Learning via Uniform Convergence

Lecture 11

Last Time

- ***Decision trees*** encode a set of rules for making predictions
- We have different learning challenges due to discrete hypothesis class
- Greedy search with pruning is usually the preferred strategy

Textbook: chapter 18

This Class

- Our next round of learning theory!
- What can we prove about the unrealizable case?
- Textbook: chapter 4

Review: PAC Learning

- ***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

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. **Up to today:** finite hypothesis class

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

3. ~~**Last time:** realizability~~

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

3. **Rest of semester:** bounded loss

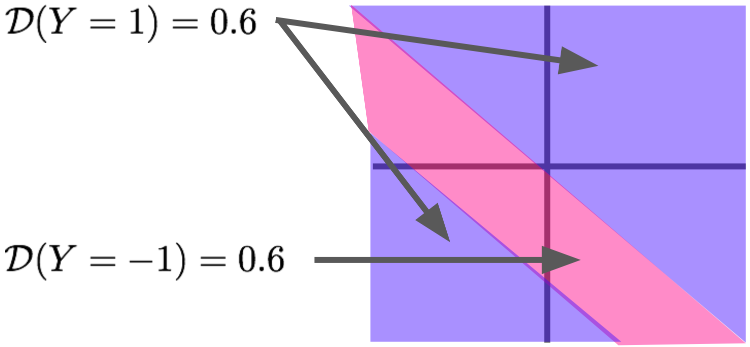
$$\exists a, b \in \mathbb{R} \quad \forall h \in \mathcal{H}, \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y} \quad \ell(h, (\mathbf{x}, y)) \in [a, b]$$

Agnostic PAC Learning

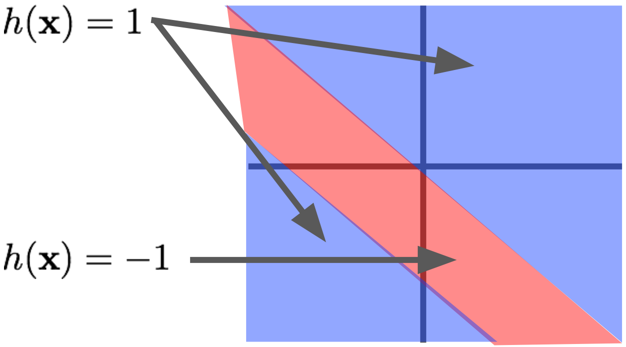
Bayes Optimal Predictor

- What's the best we could ever do in the unrealizable case?
- Let's start with intuition for the classification case, so $\mathcal{Y} = \{1, -1\}$
- Example:

True Distribution $\mathcal{D}$



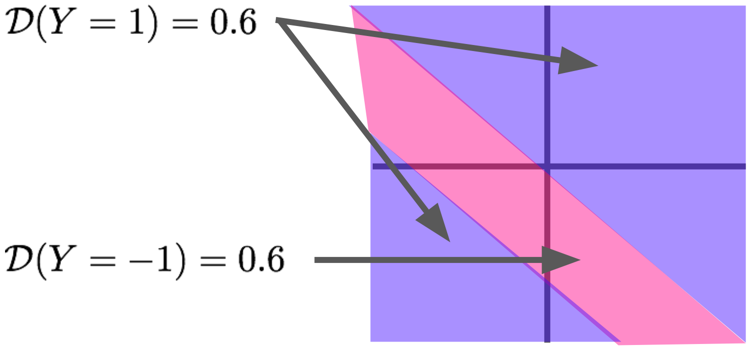
Bayes Opt. Predictor



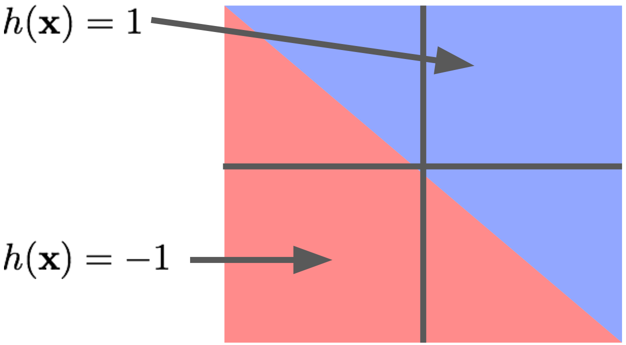
Best in Class

- But the Bayes optimal predictor might not be in our hypothesis class!
- The best hypothesis in the class is $\underset{h \in \mathcal{H}}{\arg\min}\, L_{\mathcal{D}}(h)$
- Example:

True Distribution $\mathcal{D}$



Best in class h



How should we
define success now?

Agnostic PAC Learning

- Property of a hypothesis class with respect to a data representation $\mathcal{X} \times \mathcal{Y}$ and loss ℓ, analogous to PAC, except relative to best hypothesis in the class

- There exists $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm such that, for any $\epsilon, \delta \in (0,1)$, if we have m i.i.d. examples where

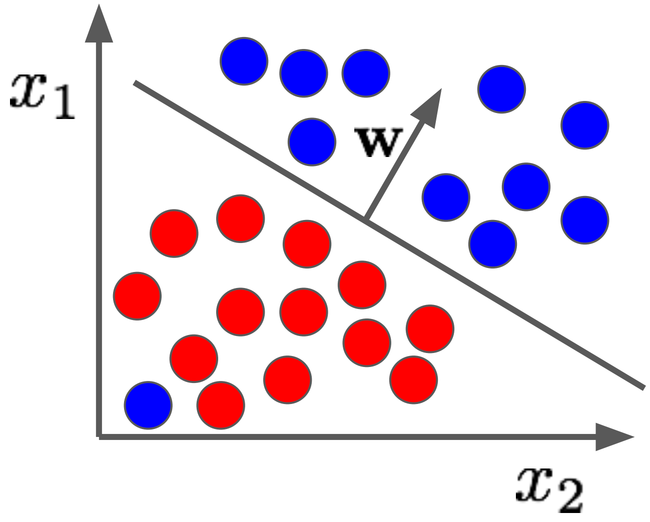
$$m \geq m_{\mathcal{H}}(\epsilon, \delta)$$

then with probability at least $1-\delta$, the learning algorithm returns h such that

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

Why is Agnostic PAC Learning Hard?

- Is there a better hypothesis than this one?
- Under the realizability assumption, we could immediately throw away this hypothesis
- Without realizability, this might be the best in class!



Uniform Convergence

Addressing the Problem Directly

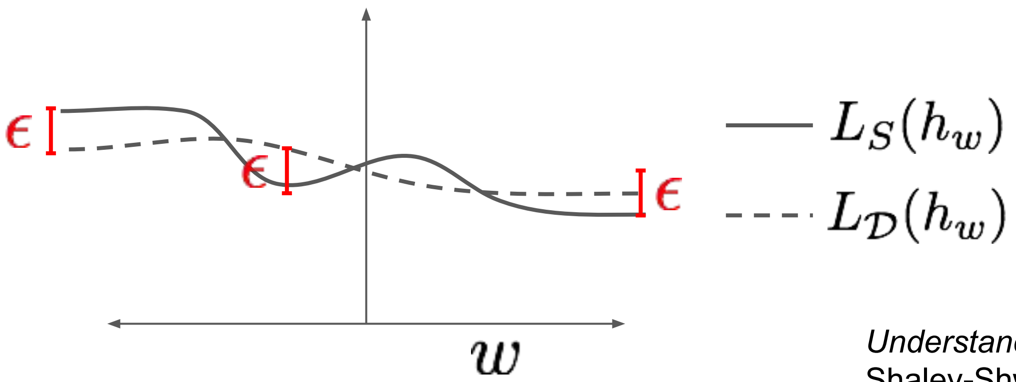
- The big challenge in machine learning is that $L_S(h) \neq L_{\mathcal{D}}(h)$
- If they were equal, then we'd just be doing optimization
- What theoretical tools can help us study this challenge?

Epsilon-Representative Sample

DEFINITION 4.1 (ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

Example:



Understanding Machine Learning.
Shalev-Shwartz and Ben-David, 2014.

If Training Data is Representative then ERM is Good

LEMMA 4.2 *Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$*, namely, any* $h_S \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$*, satisfies*

$$L_{\mathcal{D}}(h_S) \;\le\; \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Understanding Machine Learning.
Shalev-Shwartz and Ben-David, 2014.

If Training Data is Representative then ERM is Good

Proof: For every $h \in \mathcal{H}$

$$
\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\epsilon}{2} && \text{Because } S \text{ is } \frac{\epsilon}{2}\text{-representative} \\
&\leq L_S(h) + \frac{\epsilon}{2} && \text{By definition of ERM} \\
&\leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} && \text{Because } S \text{ is } \frac{\epsilon}{2}\text{-representative} \\
&= L_D(h) + \epsilon
\end{aligned}
$$

Uniform Convergence

DEFINITION 4.3 (Uniform Convergence) We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Understanding Machine Learning.
Shalev-Shwartz and Ben-David, 2014.

Uniform Convergence is Sufficient for Agnostic PAC

- Putting together Lemma 4.2 and Definition 4.3, we see that uniform convergence is sufficient for agnostic PAC learnability

- Formally stated in Corollary 4.4

Summary of Reasoning Steps

1. Assume we have a finite hypothesis class H and loss bounded in [0,1]

2. Then, H has uniform convergence

3. Then, with probability $1-\delta$, if we have a training sample S with size m, where

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta) \leq m$$

   then S is $\frac{\epsilon}{2}$- representative

4. If S is $\frac{\epsilon}{2}$- representative, then $L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$

Uniform Convergence Holds for Any Finite Hypothesis Class

Proving Uniform Convergence for Finite Classes

- Uniform convergence is such a powerful property, it's all we need to prove to show that a hypothesis class is agnostic PAC learnable via ERM

- We will follow a similar proof to PAC learning: derive an upper bound using the union bound

- We will also need another tool called Hoeffding's Inequality

Setting Up the Bound

- We want to upper bound $\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}| > \epsilon\})$

- Observe that

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}| > \epsilon\} = \cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}$$

- Then by the union bound

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_\mathcal{D}(h)| > \epsilon\})$$

What Next? Try to Concentrate...

- Now we need to upper bound $\sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$

- We'll argue that each term $\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$ gets small as m gets big

Average over Samples **Expected Value**

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$$

Concentration and Hoeffding's Inequality

Example: Sums of Random Variables Concentrate

Say we draw m random examples from a normal distribution with unknown mean:

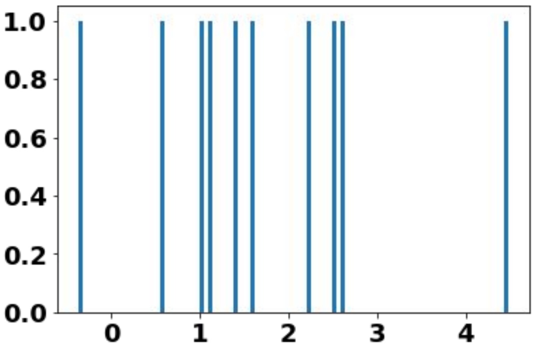
$$x = \mathcal{N}(0, 1) + \mu$$

We can estimate μ by taking the average of all m examples

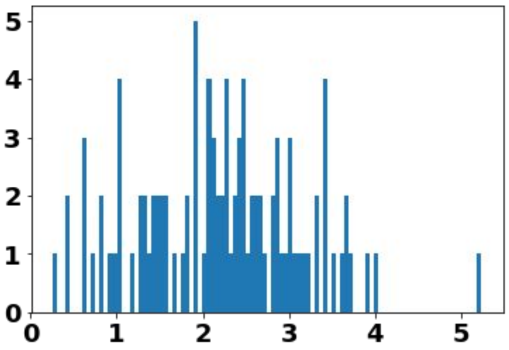
$$\hat{\mu}^m = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Example: Sums of Random Variables Concentrate

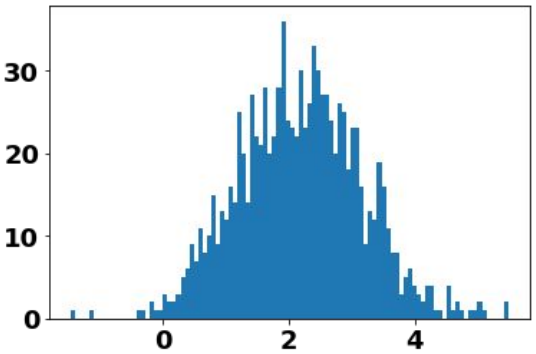
$$\hat{\mu}^{10} = 1.72$$



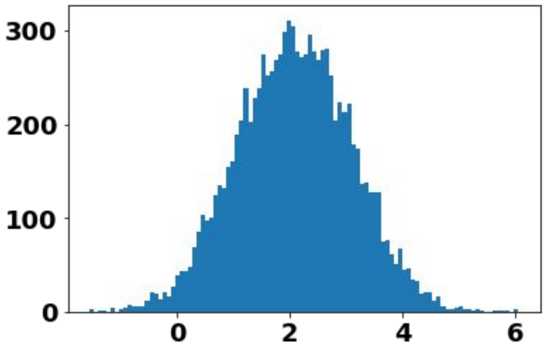
$$\hat{\mu}^{100} = 2.21$$



$$\hat{\mu}^{1000} = 2.18$$



$$\hat{\mu}^{10000} = 2.08$$



BORING!

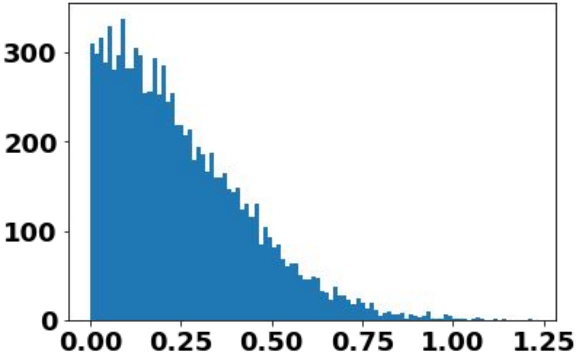
Example: Sums of Random Variables Concentrate

- A more interesting question: what if we fix m, then repeatedly collect a batch and take the average $\hat{\mu}^m$?

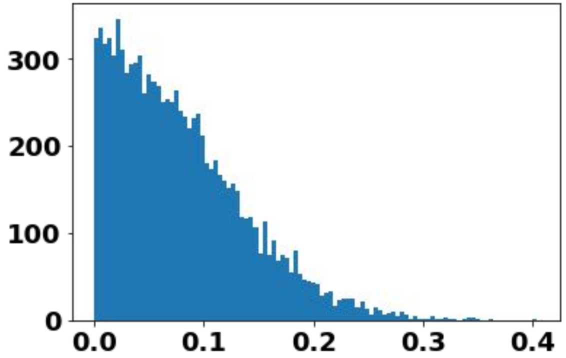
- What fraction of times would $|\hat{\mu}^m - \mu| > \epsilon$ for some ϵ ?

Example: Sums of Random Variables Concentrate

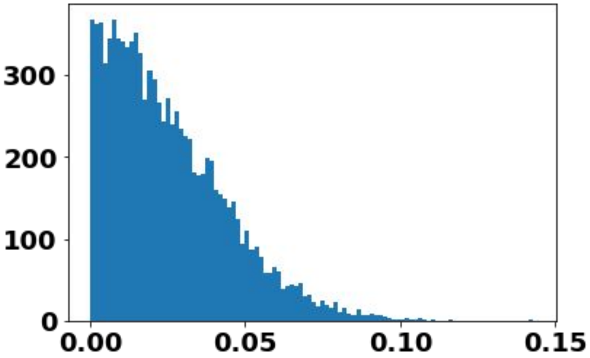
$|\hat{\mu}^{10} - \mu|$



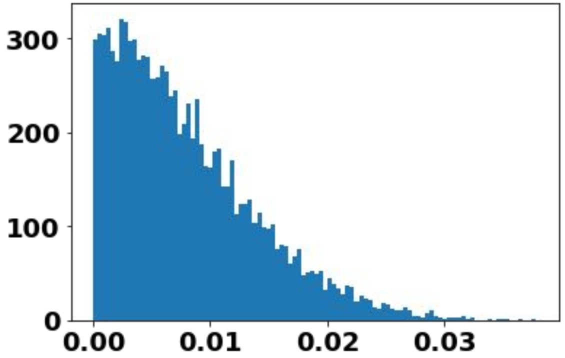
$|\hat{\mu}^{100} - \mu|$



$|\hat{\mu}^{1000} - \mu|$



$|\hat{\mu}^{10000} - \mu|$



Hoeffding's Inequality

Let $\theta \in [a, b]$, $\theta \sim \mathbb{P}$, and $\mathbb{E}_{\mathbb{P}}[\theta] = \mu$. Then for any $\epsilon > 0$:

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

(Proof in Appendix B.)

Question

Applying Hoeffding's Inequality

We just saw that if $\theta \in [a,b]$, $\theta \sim \mathbb{P}$, and $\mathbb{E}_{\mathbb{P}}[\theta] = \mu$, then for any $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} \theta - \mu\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

Which of the following is an upper bound on $\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$ implied by Hoeffding's Inequality if we assume $0 \leq L_S(h), L_D(h) \leq 1$?

A: $2\exp\left(\frac{-2m\epsilon^2}{4}\right)$

B: $2\exp\left(-2m\epsilon^2\right)$

C: $\frac{2}{m}\exp\left(\frac{-2m\epsilon^2}{4}\right)$

D: $\frac{2}{m}\exp\left(-2m\epsilon^2\right)$

Answer

Answer: B

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2\exp(-2m\epsilon^2)$$

Not A or C because $a = 0$ and $b = 1$ (as opposed to $b = 2$)

Not C or D because $L_S(h)$ is already the average over m examples

Proving Uniform Convergence

Our Final Upper Bound

Continuing with the assumption that $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$:

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} 2 \exp(-2m\epsilon^2) \\ &= 2|\mathcal{H}| \exp(-2m\epsilon^2)\end{aligned}$$

Solving for m

If we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

then

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}| > \epsilon\}) \leq \delta$$

Conclusions

- Any finite hypothesis class $\mathcal{H}$ has uniform convergence with respect to a loss $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$ with sample complexity

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

- Further, $\mathcal{H}$ is agnostically PAC learnable via ERM with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

(Corollary 4.6)

Summary of Reasoning Steps

1. Assume we have a finite hypothesis class H and loss bounded in [0,1]

2. Then, H has uniform convergence

3. Then, with probability $1-\delta$, if we have a training sample S with size m, where

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil \leq m$$

   then S is $\frac{\epsilon}{2}$- representative

4. If S is $\frac{\epsilon}{2}$- representative, then $L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$

Comparison with PAC Learning

Comparison with PAC Learning

- Compare the sample complexity of PAC learning:

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

with agnostic PAC learning:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

- Dropping realizability increases sample complexity by factor of more than $\frac{2}{\epsilon}$!

The Most Important Things

- ***Agnostic probably approximately correct (PAC) learning*** is a property of a hypothesis class $\mathcal{H}$. If it holds, there's a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and an algorithm such that if we have m i.i.d. examples where $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, then with probability at least $1 - \delta$ the algorithm returns h such that

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

- We've shown ***any finite hypothesis class is agnostic PAC learnable*** via ERM with respect to a loss function with range [0,1], with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Textbook: chapter 4

Next Time

- Our final tool of learning theory: what makes a hypothesis class learnable? Can infinite hypothesis classes ever be learnable?

- Textbook: chapters 6.0, 6.1, 6.2, 6.3, 6.4, 9.1.3