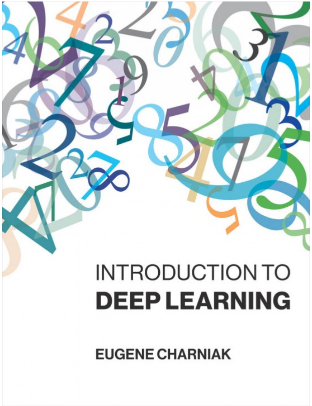
Mathematical techniques in data science

Vu Dinh

Lecture 4: Generalization bounds using covering number

February 22nd, 2019

- I and Prof. Guillot are thinking about running a seminar series on Deep Learning
- Meeting: once a week, 1 hour, starting in March



Supervised learning: standard setting

- Given: a sequence of label data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{X,Y}$
- a learning algorithm seeks a function $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space

Supervised learning: standard setting

- The function h is an element of some space of possible functions $\mathcal{H}$, usually called the *hypothesis space*
- In order to measure how well a function fits the training data, a *loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$$

  is defined

Risk and empirical risk

- With a pre-defined loss function, the "optimal hypothesis" is the minimizer over $\mathcal{H}$ of the risk function

$$R(h) = E_{(X,Y)\sim P}[L(Y, h(X))]$$

- Since P is unknown, the simplest approach is to approximate the risk function by the empirical risk

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i))$$

- The empirical risk minimizer (ERM): minimizer of the empirical risk function (in this lecture, denoted by $\hat{h}_n$)
- Let h^* denotes a minimizer of the risk function

PAC learning

Definition

The probably approximately correct (PAC) learning model typically states as follows: we say that $\hat{h}_n$ is ϵ-accurate with probability $1-\delta$ if

$$P\left[R(\hat{h}_n) - R(h^*) > \epsilon\right] < \delta.$$

In other words, we have $R(\hat{h}_n) - R(h^*) \leq \epsilon$ with probability at least $(1-\delta)$.

Exponential moment of bounded random variables

Theorem

For any random variable X, $\epsilon > 0$ and $t > 0$

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}}.$$

Theorem

If random variable X has mean zero and is bounded in $[a, b]$, then for any $s > 0$,

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

Theorem (Hoeffding's inequality)

Let $X_1, X_2, \ldots, X_n$ be i.i.d copy of a random variable $X \in [a, b]$, and $\epsilon > 0$,

$$P\left[\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X] \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Corollary:

$$P\left[\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Generalization bound for finite hypothesis space and bounded loss

Assumption

- the loss function L is bounded, that is

$$0 \leq L(y, y') \leq c \quad \forall y, y' \in \mathcal{Y}$$

- the hypothesis space is a finite set, that is

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}.$$

Key ideas

- For any $h \in \mathcal{H}$ and $\epsilon > 0$ we have

$$P[|R_n(h) - R(h)| \geq \epsilon] \leq 2 \exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

- Using a union bound on the "failure probability" associated with each hypothesis, we have

$$P[\exists h \in \mathcal{H} : |R_n(h) - R(h)| \geq \epsilon] \leq 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

Key ideas

- Using a union bound on the "failure probability" associated with each hypothesis, we have

$$P[\forall h \in \mathcal{H} : |R_n(h) - R(h)| < \epsilon]$$
$$\geq 1 - 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

- Under this "good event":

$$\begin{aligned} & R(\hat{h}_n) - R(h^*) \\ &= [R(\hat{h}_n) - R_n(\hat{h}_n)] + [R_n(\hat{h}_n) - R_n(h^*)] + [R_n(h^*) - R(h^*)] \\ &\leq 2\epsilon \end{aligned}$$

- Conclusion: $\hat{h}_n$ is (2ϵ)-accurate with probability $1 - \delta$, where

$$\delta = 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2c^2}\right)$$

PAC estimate for ERM

Theorem

For any $\delta > 0$ *and* $\epsilon > 0$*, if*

$$n \geq \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

then $\hat{h}_n$ *is* ϵ*-accurate with probability at least* $1 - \delta$*.*

PAC estimate for ERM

$$n = \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

- Fix a level of confidence δ, the accuracy ϵ of the ERM is

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\delta}\right) + \log(|\mathcal{H}|)}\right)$$

- If we want $\epsilon \to 0$ as $n \to \infty$:

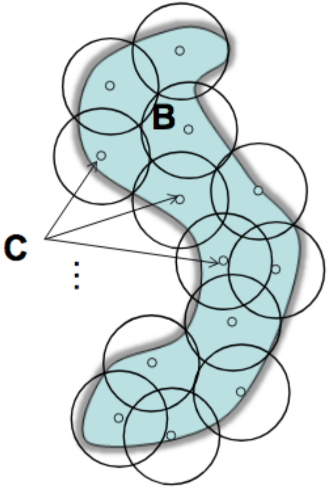
$$\log(|\mathcal{H}|) \ll n$$

- The convergence rate will not be better than $\mathcal{O}(n^{-1/2})$

Generalization bound using covering number.

Covering numbers

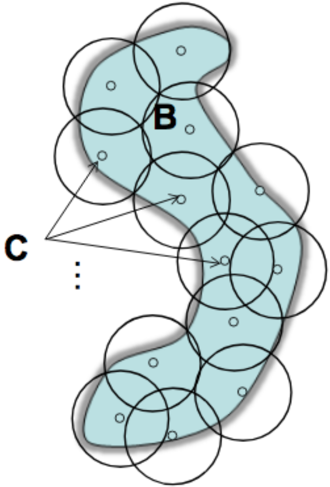
- Assumption: $\mathcal{H}$ is a metric space with distance *d* defined on it.
- For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon, \mathcal{H}, d)$ the *covering number* of $(\mathcal{H}, d)$; that is, $\mathcal{N}(\epsilon, \mathcal{H}, d)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{H}$.



Covering numbers

Remark: If $\mathcal{H}$ is a bounded $k-$dimensional manifold/algebraic surface, then we now that

$$\mathcal{N}(\epsilon, \mathcal{H}, d) = \mathcal{O}\left(\epsilon^{-k}\right)$$



Generalization bound using covering number.

- Assumption: $\mathcal{H}$ is a metric space with distance d defined on it.
- For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon, \mathcal{H}, d)$ the *covering number* of $(\mathcal{H}, d)$; that is, $\mathcal{N}(\epsilon, \mathcal{H}, d)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{H}$.
- Assumption: loss function L satisfies:

$$|L(h(x), y) - L(h'(x), y)| \leq Cd(h, h') \quad \forall, x \in \mathcal{X}; y \in \mathcal{Y}; h, h' \in \mathcal{H}$$

Key ideas

- If

$$n = \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}_\epsilon|}{\delta}\right)$$

  then the event

$$|R_n(h) - R(h)| \leq \epsilon, \forall h \in \mathcal{H}_\epsilon$$

  happens with probability at least $1 - \delta$.
- Under this event, consider any $h \in \mathcal{H}$, then there exists $h_0 \in \mathcal{H}_\epsilon$ such that $d(h, h_0) \leq \epsilon$.

Key ideas

- Since the loss function is Lipschitz

$$|R_n(h) - R_n(h_0)| \leq Cd(h, h_0)$$

  and

$$|R(h) - R(h_0)| \leq Cd(h, h_0).$$

- Conclusion:

$$|R_n(h) - R(h)| \leq (2C + 1)\epsilon \quad \forall h \in \mathcal{H}.$$

Generalization bound using covering number.

Theorem

For all $\epsilon > 0$, $\delta > 0$, *if*

$$n \geq \frac{c^2}{2\epsilon^2} \log\left(\frac{2\mathcal{N}(\epsilon, \mathcal{H}, d)}{\delta}\right)$$

then

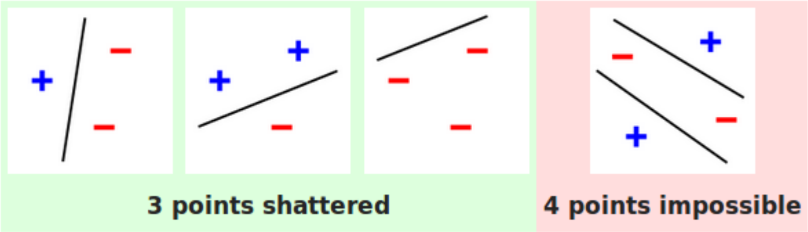
$$|R_n(h) - R(h)| \leq (2C + 1)\epsilon \quad \forall h \in \mathcal{H}.$$

with probability at least $1 - \delta$.

Example: Polynomial covering number.

- Assume that
$$\mathcal{N}(\epsilon, \mathcal{H}, d) \leq K\epsilon^{-k}$$
for some $K > 0$ and $k \geq 1$.
- $\hat{h}_n$ is ϵ-accurate with probability at least $1 - \delta$ if
$$n = \frac{c^2(4C+2)^2}{2\epsilon^2}\left(\log\left(\frac{2K}{\delta}\right) + k\log\left(\frac{4C+2}{\epsilon}\right)\right)$$
- Homework: Fix n and δ, derive an upper bound for ϵ.

Other measures of learning dimension



The set of straight lines (as a binary classification model on points) in a two-dimensional plane has VC dimension 3.

Rademacher complexity

- measures richness of a class of real-valued functions *with respect to a probability distribution*
- Given a sample $S = (x_1, x_2, \ldots, x_n)$ and a class $\mathcal{H}$ of real-valued functions defined on the input space $\mathcal{X}$, the empirical Rademacher complexity of $\mathcal{H}$ given S is defined as:

$$Rad(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right]$$

  where $\sigma_1, \sigma_2, \ldots, \sigma_m$ are independent random variables drawn from the Rademacher distribution

$$P[\sigma_i = 1] = P[\sigma_i = -1] = 1/2$$