PAC learning theory

Simplified PAC theory

The PAC (Probably Approximately Correct) model:

1. The data distribution is stationary:

$$x, y \sim P(x, y)$$

2. The training samples are drawn i.i.d. (independently, identically distributed)
3. The hypothesis space $\mathcal{H}$ is finite and has size $|\mathcal{H}|$
4. The error rate (error probability on a random sample) of an $h \in \mathcal{H}$ is:

$$error(h) = \sum_{all\ x,y} [y \neq h(x)]P(x, y) = \mathbb{E}_{P(x,y)}[y \neq h(x)]$$

5. We learn by choosing a $h_O \in \mathcal{H}$ that agrees with all training data
6. What is the probability that h_O has a low error rate: $error(h) < \epsilon$?

PAC intuition

- We can find a seriously wrong hypothesis by testing it against N examples. If we can't say h is bad after sufficiently many tests, it is unlikely that h is seriously wrong.
- We will then say it is probably approximately correct.

A PAC bound

$$\mathcal{H}_{good} = \{h \in \mathcal{H}: error(h) < \epsilon\}$$
$$\mathcal{H}_{bad} = \mathcal{H} \setminus \mathcal{H}_{good}$$

What is the prob. of not rejecting an $h_b \in \mathcal{H}_{bad}$?

$$error(h_b) > \epsilon$$
$$P(h_b \; correct \; on \; N \; samples) \leq (1-\epsilon)^N$$

What is the probability that there exists an $h_b \in \mathcal{H}_{bad}$ consistent with N samples?

$$P(\mathcal{H}_{bad} \; contains \; a \; consitent \; h_b) \leq$$
$$|\mathcal{H}_{bad}|(1-\epsilon)^N \leq |\mathcal{H}|(1-\epsilon)^N \leq |\mathcal{H}|e^{-N\epsilon}$$

A PAC bound

$$P(\mathcal{H}_{bad}\ contains\ a\ consitent\ h_b) \leq |\mathcal{H}|e^{-N\epsilon}$$

We want to ensure this is less than δ.

Solve:

$$|\mathcal{H}|e^{-N\epsilon} < \delta$$

For N:

$$N \geq \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln|\mathcal{H}|\right)$$

The space of all boolean functions

- There are 2^{2^n} boolean functions of n variables
- Therefore to learn a hypothesis from the space of all boolean functions of n variables we need to see $O(2^n)$ examples, or nearly all of them :(
- To learn from smaller number of examples we need to constrain our hypothesis space – e.g. consider only simple functions.

What about infinite $\mathcal{H}$?

- A naïve approach assumes that in a PC we never get an infinite number of models (floats have limited precision)
- The truly infinite case is solved by the Statistical Learning Theory (or the Vapnik-Chervonenkis, VC-theory)
- It introduces a measure of hypothesis complexity called VC-dimension
- PAC and VC theory are consistent
- If you are interested, see the book “Statistical Learning Theory” by Vladimir Vapnik.

How is regularization related to PAC

Intuitively, less parameters means smaller $|\mathcal{H}|$.

For infinite models, the VC dimension measures the hypothesis complexity.

The more regularized a model, the smaller its VC dimension.

Models with low VC dimension underfit, while those with a large VC dimenstion overfit.

Need to optimally regularize (find optimal VC dim) (This is called **structural risk minimization**)

Structural Risk Minimizaition