Lecture 14: Generalization

Assumptions:
- We have given that we use a model that is not overfitting.
- In particular, we can write a good model that is not overfitting.

Generalization:
- The model is overfitting, any prediction $x = y$ is chosen by having algorithm $\mathcal{A}$.
- $R(x) - R(y) = R(x) - R(x' | x = y) 
  \begin{cases} + 2 \mathbb{E}_x [C(x', y | x = y)] - C(y | x = y) + C(x | x = y) \\
  = 2 \mathbb{E}_x [C(x', y | x = y)] \\
  = 2 \mathbb{E}_x [C(x | x = y)]
  \end{cases}$

- How things to discern better with:
  1. Why is it possible?
  2. Is this a good model?

- Only in it possible
  - $\frac{1}{2} \mathbb{E}_x [C(x', y)]$
    - $\mathbb{E}_x$ of both
  - $C(x', y)$
    - Note we divide $C(x', y)$ by $Z$: $Z$.
    - $R(x) \leq C(x', y) \Rightarrow C(x, y) \leq \frac{1}{2} Z$.

- Match functions

- The bounds depend on how good $R(x)$ is.

My opinion: how to $(10^{-10})$ become success with statistical complexity.

Some papers claim learning is inherent.

- $\mathbb{E}_x$ has a good predictor $R(x)$ we assume $x$.
- $\mathbb{E}_x$ does not depend on

- $R(x)$ does not depend on $\epsilon$.

The whole is wrong

$R(x')$ due to $R(x')$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

If $R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.

$R(x)$ due to $R(x)$ high-growth.

$\epsilon$ high-growth.
Definition of measure

\[ \| \mathbf{X} \|_2 \leq \varepsilon \implies \mathbb{P}(\mathbf{X} \neq \mathbf{0}) \leq \varepsilon \]

Lemma (Markov)

\[ \mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)] \]

Proof. Note that \( f(x) = x \) for any \( x \in \mathbb{R} \).

Guarantee: \( \mathbb{E}[X] \leq \mathbb{E}[Y] \)

Example: Let \( X \) be a random variable. Then \( \mathbb{E}[X] \leq \mathbb{E}[Y] \).

\[ \mathbb{P}[X = 0] \leq \varepsilon \]

Example: Let \( X \) be a random variable. Then \( \mathbb{P}[X = 0] \leq \varepsilon \).

Theorem (Markov's inequality)

\[ \mathbb{P}[X > c] \leq \frac{\mathbb{E}[X]}{c} \]

Consider \( f(x) = e^{-x} \).

\[ \mathbb{E}[f(X)] = \mathbb{E}[e^{-X}] \]

Suppose \( X \) is a non-negative random variable.

\[ \mathbb{E}[e^{-X}] = \mathbb{E}[e^{-X}] \]

"Character-based" method

Based on the boundedness of \( X \).

References

Figures from Latex of presentation

(From Walter's, Introduction to Measure Theory)
looseness & “overhanded papers”

Nagarajan & Kolter paper gives a simple prediction problem where one can compute a good generalization bound directly, but uniform convergence gives a bound which never converges to 0 with $n \to \infty$

"fix" use diff algorithm, e.g. GD on logistic loss to select max margin solution, then apply uniform convergence

(There's also an earlier history of things we can't prove w/ uniform convergence)