

Learning Theory

Aarti Singh

Machine Learning 10-315
Nov 22, 2021

Slides courtesy: Carlos Guestrin



MACHINE LEARNING DEPARTMENT



Learning Theory

- We have explored **many** ways of learning from data
- But...
 - Can we certify how good is our classifier, really?
 - How much data do I need to make it “good enough”?

PAC Learnability

- True function space, F
- Model space, H

F is **PAC Learnable** by a learner using H if

there exists a learning algorithm s.t. for all functions in H , for all distributions over inputs, for all $0 < \epsilon, \delta < 1$, with probability $> 1 - \delta$, the algorithm outputs a model $h \in H$ s.t. $\text{error}_{\text{true}}(h) \leq \epsilon$

in time and samples that are polynomial in $1/\epsilon, 1/\delta$ and n .

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 h that gets zero error in training
 - $\text{error}_{\text{train}}(h) = 0$
- What is the probability that h has more than ε true (= test) error?
 - $\text{error}_{\text{true}}(h) \geq \varepsilon$

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

How likely is a learner to pick a bad classifier?

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

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

- 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)

$\leq$ Prob(h_1 gets 0 training error) +

Prob(h_2 gets 0 training error) + ... +

Prob(h_k gets 0 training error)

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

Union
bound

Loose but
works

How likely is a learner to pick a bad classifier?

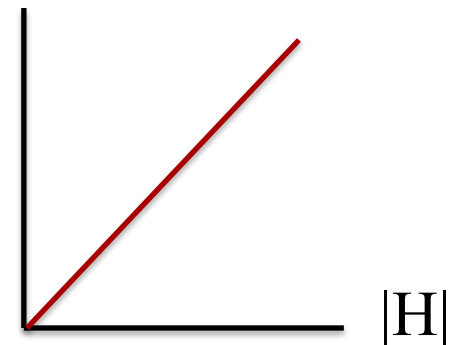
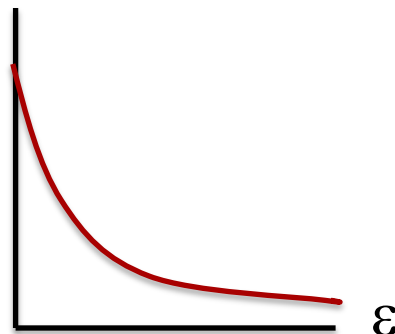
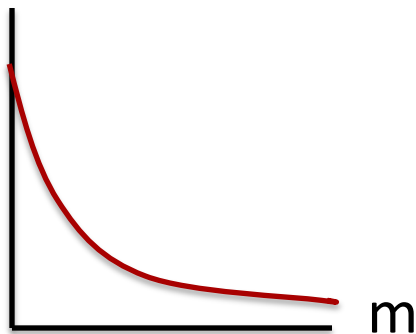
- Usually there are many many (say k) bad classifiers in the class

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

- Probability that learner picks a bad classifier

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

↙ Size of model class



PAC (Probably Approximately Correct) bound

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

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

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

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

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

Using a PAC bound

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

- Given ϵ and δ , yields sample complexity

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

- Given m and δ , yields error bound

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

Poll

Assume m is the minimum number of training examples sufficient to guarantee that with probability $1 - \delta$ a consistent learner using model class H will output a classifier with true error at worst ϵ .

Then a second learner that uses model space H' will require $2m$ training examples (to make the same guarantee) if $|H'| = 2|H|$.

A. True B. False

If we double the number of training examples to $2m$, the error bound ϵ will be halved.

C. True D. False

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 $\geq 1-\delta$,

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

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

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!

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

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

Hoeffding's bound for a single classifier

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

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

- Central limit theorem:

Hoeffding's bound for a single classifier

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

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

- For a single classifier h

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

Hoeffding's bound for $|H|$ classifiers

- For each classifier h_i :

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

- What if we are comparing $|H|$ classifiers?

Union bound

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

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

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

Summary of PAC bounds for finite model classes

With probability $\geq 1-\delta$,

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

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

Haussler's bound

2) For all $h \in H$

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

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

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

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

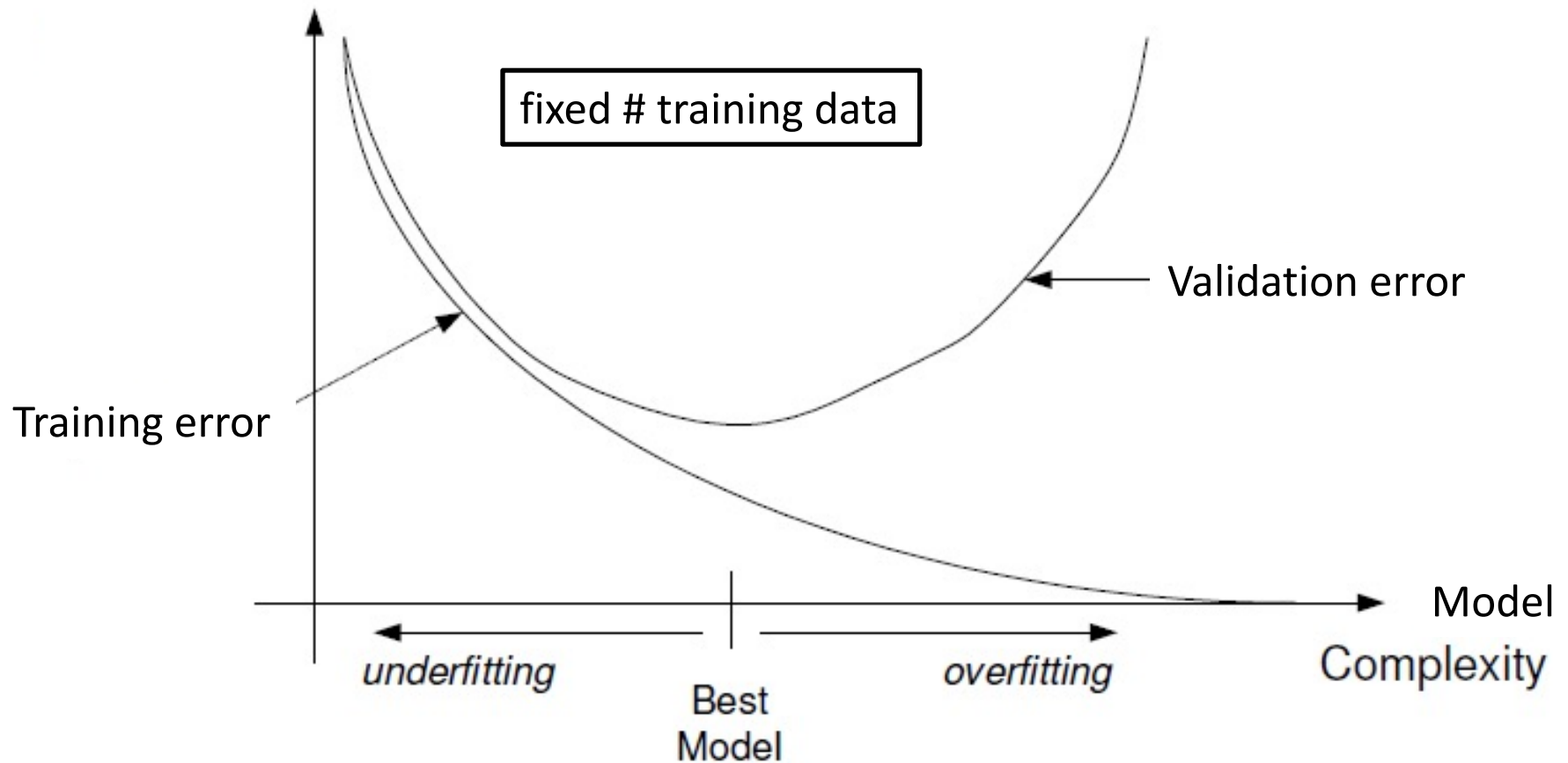
$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

- Fixed m

Model class		
complex	small	large
simple	large	small

Training vs. Test Error

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



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?

Number of decision trees of depth k

Recursive solution:

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

Given n **binary** attributes

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

$$H_0 = 2$$

H_k = (#choices of root attribute)

*(# possible left subtrees)

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

Write $L_k = \log_2 H_k$

$$L_0 = 1$$

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

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

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

PAC bound for decision trees of depth k

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

- Bad!!!
 - Number of points is exponential in depth k !
- But, for m data points, 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

$$H_1 = 2$$

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

$$\begin{aligned} & [(\text{\# left subtrees with 1 leaf}) * (\text{\# right subtrees with k-1 leaves}) \\ & + (\text{\# left subtrees with 2 leaves}) * (\text{\# right subtrees with k-2 leaves}) \\ & + \dots \\ & + (\text{\# left subtrees with k-1 leaves}) * (\text{\# right subtrees with 1 leaf})] \end{aligned}$$

$$H_k = n \sum_{i=1}^{k-1} H_i H_{k-i} = n^{k-1} C_{k-1} \quad (C_{k-1} : \text{Catalan Number})$$

Loose bound (using Sterling's approximation):

$$H_k \leq n^{k-1} 2^{2k-1}$$

Number of decision trees

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

$$\log_2 H_k \leq (k - 1) \log_2 n + 2k - 1 \quad \text{linear in } k$$

number of points m is linear in #leaves

- With depth k

$$\log_2 H_k = (2^k - 1)(1 + \log_2 n) + 1 \quad \text{exponential in } k$$

number of points m is exponential in depth

What did we learn from decision trees?

- Moral of the story:

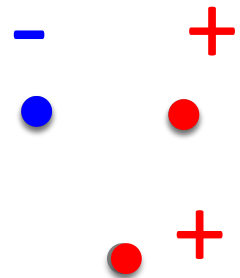
Complexity of learning not measured in terms of size of model space, but in maximum *number of points* that allows consistent classification

Rademacher Complexity

- Instead of all possible labelings, measure complexity by how accurately a model space can match a random labeling of the data.

For each data point i , draw random label

$$\sigma_i \quad \text{s.t.} \quad P(\sigma_i = +1) = \frac{1}{2} = P(\sigma_i = -1)$$



Then empirical Rademacher complexity of H is

$$\hat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i h(X_i) \right) \right]$$

Finite model class

- Rademacher complexity can be upper bounded in terms of model class size $|H|$:

$$\hat{R}_m(H) \leq \sqrt{\frac{2 \ln |H|}{m}}$$

- Often Rademacher bounds are significantly better