Learning Theory

Aarti Singh

Machine Learning 10-315
Nov 30, 2020

Slides courtesy: Carlos Guestrin

Learning Theory

- We have explored **many** ways of learning from data
- But…
  - Can we certify how good is our classifier, really? error rate / accuracy
  - How much data do I need to make it "good enough"?

A simple setting

- Classification
  - m i.i.d. data points
  - **Finite** number of possible classifiers in model class (e.g., dec. trees of depth d)
- Lets consider that a learner finds a classifier $\hat{h}$ that gets zero error in training
  - $\text{error}_{train}(\hat{h}) = 0$

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\hat{h}(x_i) \neq y_i}$$

- What is the probability that $\hat{h}$ has more than ε true (= test) error?
  - $\text{error}_{true}(\hat{h}) \geq \varepsilon$

$$P(\hat{h}(X) \neq Y) = \mathbb{E}[1_{\hat{h}(X) \neq Y}]$$

Even if *h* makes zero errors in training data, may make errors in test

How likely is a bad classifier to get m data points right?

- Consider a bad classifier h i.e. $\text{error}_{\text{true}}(h) \geq \varepsilon$

- Probability that h gets one data point right

$$\leq 1 - \varepsilon$$

- Probability that h gets m data points right

P(1st data pt right
∧ 2nd . ——
∧ . — —)
$= \prod_{i=1}^{m} P(x_i \text{ right})$

$$\leq (1-\varepsilon)^m$$

"
P(h correctly classifies all data pts)
= P(h has 0 training error)

How likely is a learner to pick a bad classifier?

- Usually there are many (say k) bad classifiers in model class

  $h_1, h_2, \ldots, h_k$ s.t. $\text{error}_{true}(h_i) \geq \varepsilon \quad i = 1, \ldots, k$

- Probability that learner picks a bad classifier = Probability that some bad classifier gets 0 training error

  Prob(h_1 gets 0 training error OR
  h_2 gets 0 training error OR … OR
  h_k gets 0 training error)

  $P(A \cup B) \leq P(A) + P(B)$
  OR

  $\leq$ Prob(h_1 gets 0 training error) +
  Prob(h_2 gets 0 training error) + … +
  Prob(h_k gets 0 training error) $\leq k(1-\varepsilon)^m$

  $\leq \; k\,(1-\varepsilon)^m$

Union bound

Loose but works

How likely is a learner to pick a bad classifier?

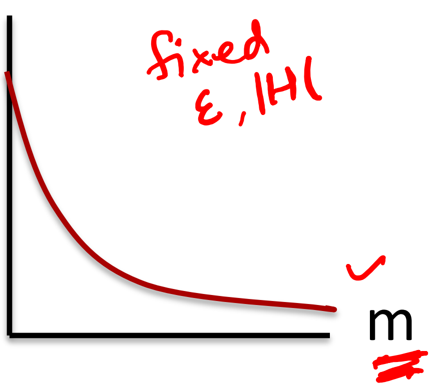
- Usually there are many many (say k) bad classifiers in the class

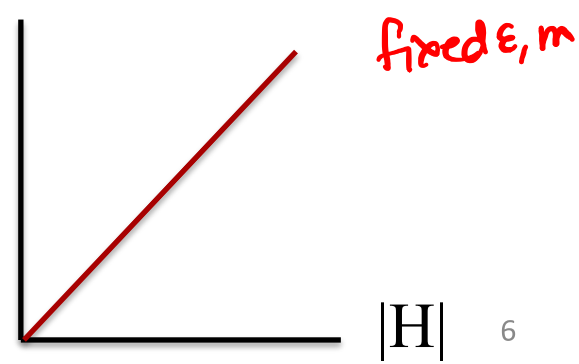
$$h_1, h_2, \ldots, h_k \qquad \text{s.t. } error_{true}(h_i) \geq \varepsilon \quad i = 1, \ldots, k$$

- Probability that learner picks a bad classifier

$$\leq \; k\,(1-\varepsilon)^m \;\leq\; |H|\,(1-\varepsilon)^m \leq\; |H|\, e^{-\varepsilon m}$$

$|H|$ → Size of model class





PAC (Probably Approximately Correct) bound

- ***Theorem [Haussler'88]***: Model class *H* finite, dataset *D* with *m* i.i.d. samples, 0 < ε < 1 : for any learned classifier $\hat{h}$ that gets 0 training error:

$$P(\text{error}_{true}(\hat{h}) \geq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(\hat{h}) \leq \epsilon$$

Important: PAC bound holds for all *h* with 0 training error, but doesn't guarantee that algorithm finds best *h*!!!

Using a PAC bound

$$\overbrace{|H|e^{-m\epsilon}}^{\le \delta} \le \delta$$

- Given ε and δ, yields sample complexity

#training data, $m \ge \dfrac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$

$\frac{|H|}{\delta} \le e^{m\varepsilon}$

- Given m and δ, yields error bound

error, $\epsilon \ge \dfrac{\ln|H| + \ln\frac{1}{\delta}}{m}$

Limitations of Haussler's bound

- Only consider classifiers with 0 training error

  h such that zero error in training, $\text{error}_{\text{train}}(h) = 0$

- Dependence on size of model class |H|

$$m \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$$

  what if |H| too big or H is continuous (e.g. linear classifiers)?

PAC bounds for finite model classes

H - Finite model class

e.g. decision trees of depth k

histogram classifiers with binwidth h

With probability ≥ 1-δ,

1) For all h $\in$ H s.t. $\text{error}_{\text{train}}(h) = 0$,

$$\text{error}_{\text{true}}(h) \leq \varepsilon = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

$\boxed{\text{Haussler's bound}}$

What if our classifier does not have zero error on the training data?

- A learner with zero training errors may make mistakes in test set
- What about a learner with $error_{train}(h) \neq 0$ in training set?
- The error of a classifier is like estimating the parameter of a coin!

$E[\mathbf{1}_{h(x)\neq y}]$

$$error_{true}(h) := P(h(X) \neq Y) \equiv P(H=1) =: \theta$$

$$error_{train}(h) := \frac{1}{m}\sum_i \mathbf{1}_{h(X_i)\neq Y_i} \equiv \frac{1}{m}\sum_i Z_i =: \widehat{\theta}$$

Hoeffding's bound for a single classifier

- Consider *m* i.i.d. flips $x_1,\ldots,x_m$, where $x_i \in \{0,1\}$ of a coin with parameter θ. For $0<\varepsilon<1$:

$$P\left(\left|\theta - \frac{1}{m}\sum_i x_i\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

$e^{-m\varepsilon}$

- For a single classifier h

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

Hoeffding's bound for |H| classifiers

- For each classifier h_i:

$$P\left(|\text{error}_{true}(h_i) - \text{error}_{train}(h_i)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- What if we are comparing |H| classifiers?

  Union bound

- ***Theorem***: Model class *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned classifier $h \in H$:

$\forall h \in H$ $\quad P(A \cup B) \leq P(A) + P(B)$

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

Important: PAC bound holds for all *h*, but doesn't guarantee that algorithm finds best *h*!!!

Summary of PAC bounds for finite model classes

$2|H|e^{-2m\varepsilon^2} \le \delta$

$\frac{2|H|}{\delta} \le e^{2m\varepsilon^2}$

With probability ≥ 1-δ,

1) For all h ∈ H s.t. $\text{error}_{\text{train}}(h) = 0$,

$$\text{error}_{\text{true}}(h) \le \varepsilon = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Haussler's bound

model complexity

$|H|$

2) For all h ∈ H

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \le \varepsilon = \sqrt{\frac{\ln 2|H| + \ln\frac{1}{\delta}}{2m}}$$

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$\forall h \in H$

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

- Fixed m

Model class		
complex	small	large
simple	large	small

What about the size of the model class?

$$2|H|e^{-2m\epsilon^2} \leq \delta$$

- Sample complexity

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

- How large is the model class?

$|H|$

Number of decision trees of depth k

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

Recursive solution:

Given *n* **binary** attributes

H_k = Number of **binary** decision trees of depth k

$H_0 = 2$

$H_k =$ (#choices of root attribute) ← n

*(# possible left subtrees) ← H_{k-1}

*(# possible right subtrees) $= n * H_{k-1} * H_{k-1}$

↑H_{k-1}

$\log_2 \quad = \log_2 n + 2 \underbrace{\log_2 H_{k-1}}_{L_{k-1}}$

Write $L_k = \log_2 H_k$

$L_0 = 1$

$L_k = \log_2 n + 2L_{k-1} = \log_2 n + 2(\log_2 n + 2L_{k-2})$

$= \log_2 n + 2\log_2 n + 2^2\log_2 n + \ldots + 2^{k-1}(\log_2 n + 2L_0)$

So $L_k = (2^k-1)(1+\log_2 n) + 1$

PAC bound for decision trees of depth k

$$m \geq \frac{\ln 2}{2\epsilon^2}\left(\underline{(2^k - 1)(1 + \log_2 n) + 1} + \log_2 \frac{2}{\delta}\right)$$

- Bad!!!
  - Number of points is exponential in depth k!

- But, for <u>*m* data points</u>, decision tree can't get too big…

Number of leaves never more than number data points

Number of decision trees with k leaves

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

H_k = Number of binary decision trees with k leaves

H_1 =2

H_k = (#choices of root attribute $\overset{n}{)}$ *

[(# left subtrees wth 1 $\overset{H_1}{\text{leaf}}$)*(# right $\overset{H_{k-1}}{\text{subtrees}}$ wth k-1 leaves)

+ (# left subtrees wth 2 $\overset{H_2}{\text{leaves}}$)*(# right $\overset{H_{k-2}}{\text{subtrees}}$ wth k-2 leaves)

+ ...

+ (# left subtrees wth $\overset{H_{k-1}}{\text{k-1}}$ leaves)*(# right $\overset{H_1}{\text{subtrees}}$ wth 1 leaf)]

$$H_k = n\sum_{i=1}^{k-1} H_i H_{k-i} = n^{k-1}\, C_{k-1}$$

(C_{k-1} : Catalan Number)

Loose bound (using Sterling's approximation):

depth k

$\log_2 H_k \sim 2^k$

$$\log_2 \quad H_k \leq \underbrace{n^{k-1}}_{\log_2} 2^{2k-1}$$

Number of decision trees

- With k leaves

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

$$\log_2 H_k \leq (k-1)\log_2 n + 2k - 1$$

linear in k

number of points m is linear in #leaves

- With depth k

$$\log_2 H_k = (2^k - 1)(1 + \log_2 n) + 1$$

exponential in k

number of points m is exponential in depth

PAC bound for decision trees with k leaves – Bias-Variance revisited

With prob $\underline{\geq 1\text{-}\delta}$ $\quad \text{error}_{true}(h) \leq \text{error}_{train}(h) + \underbrace{\sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}}_{\varepsilon}$

With $H_k \leq n^{k-1}2^{2k-1}$, we get

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{(k \underline{-1})\ln n + (2k-1)\ln 2 + \ln\frac{2}{\delta}}{2m}}$$

k = m	0	- large (~ > ½)
k < m	>0	= small (~ <½)

What did we learn from decision trees?

- Moral of the story:

  Complexity of learning not measured in terms of size of model space, but in maximum *number of points* that allows consistent classification

Summary of PAC bounds for finite model class

With probability $\geq 1-\delta$,

1) For all $h \in H$ s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln \boxed{|H|} + \ln \frac{1}{\delta}}{m}$$

Haussler's bound

2) For all $h \in H$

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

Hoeffding's bound