Learning Theory

Aarti Singh & Geoff Gordon

Machine Learning 10-701
Apr 14, 2021

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?
  - 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 *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

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

- 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)

  ≤ Prob(h_1 gets 0 training error) +

  Prob(h_2 gets 0 training error) + … +

  Prob(h_k gets 0 training error)

  ≤ $k\,(1-\varepsilon)^m$

Union bound

Loose but works

How likely is a learner to pick a bad classifier?

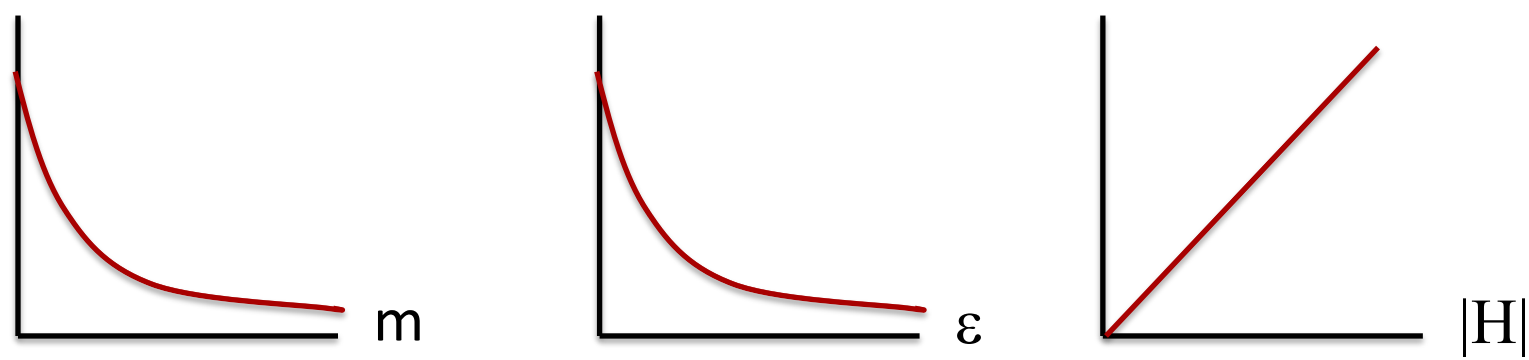
- Usually there are many many (say k) bad classifiers in the class

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

- Probability that learner picks a bad classifier

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

  Size of model class



PAC (Probably Approximately Correct) bound

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

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

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

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

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

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

- Given m and δ, yields error bound

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

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)?

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!

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

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

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 $0<\varepsilon<1$:

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

- Central limit theorem:

$Z_1 \ldots Z_m$ iid $\quad E[Z_i] = \mu$, $var(Z_i) = \sigma^2$

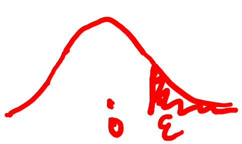
$X_i \sim Ber(\theta)$

$E[X_i] = \theta$

$var(X_i) = \theta(1-\theta) \leq \frac{1}{4}$

$$\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right) \longrightarrow N(0,\sigma^2)$$

$$\frac{1}{m}\sum_{i=1}^{m} X_i - \theta \rightarrow N\left(0,\frac{\sigma^2}{m}\right) \rightarrow N\left(0,\frac{1}{4m}\right)$$



$$e^{-\varepsilon^2/2(\sigma^2/m)} = e^{-2m\varepsilon^2}$$

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 $0<\varepsilon<1$:

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

- 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 < ε < 1 : for any learned classifier $h \in H$:

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

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