Statistical Learning theory: PAC Learning

Machine Learning Theory (MLT)

Edinburgh

Rik Sarkar

Today’s topic

- ML models are trained using random samples
- How many random samples do we need to train a good model?
- On what quantities does this number depend? What is the formula?

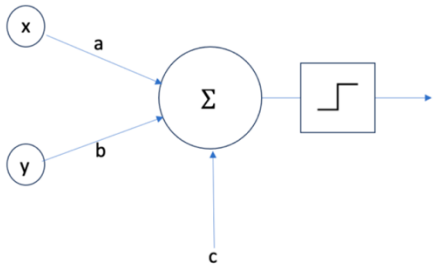
What is a good model?

- Given a sample S, there is always perfect model for S!

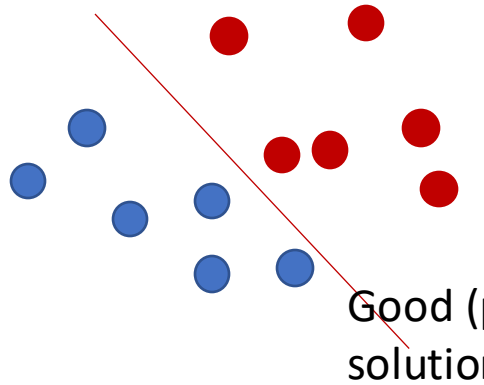
- What is the trivial way to build a perfect model, and why don’t we use it?

Models in a hypothesis class

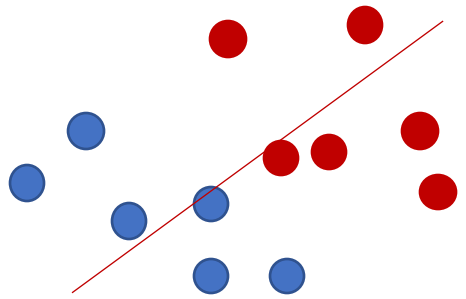
- We use models from a hypothesis class $\mathcal{H}$
- E.g. Linear separators (a, b, c)
  - Representing $ax + by + c \geq 0$



- There are an infinite number of lines $|\mathcal{H}| = \infty$
  - Perfect solution may not be always easy

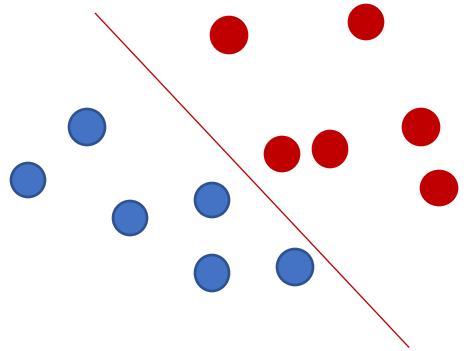


Good (perfect) solution

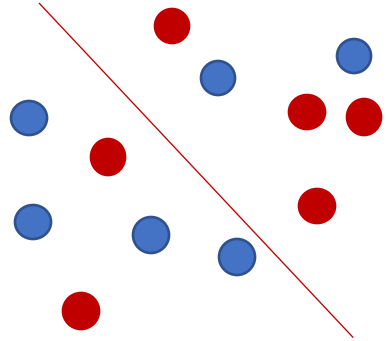


Bad solution

Realizable and agnostic case



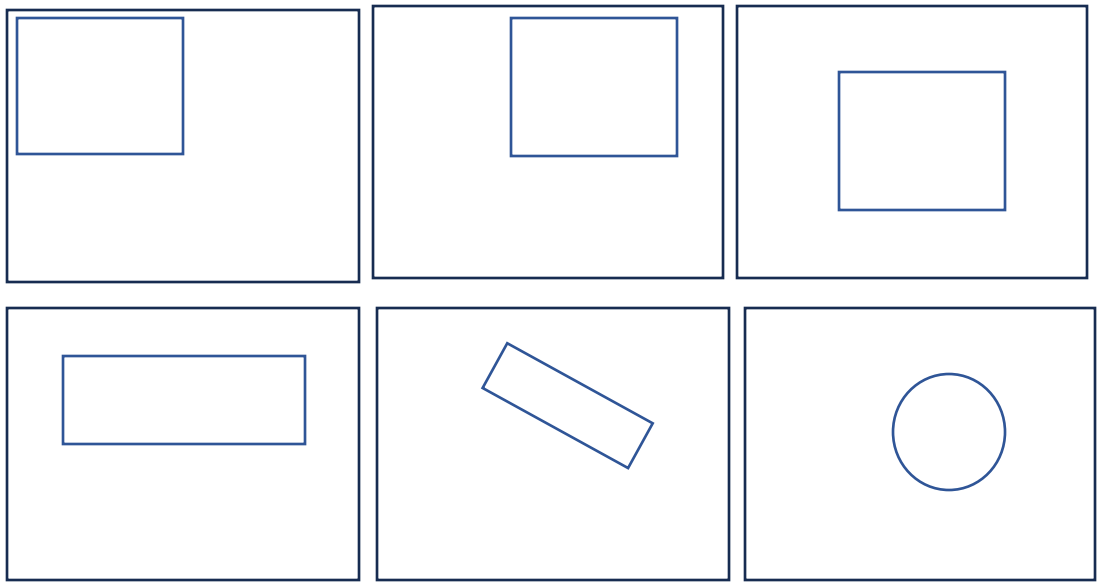
- Realizable (Separable) case
  - When there is a model that separates perfectly
  - There is $h^* \in \mathcal{H}$ that achieves perfect separation between classes
  - i.e. zero *true* loss: $L_{(\mathcal{D},f)}(h^*) = 0$
  - So the *in-sample* loss $L_S(h^*) = 0$



- Agnostic case
  - When there is no perfect separator
  - We just have to find the best imperfect one

We start with realizable and finite $\mathcal{H}$

- To simplify, suppose $\mathcal{H}$ is finite
  - That is, it consists of a few fixed shapes at fixed locations
- See example with $\mathcal{H} = 6$



Finite hypothesis classes

- Suppose the sensor values are in range [0,100] and we can choose thresholds at only integer positions. What is $|\mathcal{H}|$?

- Suppose sensor values are in range [0, 1] and we are choosing from pre-fixed thresholds at intervals of ϵ. How many thresholds (classes) are there?

Empirical risk

- Empirical loss or risk of any hypothesis $h \in \mathcal{H}$ as:

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

Sampling and simple algorithm

- Assumption (iid):
  - Examples in training set are independent and identically distributed according to $\boldsymbol{\mathcal{D}}$
  - Written as $S \sim \boldsymbol{\mathcal{D}}^m$

- ERM Algorithm A:
  - Check all $h \in \mathcal{H}$
  - Pick $h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$
- This h_S is the best we can do with the data, but may not be perfect on unseen data

Question

- Assuming realizability
  - There is a perfect h^* corresponding to unseen data $\mathcal{D}$
  - But we do not know which one

- Can we make sure that we get an h that is close to h^*?
  - That is, may be not zero error, but small error?

Recap: General ML Notations

- **Domain set $\mathcal{X}$.**
- **Label Set $\mathcal{Y}$.** Eg. {0,1} or {-1, +1} red or blue.
- **Training data (sample set): $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$**
- **Model, hypothesis, classifier, predictor h**:
  - A function h: $\mathcal{X} \rightarrow \mathcal{Y}$. That is, $h(x)$ returns a predicted label y
- **Hypothesis class $\mathcal{H}$**: The set of functions from which h is chosen
- **Algortihm A:** Chooses hypothesis h based on S
- **Data generating distribution $\mathcal{D}$**
- **Error on a single item: $\ell(x_i)$**
- **Success measure: Overall Loss/error function *L***

Sampling bound in realizable finite case

- With assumptions of realizability and finite $\mathcal{H}$, we can show that

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$$

- Samples suffice for ϵ, δ guarantee: $\mathbb{P}\left[L_{\mathcal{D},f}(h_S) \leq \epsilon\right] \geq 1 - \delta$
  - The best hypothesis on training data has small true loss
  - With probability $1 - \delta$,

Useful relations

- For $0 < p < 1$ (e.g. p is a probability)
- $(1-p)^{\frac{1}{p}} \leq 1/e$

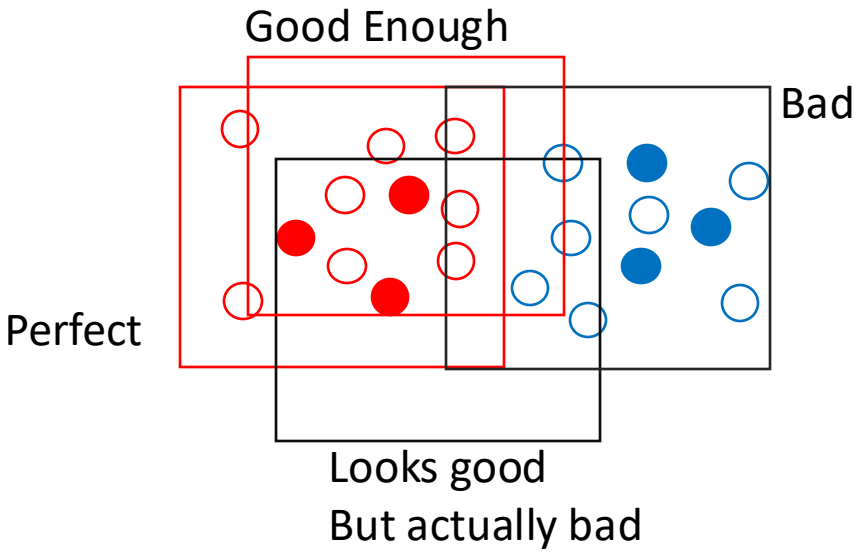
- Union bound:
  - If A and B are event, then: $P(A\ or\ B) \leq P(A) + P(B)$
  - Writing A and B as sets: $P(A \cup B) \leq P(A) + P(B)$

Proof

- The algorithm expects and finds 0 empirical loss in the training set
  - Outputs an h with 0 empirical loss (there can be many of these)
  - All these "Look good" in data
- A Perfect hypothesis also has 0 true loss in $\mathcal{D}$ (realizability in the general case)
- Certain hypothesis are "bad": have a true loss $L_{\mathcal{D},f}(h) > \epsilon$
- We need one that is Good enough: true loss $L_{\mathcal{D},f}(h) \leq \epsilon$

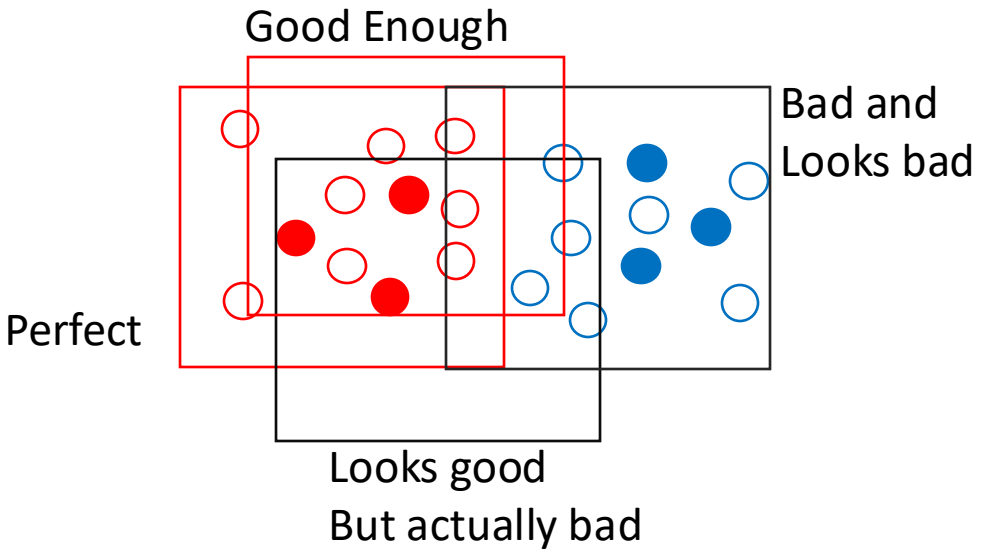
Data in training sample

Unseen Data not in training sample



Proof

- We get a bad output only if a bad hypothesis has zero empirical loss in the sample. Let's compute the probability
- For a bad hypothesis h, $L_{\mathcal{D},f}(h) > \epsilon$, so the probability of getting one training label right is:
  - $1 - L_{\mathcal{D},f}(h) \leq 1 - \epsilon$
- The probability of h getting m labels right is $\leq (1-\epsilon)^m \leq e^{-\epsilon m}$
  - This is the probability that a bad hypothesis h *looks good*



- If H_B is the subset of bad hypotheses
- Probability that a single bad $h \in H_B$ looks good is $e^{-\epsilon m}$
- Probability of some bad hypothesis looking good is
  - $\leq |H_B| e^{-\epsilon m} \leq |\boldsymbol{\mathcal{H}}| e^{-\epsilon m}$
  - Using union bound
  - So we want $|\boldsymbol{\mathcal{H}}| e^{-\epsilon m} \leq \delta$: Probability that is bad hypothesis looks good is small
  - The probability of not getting a bad result is $\geq 1 - \delta$
- Solve for m to to get $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$

QED

Observe

- The proof says that if h^* is the best hypothesis in a finite $\mathcal{H}$,
  - It is always possible to get as close to h^* in accuracy as we want
  - Just need large enough m

- That is, with some assumptions a good enough h_S can always be "learned" from big enough dataset

PAC Learnability

- We have just seen that every finite class is “PAC learnable”

- If $\mathcal{H}$ is finite and realizable, then there is an algorithm that can
  - get as close to the optimum model as we want (small ϵ),
  - with as high a probability as we want (small δ)
  - Provided we give it enough data
    - (and happily, that data is not too much!)

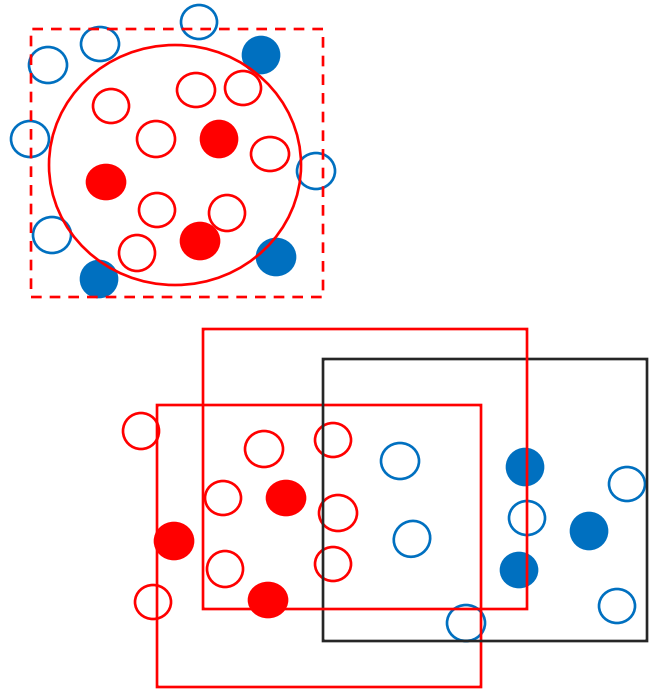
PAC learnability (formal definition)

- A hypothesis class $\mathcal{H}$ is PAC learnable if
  - There exists a function $m_{\mathcal{H}}(0,1)^2 \to \mathbb{N}$ (means: depending on ϵ, δ, there is a suitable number of samples)
  - And an there exists an algorithm that:
    - For every ϵ, δ
    - For $\mathcal{D}$ over $\mathcal{X}$
    - With realizability assumption
    - On $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d samples from $\mathcal{D}, f$
    - Finds an h that satisfies
      - $L_{(\mathcal{D},f)}(h) \leq \epsilon$ (finds a good h)
      - with probability at least $1 - \delta$

Non realizable, or agnostic case

- In general, realizability is not true
  - There may be no perfect $h = f$
- We just have to do our best and find the best $h \in \mathcal{H}$ instead of the ideal one
  - Called Agnostic PAC learning

- E.g. Our $\mathcal{H}$ consists of squares
  - But the data needs a circle to separate classes
  - Or, separation can be achieved by a square but that square is not in our selected fixed set of squares in $\mathcal{H}$
- To extend to more general scenarios, let's change our assumptions



More general model – agnostic learning

- Modified data generating distribution:
  - Define $\mathcal{D}$ to be probability distribution over $\mathcal{X} \times \mathcal{Y}$
  - Consequence: The same $x \in \mathcal{X}$ may have labels 0 or 1 probabilistically
- Redefine true risk:

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\}).$$

- (homework: compare this with how we defined true risk earlier)

- Observe: $\mathcal{D}$ is a probability distribution over $\mathcal{X} \times \mathcal{Y}$ allows same x to have different labels! -- Suggest an example where this is possible

Agnostic PAC learnability

- A hypothesis class $\mathcal{H}$ is Agnostic PAC learnable if
  - There exists a function $m_{\mathcal{H}}(0,1)^2 \to \mathbb{N}$
  - And an algorithm that:
    - For every ϵ, δ
    - For $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
    - ~~With realizability assumption~~
    - On $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d samples from $\mathcal{D}$
    - Finds an h that satisfies
      - $\boldsymbol{L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon}$ (gets ϵ close to the best $h' \in \mathcal{H}$)
      - with probability at least $1 - \delta$

Other types of learning problems (defined by suitable loss)

- We have looked at binary classification
- Other possibilities:
- Multi-class classification
  - E.g, Measure loss as the probability of predicting a wrong label
- Regression: labels are real numbers i.e. $\mathcal{Y} = \mathbb{R}$

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} (h(x) - y)^2$$

Generalised loss

- Instead of $\mathcal{X} \times \mathcal{Y}$, we consider a single domain $\mathcal{Z}$ (which may be $\mathcal{X} \times \mathcal{Y}$, or something else)
  - Loss functions are: $\ell: \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$
    - The loss measured for a single item z on hypothesis h is written as $\ell(h, z)$
- Generalises to more ML problems e.g. clustering (unsupervised learning)
- True risk function: Expected loss: $L_{\mathcal{D}}(h) = \mathbb{E}_{z \in \mathcal{D}}[\ell(h, z)]$
- Empirical risk function: $L_{S(h)} = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$

- Exercise: Define k-means clustering as a formal ML problem, with hypothesis class, loss function etc.

Representative data sets

- We use S as a representative of $\mathcal{D}$
- We hope that
  - We will find an h that does well outside training data,
  - the performance on S matches general performance on $\mathcal{D}$
- When it does, we say S is a representative sample

Representative sample

- S is ϵ −representative w.r.t $(\mathcal{Z}, \mathcal{H}, \mathcal{D})$ if:
  - $\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$

Representative sample

- S is ϵ −representative w.r.t $(\mathcal{Z}, \mathcal{H}, \mathcal{D})$ if:
  - $\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$
- S gives a good estimate of the true loss for each h

- Observe:
  - A sample is representative with respect to $\mathcal{H}, \mathcal{Z}$
  - That is, it is representative with respect to a specifc problem and hypothesis class
- Question: Can there be a notion of represenativeness independent of $\mathcal{H}, \mathcal{Z}$?

Representative sample

- S is ϵ −representative w.r.t $(\mathcal{Z}, \mathcal{H}, \mathcal{D})$ if:
  - $\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$
- S gives a good estimate of the true loss for each h

- Lemma:
  - If S is $\frac{\epsilon}{2}$ −representative, and $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$, then
  - $L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$

- With representative data, the best empirical (trained) model (h_S) is almost as good as the best model for true data

Uniform convergence

- $\mathcal{H}$ has uniform convergence if there is function $m_{\mathcal{H}}^{UC}: (0,1)^2 \to \mathbb{N}$
  - Such that a random sample $S \sim \mathcal{D}^m$ of size $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$
  - Is ϵ −representative with probability at least $1 - \delta$

- When $\mathcal{H}$ has uniform convergence, it means we know a large enough m that gives accurate estimates for all h

Corollary

- If $\mathcal{H}$ has uniform convergence with $m_{\mathcal{H}}^{UC}$,
  - Then $\mathcal{H}$ is PAC learnable with $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$

- Theorem:
- Every finite $\mathcal{H}$ has uniform convergence
  - i.e. Given a random suitable sized S, $\mathbb{P}[\exists h \in \mathcal{H}: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq \delta$

- And therefore every finite $\mathcal{H}$ is agnostic PAC-learnable

- Proof, using Chernoff-hoeffding bound

Chernoff-Hoeffding bound

- Very important result in theoretical CS and ML
- Suppose θ_i are random variables with average $\frac{1}{m}\sum_{i=1}^{m}\theta_i$
- Suppose μ is the expected value of a random θ
- Law of large numbers: with increasing m, $\frac{1}{m}\sum_{i=1}^{m}\theta_i$ approaches μ
  - Ie, $\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right|$ becomes smaller
- But how fast? What m do we need to get ϵ-close to μ?
- Chernoff-Hoeffding bound:
- $\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$

- Theorem:
- Every finite $\mathcal{H}$ has uniform convergence
  - i.e. Given a random S, $\mathbb{P}[\exists h \in \mathcal{H}: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq \delta$

- (And therefore every finite $\mathcal{H}$ is agnostic PAC-learnable)

- To prove this, we need the Chernoff-hoeffding bound

- Proof that $\mathbb{P}[\exists h \in \mathcal{H}: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq \delta$ [from book]
- Take any $h \in \mathcal{H}$
- Now take a random sample S
- Let us write $\mu = \mathbb{E}[L_S(h)] = L_{\mathcal{D}}(h)$
  - I.e. note that the expected value of empirical loss is the true loss
- For every $z_i \in S$, we write its loss on h as θ_i. I.e. $\theta_i = \ell(h, z_i)$
- Then the empirical loss is $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \theta_i$
- So, what is the probability that $\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| > \epsilon$?

- What is the probability that $\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon$?
- Using Chernoff bound, probability that any one h has large error is:
  - $\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$
- Summing over all $h \in \mathcal{H}$, probability that one or more has large error is:
  - $\leq 2|\mathcal{H}|e^{-2m\epsilon^2}$ (by union bound)
- Substitute $m \geq \frac{1}{2\epsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)$ to get a probability bound δ

- So, we can proved finite hypothesis classes are all PAC learnable (see next lecture)

- Next week, we will cover
  - No free lunch theorem: There is no universal learner
  - Bias-complexity tradeoff
  - Infinite hypothesis classes and fundamental theorem of statistical learning
  - Starting with ML algorithms/models

- Read chapters 3 & 4
- Lecture notes for last week up now.