Machine Learning

Rademacher complexity

Kristóf Karacs

Motivation

Do we have other generalization bounds beside VC?

Intuition: Growth function: worst case over all possible inputs

Goal: Grabbing the probability distribution of the input space

Idea: model complexity $\approx$ how well can the model fit to *random* data

Advantages

- Can make use of probability theory, statistics and calculus instead of combinatorial analysis
- Can be computed for more function classes than the VC-dimension (linear classes, polygons, etc.)

Empirical Rademacher complexity

Definition

Let $x_1, \ldots, x_N$ be the dataset (drawn i.i.d. from the data distribution $\mathcal{D}$), $\mathcal{F}$ a function class.
Then the *empirical Rademacher complexity* of $\mathcal{F}$ is:

$$\widehat{Rad}_N(\mathcal{F}) = \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{F}} \left[\frac{1}{N} \sum_{i=1}^{N} \sigma_i f(x_i) \right] \right\},$$

where $\sigma_1, \ldots, \sigma_N \in \{-1, 1\}$ are independent, uniform random variables.

Rademacher complexity

Definition

The *Rademacher complexity* of a function class $\mathcal{F}$ is defined as

$$Rad_N(\mathcal{F}) = \mathbb{E}_{\mathcal{D}}\left\{\widehat{Rad}_N(\mathcal{F})\right\},$$

where $\mathcal{D}$ is the data distribution.

Intuition behind Rademacher complexity

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

- Consider the correlation (cosine distance) between $f(x_i)$ and σ_i
- Take the maximum of this correlation over all $f \in \mathcal{F}$.
- Take the expectation: measure the ability of hypotheses from $\mathcal{F}$ to fit random noise.

Rademacher bound

Theorem

Fix a parameter $\delta \in (0,1)$*. Then with probability at least* $1-\delta$*,*

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{f(x)\right\} \leq \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right) + 2Rad_N(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{N}},$$

for all $f \in \mathcal{F}$*.*

In addition, with probability at least $1-\delta$*,*

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{f(x)\right\} \leq \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right) + 2\widehat{Rad}_N(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{N}},$$

for all $f \in \mathcal{F}$*.*

Application for loss

Compose the model and loss:
$f(x) = L(h(x), x)$,
where $h \in \mathcal{H}, h : X \to \{-1, 1\}$ and $L : \{-1, 1\} \to \mathbb{R}$

Then
$E_{out} = \mathbb{E}_{\mathcal{D}} \{f(x)\}$
and
$E_{in} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$

using the Rademacher bound:

$$E_{out} \leq E_{in} + 2Rad_N(\mathcal{L}(\mathcal{H})) + \sqrt{\frac{1/\delta}{N}},$$

where $\mathcal{L}(\mathcal{H})$ is the function class of the loss and hypothesis combined.

Application for loss

In general:

$$E_{out} \leq E_{in} + 2Rad_N(\mathcal{L}(\mathcal{H})) + \sqrt{\frac{1/\delta}{N}},$$

Assuming 1-0 error:

$$E_{out} \leq E_{in} + Rad_N(\mathcal{H}) + \sqrt{\frac{1/\delta}{N}}$$