypothesis class $\mathcal{H} = \{h_\theta\}$ of threshold functions $h_\theta : \mathbb{R} \mapsto \{0, 1\}, \theta \in \mathbb{R}$:

$$h_\theta(x) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{otherwise} \end{cases}$$

What is the VC dimension of this hypothesis class?

Shattering One Point: Consider any single point x_1 in $\mathbb{R}$. We can always find a threshold θ such that $h_\theta(x_1) = 1$ (by choosing $\theta < x_1$) and another threshold such that $h_\theta(x_1) = 0$ (by choosing $\theta \geq x_1$). This means we can realize both possible labelings (0 or 1) for this single point. Therefore, the class can shatter one point.

Failing to Shatter Two Points: Now consider any two distinct points x_1 and x_2 on the real line, without loss of generality, assume $x_1 < x_2$. We can label these points in four ways: (00), (01), (10), (11). While the threshold functions can realize the labelings (00), (01), and (11) by appropriately choosing θ, they cannot realize the labeling (10) (where x_1 is labeled 1 and x_2 is labeled 0). This is because if $x_1 > \theta$ (to make $h_\theta(x_1) = 1$), then $x_2 > \theta$ must also be true (making $h_\theta(x_2) = 1$), given that x_2 is greater than x_1.

3 VC-dim for convex polygons

Let $\mathcal{H}$ be the set of all convex polygons in Euclidean space. This has infinite VC dimension, since the cardinality of shattered subsets is unbounded: for example take equally spaced points on the unit sphere.
You can place arbitarily many points on the surface of the sphere, and still find a convex polygon which includes any subset of those points, and excludes the rest.

3.1 Generalization bound based on VC-dim

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then for any $\delta > 0$, with probability at least $1 - \delta$ the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2\log(em/d)}{m/d}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

m means the number of examples.

The first root reflects the complexity of the hypothesis class relative to the size of the training set. It decreases as the number of examples m increases, especially in relation to the VC-dimension d
The second root is a measure of the confidence in the bound. As δ decreases (meaning we want more confidence), this term increases.

Manifestation of the Occam's razor principle: simpler hypotheses (those with lower VC-dimension) should be preferred, as they require less data to achieve a similar level of generalization. A more complex hypothesis (with higher VC-dim) needs proportionally more training data to ensure the true error rate remains low.

4 Rademacher complexity

Rademacher complexity defines complexity as the capacity of hypothesis class to fit random noise
For binary classification with labels $\mathcal{Y} = \{-1, +1\}$ empirical Rademacher complexity can be defined as

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{2} E_\sigma \left(\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma^i h\left(\mathbf{x}_i\right) \right)$$

The expression $\frac{1}{m}\sum_{i=1}^{m} \sigma^i h\left(\mathbf{x}_i\right)$ calculates the correlation between the random labels σ_i and the predictions made by a hypothesis h for the data points $\mathbf{x}_i$.
The $\sup_{h\in\mathcal{H}}$ part finds the maximum correlation over all hypotheses in $\mathcal{H}$. This represents the best fit that any hypothesis in $\mathcal{H}$ can achieve on the randomly labeled data. supremum: the least upper bound of that set.
A high Rademacher complexity indicates that the class $\mathcal{H}$ can fit the data very well, even if the data labels are just random noise. This suggests a high risk of overfitting.

Let us rewrite $\hat{\mathcal{R}}_S(\mathcal{H})$ in terms of empirical error
Note that with labels $\mathcal{Y} = \{+1, -1\}$,

$$\sigma_i h\left(\mathbf{x}_i\right) = \begin{cases} 1 & \text{if } \sigma_i = h\left(\mathbf{x}_i\right) \\ -1 & \text{if } \sigma_i \neq h\left(\mathbf{x}_i\right) \end{cases}$$

Thus

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\sigma_i h\left(\mathbf{x}_i\right) \\
&= \frac{1}{m}\left(\sum_i \mathbf{1}_{\{\mathbf{h}(\mathbf{x}_i)=\sigma_i\}} - \sum_i \mathbf{1}_{\{\mathbf{h}(\mathbf{x}_i)\neq\sigma_i\}}\right) \\
&= \frac{1}{m}\left(m - 2\sum_i \mathbf{1}_{\{\mathbf{h}(\mathbf{x}_i)\neq\sigma_i\}}\right) = 1 - 2\epsilon(\hat{h})
\end{aligned}
$$

Plug in

$$
\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{H}) &= \frac{1}{2}E_\sigma\left(\sup_{h\in\mathcal{H}}(1-2\hat{\epsilon}(h))\right) \\
&= \frac{1}{2}\left(1 - 2E_\sigma \inf_{h\in\mathcal{H}}\hat{\epsilon}(h)\right) = \frac{1}{2} - E_\sigma \inf_{h\in\mathcal{H}}\hat{\epsilon}(h)
\end{aligned}
$$

where $\epsilon(\hat{h})$ is the empirical error rate of the hypothesis h
The infimum represents the lowest possible empirical error across all hypotheses in the class $\mathcal{H}$. infimum : greatest lower bound of that set.

(Mohri et al. 2018): For any $\delta > 0$, with probability at least $1-\delta$ over a sample drawn from an unknown distribution D, for any $h \in \mathcal{H}$ we have:

$$
R(h) \leq \hat{R}_S(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}
$$

The bound is composed of the sum of :
- The empirical risk of h on the training data S (with the original labels): $\hat{R}_S(h)$
- The empirical Rademacher complexity: $\hat{\mathcal{R}}_S(\mathcal{H})$
- A term that tends to zero as a function of size of the training data as $O(1/\sqrt{m})$ assuming constant δ.

5 Rademacher vs. VC

VC dimension is independent of any training sample or distribution generating the data: it measures the worst-case where the data is generated in a bad way for the learner.
Rademacher complexity depends on the training sample thus is dependent on the data generating distribution.
VC dimension focuses the extreme case of realizing all labelings of the data.
Rademacher complexity measures smoothly the ability to realize random labelings.

Generalization bounds based on Rademacher Complexity are applicable to any binary classifiers (SVM, neural network, decision tree).
It motivates state of the art learning algoritms such as support vector machines.
But computing it might be hard, if we need to train a large number of classifiers.

VC-dim is an alternative that is usually easier to derive analytically.