PAC Learning

Lekshmi Ramesh

Indian Institute of Science
Bangalore

October 26, 2019

PAC Learning: Introduction

- Probably Approximately Correct learning: A model of learning

PAC Learning: Introduction

- Probably Approximately Correct learning: A model of learning
- Seeks to find algorithms that try to learn a concept (e.g. classifying emails as spam/not spam) from labeled examples

PAC Learning: Introduction

- Probably Approximately Correct learning: A model of learning
- Seeks to find algorithms that try to learn a concept (e.g. classifying emails as spam/not spam) from labeled examples
- Goal of the algorithm is to approximate the true concept with high probability over training samples

PAC Learning: Some terminology

- Input space/Domain: The set $\mathcal{X}$ of objects that we wish to label/classify

PAC Learning: Some terminology

- Input space/Domain: The set $\mathcal{X}$ of objects that we wish to label/classify
- Output space/Label set: A set $\mathcal{Y}$ that describes the labels

PAC Learning: Some terminology

- Input space/Domain: The set $\mathcal{X}$ of objects that we wish to label/classify
- Output space/Label set: A set $\mathcal{Y}$ that describes the labels
- Data generation process: An unknown distribution D on $\mathcal{X}$ generates sample x which is then labeled by the "true" labeling function f to get label $y = f(x)$

PAC Learning: Some terminology

- Input space/Domain: The set $\mathcal{X}$ of objects that we wish to label/classify
- Output space/Label set: A set $\mathcal{Y}$ that describes the labels
- Data generation process: An unknown distribution D on $\mathcal{X}$ generates sample x which is then labeled by the "true" labeling function f to get label $y = f(x)$
- Training data: The input S to the learning algorithm consisting of iid samples $\{(x_i, y_i)\}_{i=1}^{n}$ from D_n

PAC Learning: Some terminology

- Concept class: The set $\mathcal{C}$ of all target functions $f : \mathcal{X} \to \mathcal{Y}$ that could have generated the samples

PAC Learning: Some terminology

- Concept class: The set $\mathcal{C}$ of all target functions $f : \mathcal{X} \to \mathcal{Y}$ that could have generated the samples
- Hypothesis class: The set $\mathcal{H}$ of all possible hypotheses $h : \mathcal{X} \to \mathcal{Y}$ that the algorithm can choose from
  e.g. linear classifiers parameterized by θ:
  $\mathcal{H} = \{h_\theta : h_\theta(x) = 1_{\{\theta^\top x \geq 0\}}\}$

Assessing a learning algorithm

- Generalization error/True error of h

$$\begin{aligned} L_{D,f}(h) &= \mathrm{P}_{x\sim D}(h(x) \neq f(x)) \\ &= D(\{x : h(x) \neq f(x)\}) \end{aligned}$$

Cannot compute since D unknown

Assessing a learning algorithm

- Generalization error/True error of h

$$
\begin{aligned}
L_{D,f}(h) &= \mathrm{P}_{x\sim D}(h(x) \neq f(x)) \\
&= D(\{x : h(x) \neq f(x)\})
\end{aligned}
$$

  Cannot compute since D unknown

- Training error/empirical risk as

$$
L_S(h) = \frac{1}{|S|} \sum_{i \in S} 1_{\{h(x_i) \neq y_i\}}
$$

  This is the fraction of samples misclassified by the algorithm

An example

- Let $\mathcal{X} = \{0,1\}^n$ and $D = \mathrm{Unif}(\mathcal{X})$

An example

- Let $\mathcal{X} = \{0, 1\}^n$ and $D = \mathrm{Unif}(\mathcal{X})$
- Let the true labeling function be f and $f(x) \sim \mathrm{Ber}(1/2)$ for all $x \in \mathcal{X}$

An example

- Let $\mathcal{X} = \{0,1\}^n$ and $D = \text{Unif}(\mathcal{X})$
- Let the true labeling function be f and $f(x) \sim \text{Ber}(1/2)$ for all $x \in \mathcal{X}$
- Given training data $S = \{x_i, y_i\}_{i=1}^m$, the algorithm outputs hypothesis h where

$$h(x) = \begin{cases} y_i, & \text{if } x = x_i \text{ for some } i \in S \\ Z, & \text{otherwise} \end{cases}$$

where $Z \sim \text{Ber}(1/2)$

An example

- Let $\mathcal{X} = \{0,1\}^n$ and $D = \text{Unif}(\mathcal{X})$
- Let the true labeling function be f and $f(x) \sim \text{Ber}(1/2)$ for all $x \in \mathcal{X}$
- Given training data $S = \{x_i, y_i\}_{i=1}^m$, the algorithm outputs hypothesis h where

$$h(x) = \begin{cases} y_i, & \text{if } x = x_i \text{ for some } i \in S \\ Z, & \text{otherwise} \end{cases}$$

  where $Z \sim \text{Ber}(1/2)$
- What is the training error and true error of this hypothesis?

An example

- On the training set, there is no error

An example

- On the training set, there is no error
- True error

$$
\begin{aligned}
L_S(h) =& \Pr\left(h(x) \neq f(x) | x \in S\right) \Pr\left(S\right) \\
&+ \Pr\left(h(x) \neq f(x) | x \in \mathcal{X} \backslash S\right) \Pr\left(\mathcal{X} \backslash S\right) \\
&= \frac{1}{2}\left(\frac{2^n - m}{2^n}\right)
\end{aligned}
$$

An example

- On the training set, there is no error
- True error

$$\begin{aligned} L_S(h) = & \Pr\left(h(x) \neq f(x) | x \in S\right) \Pr\left(S\right) \\ & + \Pr\left(h(x) \neq f(x) | x \in \mathcal{X} \backslash S\right) \Pr\left(\mathcal{X} \backslash S\right) \\ & = \frac{1}{2}\left(\frac{2^n - m}{2^n}\right) \end{aligned}$$

- For m small, this can be bad

An example

- On the training set, there is no error
- True error

$$\begin{aligned} L_S(h) =& \Pr\left(h(x) \neq f(x) | x \in S\right) \Pr\left(S\right) \\ &+ \Pr\left(h(x) \neq f(x) | x \in \mathcal{X} \setminus S\right) \Pr\left(\mathcal{X} \setminus S\right) \\ &= \frac{1}{2}\left(\frac{2^n - m}{2^n}\right) \end{aligned}$$

- For m small, this can be bad
- To avoid overfitting, we restrict the search space to certain hypothesis classes: Empirical risk minimization

$$\arg\min_{h \in \mathcal{H}} L_S(h)$$

PAC learnability

- A concept class $\mathcal{C}$ is PAC learnable using hypothesis class $\mathcal{H}$ if, for all $f \in \mathcal{C}$, $\varepsilon > 0$, $\delta > 0$ and all distributions D_n, there exists an algorithm that produces hypothesis $h \in \mathcal{H}$ such that the following holds:

$$\mathrm{P}_{S \in D_n}(L_{D,f}(h) \geq \varepsilon) \leq \delta.$$

PAC learnability

- A concept class $\mathcal{C}$ is PAC learnable using hypothesis class $\mathcal{H}$ if, for all $f \in \mathcal{C}$, $\varepsilon > 0$, $\delta > 0$ and all distributions D_n, there exists an algorithm that produces hypothesis $h \in \mathcal{H}$ such that the following holds:

$$\mathrm{P}_{S \in D_n}(L_{D,f}(h) \geq \varepsilon) \leq \delta.$$

- True error of the hypothesis is small with high probability

Questions

- Generalization bounds
  Can we bound the true error of a hypothesis given its training error?

Questions

- Generalization bounds
  Can we bound the true error of a hypothesis given its training error?

- PAC learnability
  Can we find hypotheses that have small true error with high probability?

Questions

- Generalization bounds
  Can we bound the true error of a hypothesis given its training error?

- PAC learnability
  Can we find hypotheses that have small true error with high probability?

- Sample complexity
  How many training samples are needed for PAC learnability (for a given concept)?

Learning an interval

- Let $\mathcal{X} = \mathbb{R}$ and the true concept be $f_a(x) = \mathbb{1}_{\{0 \le x \le a\}}$ for some $a \in \mathbb{R}$

Learning an interval

- Let $\mathcal{X} = \mathbb{R}$ and the true concept be $f_a(x) = \mathbb{1}_{\{0 \le x \le a\}}$ for some $a \in \mathbb{R}$
- Given $\{x_i, y_i\}_{i=1}^n$, let

$$\hat{a} = \max_{i:\ y_i=1} x_i$$

Learning an interval

- Let $\mathcal{X} = \mathbb{R}$ and the true concept be $f_a(x) = \mathbb{1}_{\{0 \leq x \leq a\}}$ for some $a \in \mathbb{R}$
- Given $\{x_i, y_i\}_{i=1}^n$, let
$$\hat{a} = \max_{i:\ y_i = 1} x_i$$
- Let $[c, a]$ be an interval that has probability ε. Our hypothesis $h_{\hat{a}}(x) = \mathbb{1}_{\{0 \leq x \leq \hat{a}\}}$ has true error less than ε if we see at least on example from $[c, a]$

Learning an interval

- Thus, $L_S(h_{\hat{a}}) \geq \varepsilon$ if all n examples lie outside $[c, a]$

$$\Pr\left(L_S(h_{\hat{a}}) \geq \varepsilon\right) = (1-\varepsilon)^n \leq e^{-n\varepsilon}$$

Learning an interval

- Thus, $L_S(h_{\hat{a}}) \geq \varepsilon$ if all n examples lie outside $[c, a]$

$$\Pr\left(L_S(h_{\hat{a}}) \geq \varepsilon\right) = (1-\varepsilon)^n \leq e^{-n\varepsilon}$$

- This can be made small for sufficiently large n

$$n \geq \frac{1}{\varepsilon} \ln \frac{1}{\delta}$$

PAC learnability for finite hypothesis space

- Realizability: For some $h \in \mathcal{H}$, $L_{D,f}(h) = 0$

PAC learnability for finite hypothesis space

- Realizability: For some $h \in \mathcal{H}$, $L_{D,f}(h) = 0$
- h_S is the output of ERM

PAC learnability for finite hypothesis space

- Realizability: For some $h \in \mathcal{H}$, $L_{D,f}(h) = 0$
- h_S is the output of ERM
- Let $\mathcal{H}$ be a finite hypothesis class, $0 < \delta, \varepsilon < 1$ and

$$n \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}.$$

Then, for any concept f and any distribution D for which realizability holds, with probability $1 - \delta$ over training samples S of size n, it holds that

$$L_{D,f}(h_S) \leq \varepsilon.$$

References

- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.*
- Andrew Ng. *Learning theory, CS229 Lecture notes.*

Thank you