PAC Learning

Lecture 6

Last Time

- Stochastic gradient descent (SGD) is a variant of gradient descent that is faster and more scalable.
  - Just like gradient descent, except we compute the gradient on small batches of training data
- In addition to step size, other hyperparameters include batch size and the convergence criterion
- How we represent our data affects which hypotheses we learn
- Normalizing your data is often needed for best performance
- Textbook: sections 12.1.1, 14.0, 14.1.0, 14.3.0, 14.5.1

This Class

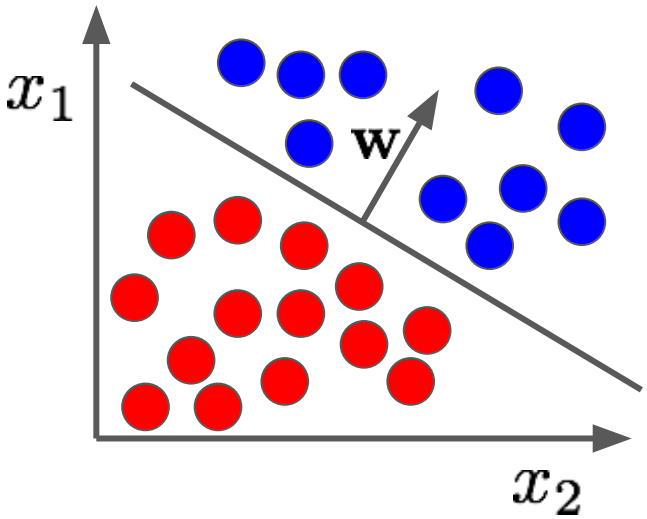
- Now that we have some hypothesis classes in our toolbox, what can we prove formally about machine learning algorithms that use the ERM principle?

- Textbook: chapters 2.3, 3

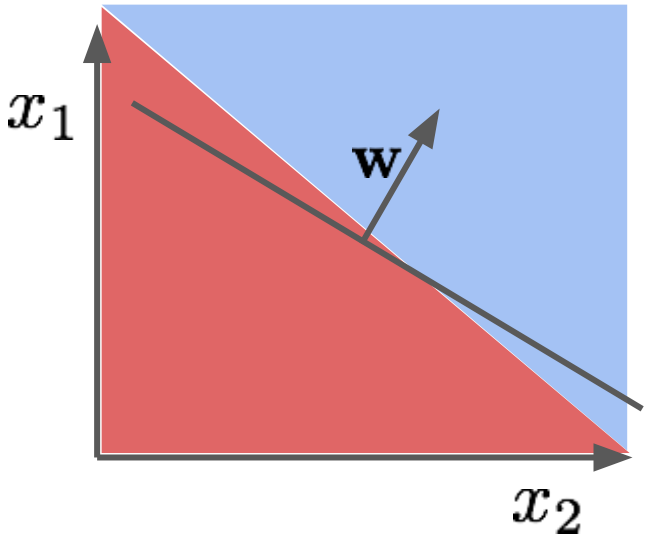
Motivation

Motivation

- What does it mean for a machine learning algorithm to be correct?



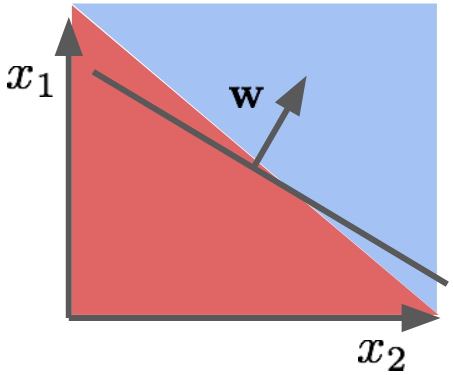
$$L_S(h_S) = 0$$



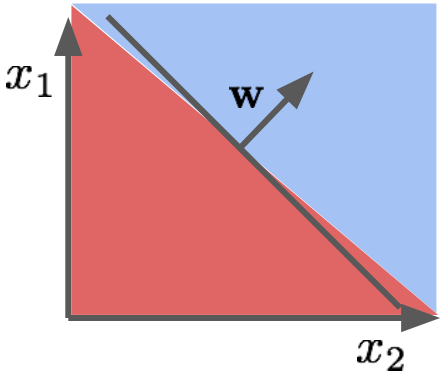
$$L_{\mathcal{D}}(h_S) > 0$$

Approximately Correct

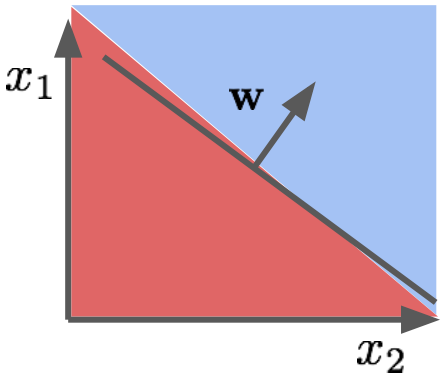
- We want some guarantee of the quality of the hypothesis, h_S, that we get from an algorithm



$L_{\mathcal{D}}(h_S) \leq \epsilon$



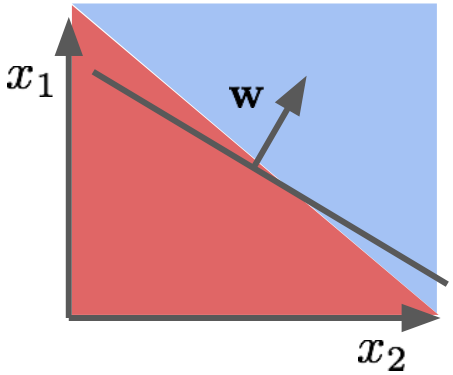
$L_{\mathcal{D}}(h_S) \leq \epsilon$



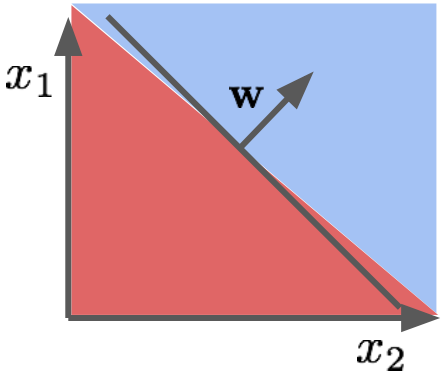
$L_{\mathcal{D}}(h_S) \leq \epsilon$

Probably Approximately Correct

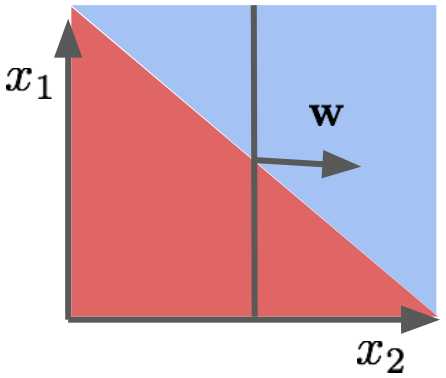
- Since $S \sim \mathcal{D}^m$, there is some chance that the training data is not representative of the true distribution. We want to bound the probability that this happens.



$L_{\mathcal{D}}(h_S) \leq \epsilon$



$L_{\mathcal{D}}(h_S) \leq \epsilon$



Well, 2 out of 3 ain't bad

Probably Approximately Correct (PAC) Learning

Goal

- Want to bound the frequency that the model returns a “bad” hypothesis given some $S \sim \mathcal{D}^m$
- In other words, we want to bound this probability:

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D}}(h_S) > \epsilon\})$$

- $S|_x = (x_1, x_2, \ldots, x_m)$ is the attribute instances of the training set
- Once we have this bound, we can see how many examples we need (i.e., how big m has to be) for the probability to be small

Analysis Assumptions

1. **All semester:** independently and identically distributed (i.i.d.) data

$$\mathcal{D}^2(z_1, z_2) = \mathcal{D}(z_1)\mathcal{D}(z_2) \quad \forall z \in \mathcal{X} \times \mathcal{Y}$$

2. **Just today:** finite hypothesis class

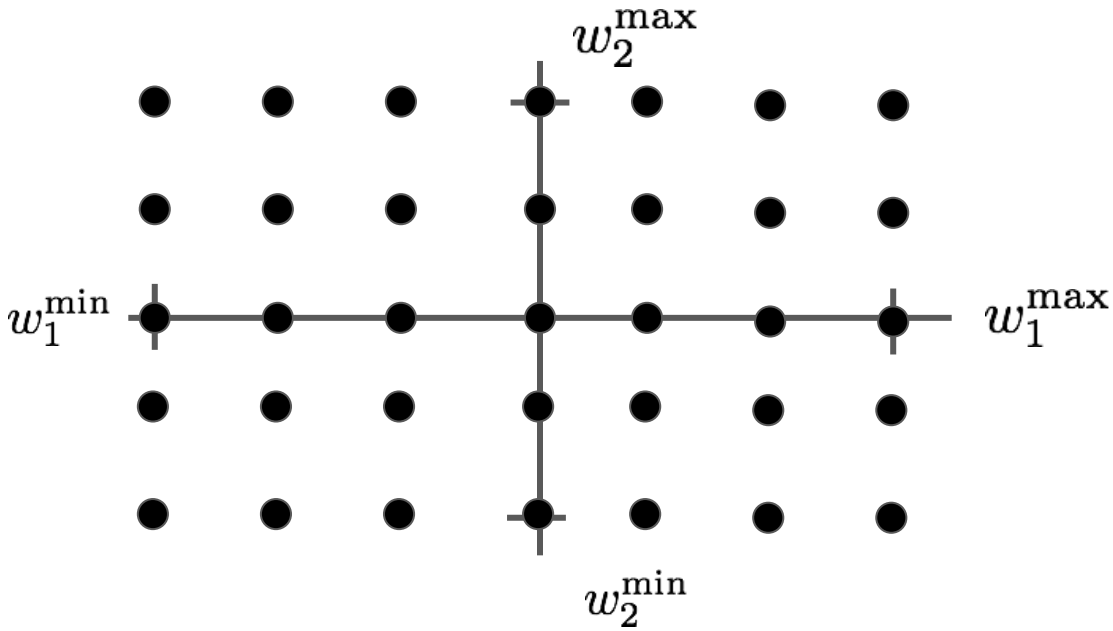
$$|\mathcal{H}| < \infty$$

3. **Just today:** realizability

$$\exists h^\star \in \mathcal{H} : L_{\mathcal{D}}(h^\star) = 0$$

Finite Hypothesis Classes

Think of a halfspace or logistic regression classifier where the weights are represented with finite precision



What We Don’t Assume

- We won’t place any additional assumptions on $\mathcal{D}$
- So our further goal is to bound the probability of failure with a quantity that doesn’t refer to $\mathcal{D}$
- Called “distribution free” learning

What's a Bad Hypothesis?

- We say that learning fails if a learning algorithm returns a hypothesis in

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon\}$$

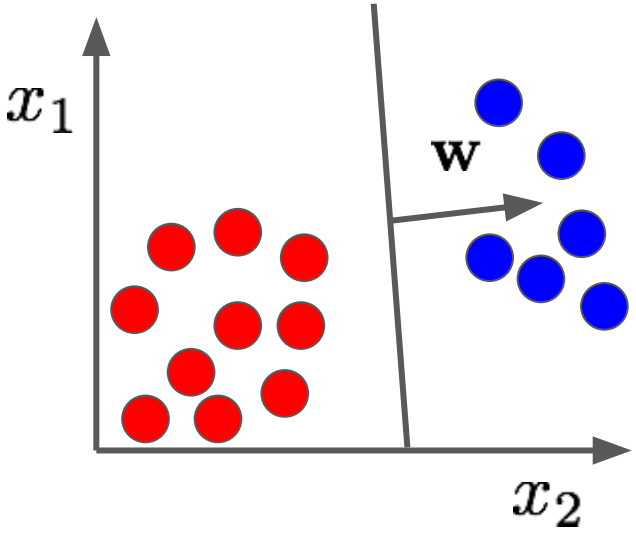
How could we Learn a Bad Hypothesis?

- Realizability implies that $L_S(h_S) = 0$

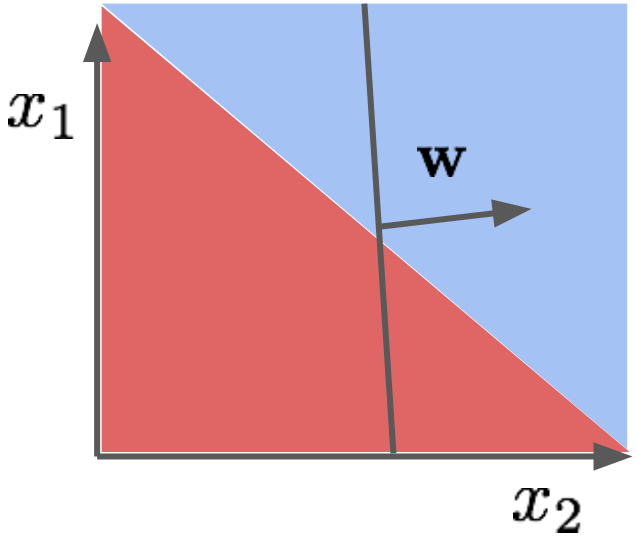
- $L_{\mathcal{D}}(h_S) > \epsilon$ can only occur when the training samples are perfectly predicted by a bad hypothesis in $\mathcal{H}_B$

- In other words, we can rephrase the question as “What’s the probability we get a misleading sample of training data?”

A Misleading Sample



$$L_S(h_S) = 0$$



$$L_{\mathcal{D}}(h_S) > \epsilon$$

All Misleading Samples

- Let

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

  be the set of all misleading samples

- In other words, M is the set of training sets that are correctly classified by at least one bad hypothesis

Learning Can Fail When the Sample is Misleading

- The probability of getting a misleading sample is $\mathcal{D}^m(M)$

  where

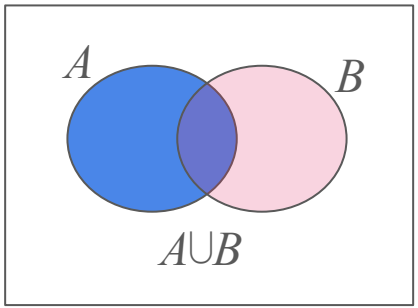
$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

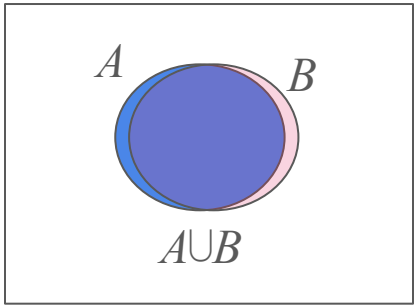
- Now we can begin constructing an upper bound:

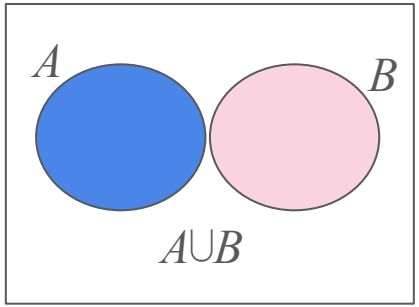
$$\begin{aligned} \mathcal{D}^m(\{S|_x : L_{\mathcal{D}}(h_S) > \epsilon\}) &\leq \mathcal{D}^m(M) \\ &= \mathcal{D}^m(\cup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}) \end{aligned}$$

A Useful Tool: Union Bound

For any two events A and B and any distribution $\mathcal{D}$, what can we say about $\mathcal{D}(A \cup B)$ in relation to $\mathcal{D}(A)$ and $\mathcal{D}(B)$?







$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

Question

We’ll Have to Wait for the Question

Bounding the Probability of Perfection

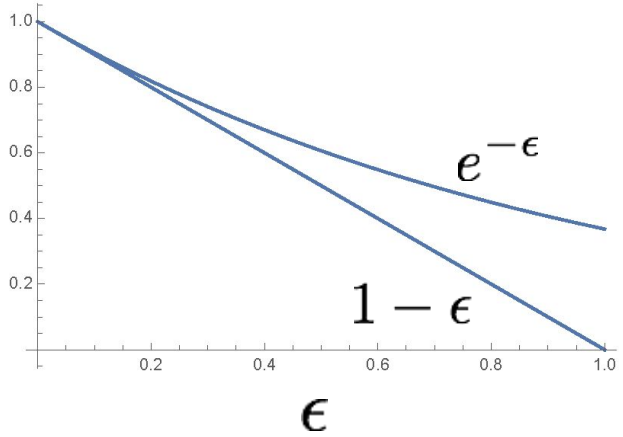
$$\sum_{h \in \mathcal{H}_B} \underbrace{\mathcal{D}^m(\{S|_x : L_S(h) = 0\})}_{\text{“Probability of Perfection”}}$$

For the empirical risk to be zero, the hypothesis must correctly label every example $(\mathbf{x}_i, y_i) \in S$

$$\begin{aligned}
\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i \in [m], h(\mathbf{x}_i) = y_i\}) \\
&= \prod_{i=1}^{m} \mathcal{D}(\{\mathbf{x}_i : h(\mathbf{x}_i) = y_i\}) \quad \text{since examples are i.i.d.} \\
&\leq \prod_{i=1}^{m} (1 - \epsilon) \quad \text{by definition of } \mathcal{H}_B!
\end{aligned}$$

A Convenient Trick

Using the inequality $1 - \epsilon \leq e^{-\epsilon}$, we can make the bound more convenient:



$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \prod_{i=1}^{m}(1-\epsilon) \\ &= (1-\epsilon)^m \\ &\leq e^{-\epsilon m}\end{aligned}$$

Putting it All Together

Now, using this in our original inequality:

$$\begin{aligned}
\mathcal{D}^m(\{S|_x : L_{\mathcal{D}}(h_S) > \epsilon\}) &\leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \\
&\leq \sum_{h \in \mathcal{H}_B} e^{-\epsilon m} \\
&= |\mathcal{H}_B| e^{-\epsilon m} \\
&\leq |\mathcal{H}| e^{-\epsilon m} \qquad \text{since } \mathcal{H}_B \subseteq \mathcal{H}
\end{aligned}$$

A Sufficient Condition on m

Recall our goal to upper bound the probability of a bad hypothesis: $\mathcal{D}^m(L_{\mathcal{D}}(h_S) > \epsilon)$

Let's restate what we've shown as a sufficient condition (lower bound) on m:

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

$$\log \delta \geq \log |\mathcal{H}| - \epsilon m$$

$$0 \geq \log \frac{|\mathcal{H}|}{\delta} - \epsilon m$$

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

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

PAC Learnability

- See Definition 3.1 and Corollary 3.2 to behold it in its full glory

- Formally states what we've just shown: given any δ, ϵ, and finite hypothesis class, $L_{\mathcal{D}}(h_S) \leq \epsilon$ with probability at least $1 - \delta$ for any $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ where

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

- This upper bound on $m_{\mathcal{H}}(\epsilon, \delta)$ is called the sample complexity

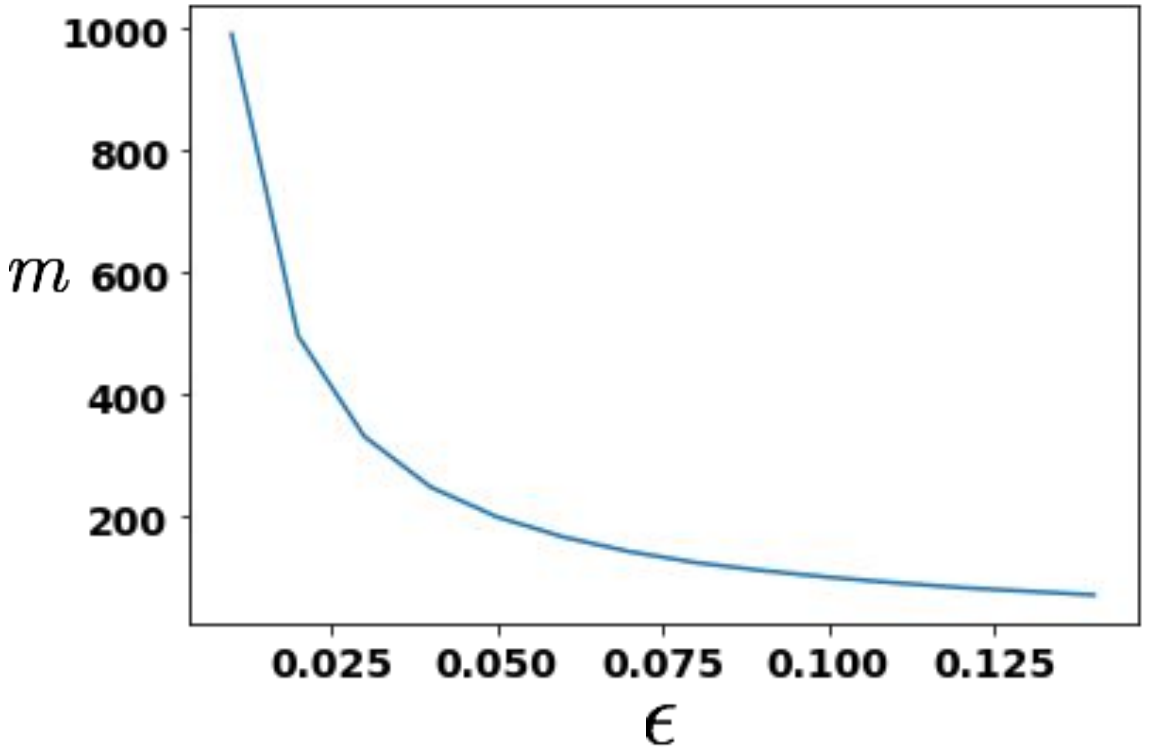
Examples

Varying Epsilon

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

$$|\mathcal{H}| = 1000$$

$$\delta = 0.05$$

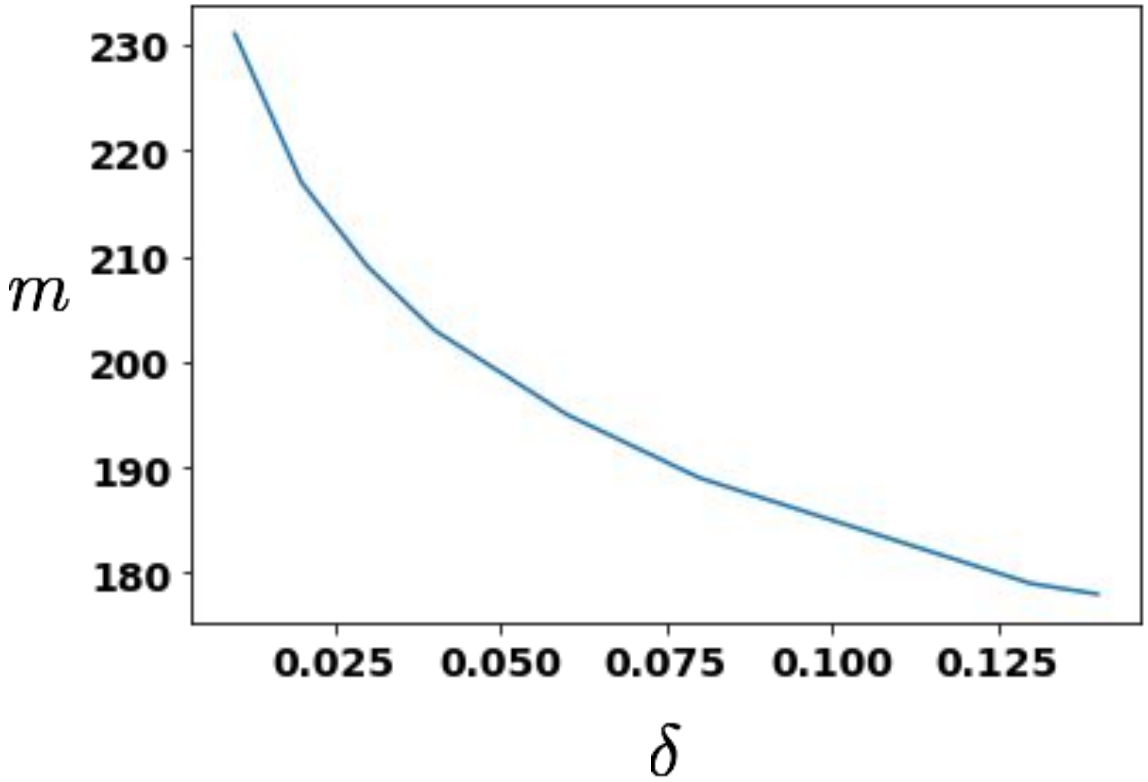


Varying Delta

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

$$|\mathcal{H}| = 1000$$

$$\epsilon = 0.05$$

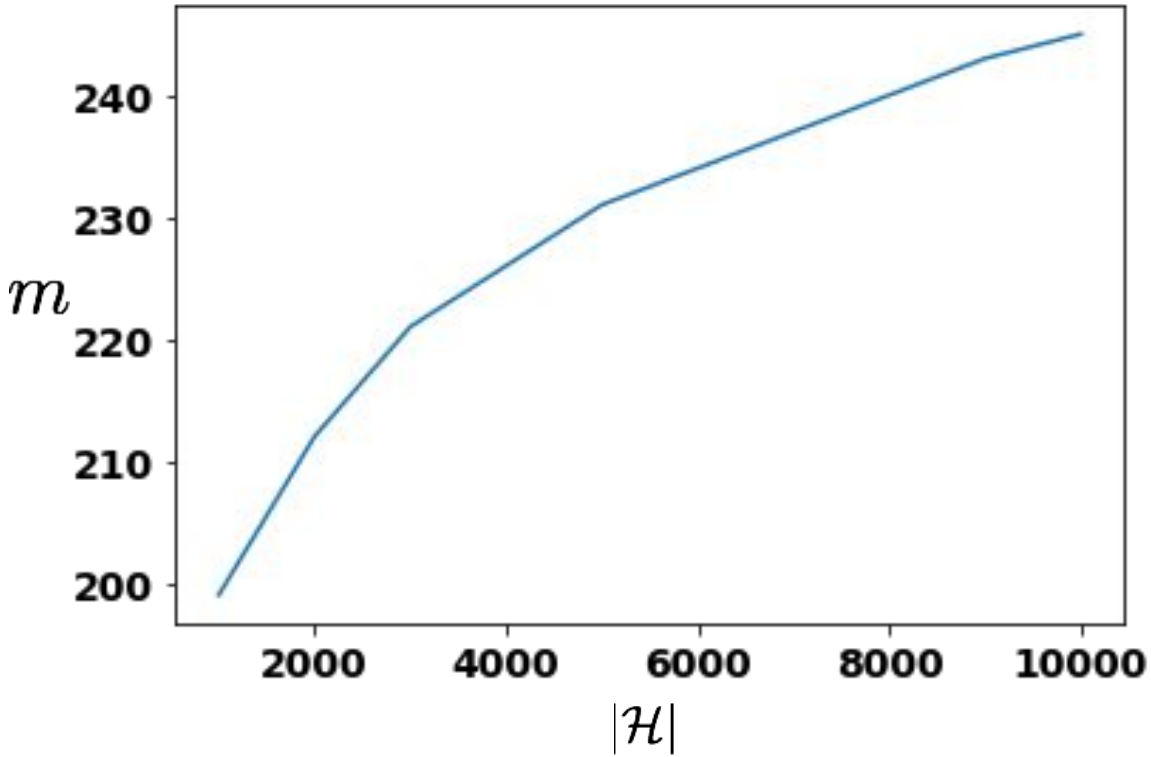


Varying H

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

$$\epsilon = 0.05$$

$$\delta = 0.05$$



The Unrealizable Case

What happens in the unrealizable case?

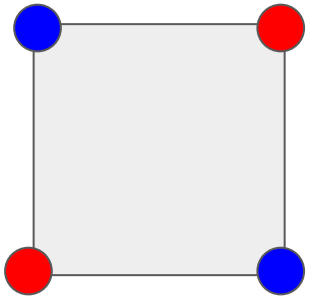
- Recall, under the realizability assumption, there exists $h^* \in \mathcal{H}$ such that

$$P_{\mathbf{x},y\sim\mathcal{D}}[h^\star(\mathbf{x}) = y] = 1$$

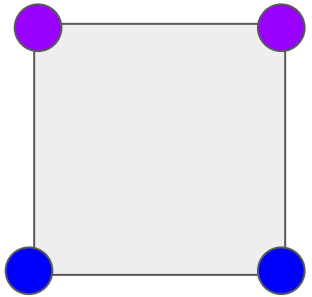
- Do we have the same guarantee when this assumption does not hold?

Two Things Can Happen

1. Hypothesis class does not fit perfectly, e.g., halfspace on data that isn't linearly separable



2. Label noise



We Need More Power!

- Either way, we can no longer identify a hypothesis as bad if it makes a single mistake on training data

- We'll need to develop a more powerful theory to relax the realizability assumption!

- We'll revisit this question in March

The Most Important Things

- ***Probably approximately correct (PAC) learnability*** is a property of a hypothesis class $\mathcal{H}$. If it holds, there's some function that gives a number of i.i.d. training examples m that are sufficient to guarantee that $L_{\mathcal{D}}(h_S) \leq \epsilon$ with probability at least $1 - \delta$ (for arbitrary ϵ and δ, and some algorithm)

- We've shown that any finite, realizable $\mathcal{H}$ is PAC learnable via ERM, with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

- Textbook: chapters 2.3, 3

Next Time

- Can we hope to find a universal learner that works best on all problems?
- If not, what tradeoffs must we make when selecting a learning algorithm?

  Textbook: chapter 5