10-701: Introduction to Machine Learning

Lecture 19 – Learning Theory (Finite Case)

Hoda Heidari

* Slides adopted from F24 offering of 10701 by Henry Chai.

Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- Learning Theory
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- **Learning Theory**
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

Statistical Learning Theory Model

1. Data points are generated i.i.d. from some *unknown* distribution

$$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$

2. Labels are generated from some *unknown* function

$$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right)$$

3. The learning algorithm chooses the hypothesis (or classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$

4. Goal: return a hypothesis (or classifier) with low *true* error rate

Recall: Types of Error

- True error rate
  - Actual quantity of interest in machine learning
  - How well your hypothesis will perform on average across all possible data points $\boxed{\mathbb{E}_{x \sim p^*}\left(h(x) \neq c^*(x)\right)}$
- Test error rate: used to evaluate hypothesis performance
  - Good estimate of the true error rate $\mathcal{D}' \sim p^*$
- Validation error rate: used to set model hyperparameters
  - Slightly "optimistic" estimate of the true error rate
- Training error rate: used to set model parameters $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(h(x^{(i)}) \neq c^*(x^{(i)})\right)$
  - Very "optimistic" estimate of the true error rate

$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ w. $(x^{(i)} \sim p^*)$

Types of Risk (a.k.a. Error)

- Expected *risk* of a hypothesis h (a.k.a. true error)

$$\underline{R(h)} = P_{\underline{\boldsymbol{x} \sim p^*}}\left(\underline{c^*(\boldsymbol{x})} \neq \underline{h(\boldsymbol{x})}\right)$$

- Empirical risk of a hypothesis h (a.k.a. training error)

$$\begin{aligned}\underline{\hat{R}(h)} &= P_{\boldsymbol{x} \sim \underline{\mathcal{D}}}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right) \\ &= \frac{1}{N}\sum_{\underline{n=1}}^{N} \mathbb{1}\left(c^*\left(\boldsymbol{x}^{(n)}\right) \neq h\left(\boldsymbol{x}^{(n)}\right)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(y^{(n)} \neq h\left(\boldsymbol{x}^{(n)}\right)\right)\end{aligned}$$

where $\underline{\mathcal{D}} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$ is the training data set with $\boldsymbol{x}^{i}$ denoting a point sampled uniformly at random from $\underline{p^*}$

Three Hypotheses of Interest

1. The *true function*, c^*

2. The *expected risk minimizer,*

$$\underbrace{h^*} = \underset{\underbrace{h \in \mathcal{H}}}{\operatorname{argmin}}\, \underbrace{R(h)}$$

3. The *empirical risk minimizer,*

$$\hat{h} = \underset{\underline{h} \in \underbrace{\mathcal{H}}}{\operatorname{argmin}}\, \underline{\hat{R}(h)}$$

Empirical risk minimization paradigm

Key Question

- Given a hypothesis with zero/low training error, what can we say about its true error?

PAC Learning

- PAC = **P**robably **A**pproximately **C**orrect
- PAC-learning is a mathematical framework for analysis learning algorithms:
  - The learner receives samples ($\mathcal{D}$)
  - It must select a *hypothesis* h from a certain hypothesis class $\mathcal{H}$.
  - The goal is that, **with high probability,** the selected function will have **low error,**
  - No matter what the underlying distribution of samples p^* is.

PAC Learning

- PAC = **P**robably **A**pproximately **C**orrect
- PAC Criterion:

$$\underbrace{P\left(\left|R(h)-\hat{R}(h)\right| \le \overset{\text{error tolerance}}{\epsilon}\right)}_{\text{Generalization error}} \ge \overset{\text{confidence}}{\underbrace{1-\delta}} \;\forall\, h \in \mathcal{H}$$

for some ϵ (difference between expected and empirical risk) and δ (probability of "failure")

  - We want the PAC criterion to be satisfied for $\mathcal{H}$ with *small* values of ϵ and δ

Sample Complexity

- The **sample complexity** of a learning algorithm operating on hypothesis set, $\mathcal{H}$, is the number of labelled training data points needed to satisfy the PAC criterion for some δ and ϵ.

sampled i.i.d from p^*

$m(\varepsilon, \delta)$: # of samples needed to satisfy the PAC criteria for a given ε, δ.

Sample Complexity & PAC Learnability

- A hypothesis class $\mathcal{H}$ is PAC-learnable if for every $\epsilon , \delta \in (0 , 1)$, there exists a sample size $m (\epsilon , \delta)$ polynomial in $1/\epsilon$ and $1/\delta$, such that with m i.i.d. samples from ANY distribution p^* the algorithm outputs a hypothesis whose generalization error is at most ϵ with probability at least $1 - \delta$.

↳ PAC-learning criteria

Poll: PAC-learning

- Which of statement most precisely captures what it means for a hypothesis class $\mathcal{H}$ to be PAC learnable?

Sample Complexity

	Realizable	Agnostic
Finite	today	
infinite		

- Four cases
  - Realizable vs. Agnostic
    - Realizable → $c^* \in \mathcal{H}$
    - Agnostic → c^* might or might not be in $\mathcal{H}$
  - Finite vs. Infinite
    - Finite → $|\mathcal{H}| < \infty$
    - Infinite → $|\mathcal{H}| = \infty$

Theorem 1: Finite, Realizable Case

- Consider a finite hypothesis set $\mathcal{H}$ s.t. $\underline{c^* \in \mathcal{H}}$ and arbitrary distribution p^* . If the number of labelled training data points (sampled i.i.d from p^*) satisfies

$$\text{\# of samples} \leftarrow \boxed{M} \geq \underline{\frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)}$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \underline{\leq \epsilon}$.

Proof of Theorem 1: Finite, Realizable Case

Proof strategy:

- For any given M, we will provide an upper bound on the probability of $\exists h \in \mathcal{H}$ s.t. $\hat{R}(h) = 0$ but $R(h) > \epsilon$.
- We show that if M is "large enough", this upper bound $\leq \delta$
- From there, we will conclude that w.p. $1-\delta$, any $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ has $R(h) < \epsilon$.

Proof of Theorem 1: Finite, Realizable Case

- Consider a "bad" hypothesis $h \in \mathcal{H}$ s.t. $\hat{R}(h) = 0$ but $R(h) > \varepsilon$.
- The probability that $\hat{R}(h) = 0$ (given $R(h) > \varepsilon$) is $\leq (1-\varepsilon)^M$ b.c.
  - The prob of h misclassifying a randomly selected $x^{(i)} \sim p^*$ is going to be at least ε.
  - Therefore, the prob of h correctly classifying $x^{(i)}$ is at most $(1-\varepsilon)$
  - The prob that h correctly classifies all M samples in D is at most $(1-\varepsilon)^M$.

Proof of Theorem 1: Finite, Realizable Case

Suppose we have K hypotheses $\overset{h_1, \dots, h_K}{}$ in $\mathcal{H}$ with $R(h) > \varepsilon$.

The probability that at least one of them achieves 0 empirical error (i.e. $\hat{R}(h) = 0$) is going to be $\le \underbrace{K}(1-\varepsilon)^M$.

$$\rightarrow P(\underbrace{\hat{R}(h_1) = 0} \vee \underbrace{\hat{R}(h_2) = 0} \vee \dots \vee \underbrace{\hat{R}(h_K) = 0})$$

$$\le \sum_{i=1}^{K} \underbrace{P(\hat{R}(h_i) = 0)} \qquad \text{(union bound)}$$

$$\le K(1-\varepsilon)^M \le \boxed{|\mathcal{H}|(1-\varepsilon)^M}$$

$$\boxed{\begin{aligned} P(A \cup B) &= P(A) + P(B) - \underbrace{P(A \cap B)}_{0 \le} \\ &\le P(A) + P(B) \end{aligned}}$$

Proof of Theorem 1: Finite, Realizable Case

$$\forall \varepsilon \in \mathbb{R} \quad \underbrace{(1-\varepsilon) \leq e^{-\varepsilon}}$$

$$\text{Prob}(\ldots \text{bad event} \ldots) \leq |\mathcal{H}|(1-\varepsilon)^M$$

$$\leq |\mathcal{H}| e^{-\varepsilon M}$$

Proof of Theorem 1: Finite, Realizable Case

We want this probability (of bad event) to be bounded by δ:

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

$$\Leftrightarrow \frac{|\mathcal{H}|}{\delta} \leq e^{\epsilon M}$$

$$\Leftrightarrow \ln|\mathcal{H}| + \ln\left(\frac{1}{\delta}\right) \leq \epsilon M$$

$$\Leftrightarrow \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right) \leq M \qquad \blacksquare$$

Theorem 1: Finite, Realizable Case

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$\underline{M} \geq \underbrace{\boxed{\frac{1}{\epsilon}}\left(\ln(|\mathcal{H}|) + \ln\left(\boxed{\frac{1}{\delta}}\right)\right)}$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- Making the bound tight (setting the two sides equal to each other) and solving for ϵ gives...

Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = \underbrace{M}$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$\underbrace{R(h)}_{\text{true error}} \leq \underline{\frac{1}{M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

Theorem 2: Finite, Agnostic Case

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

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

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ satisfy $\left|R(h) - \hat{R}(h)\right| \leq \epsilon$

- Bound is inversely quadratic in ϵ, e.g., halving ϵ means we need four times as many labelled training data points
- Again, making the bound tight and solving for ϵ gives…

Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$\underset{|R(h) - \hat{R}(h)|}{R(h) \leq \hat{R}(h)} + \underbrace{\sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}}_{\varepsilon}$$

with probability at least $1 - \delta$.

What happens when $|\mathcal{H}| = \infty$?

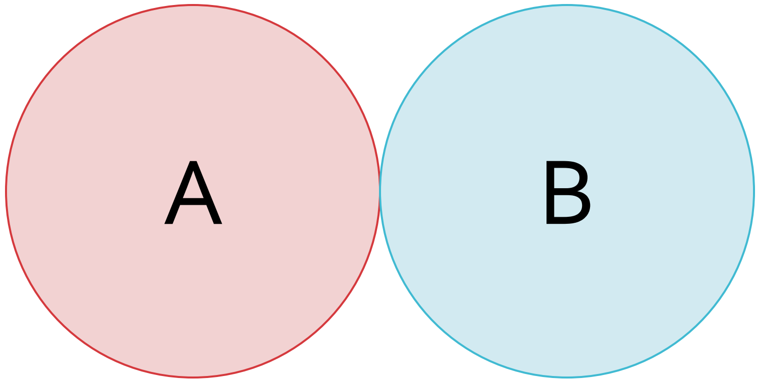
- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

The Union Bound…

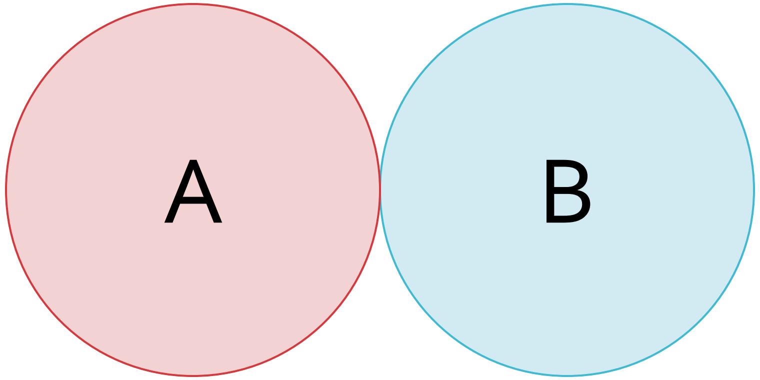
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

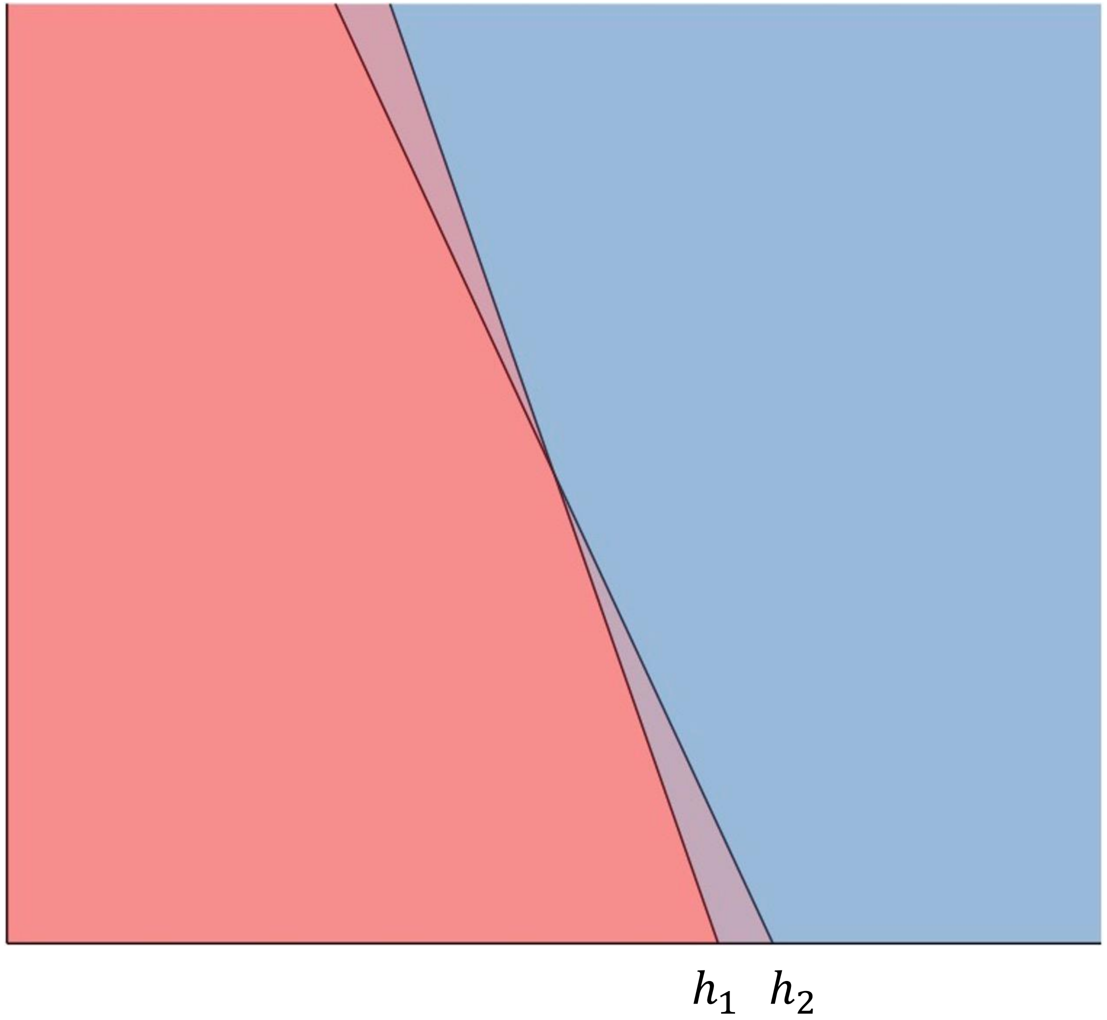


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!

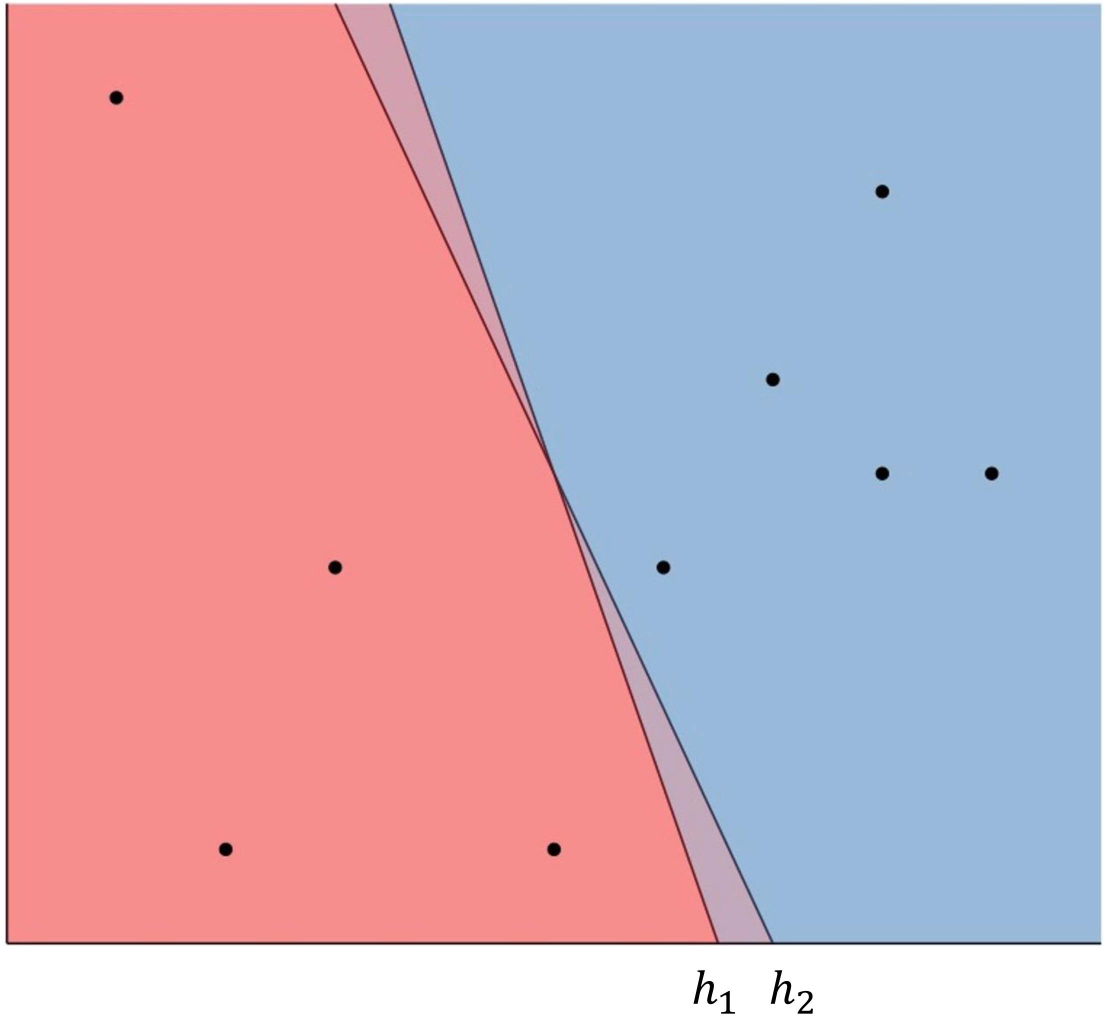


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!



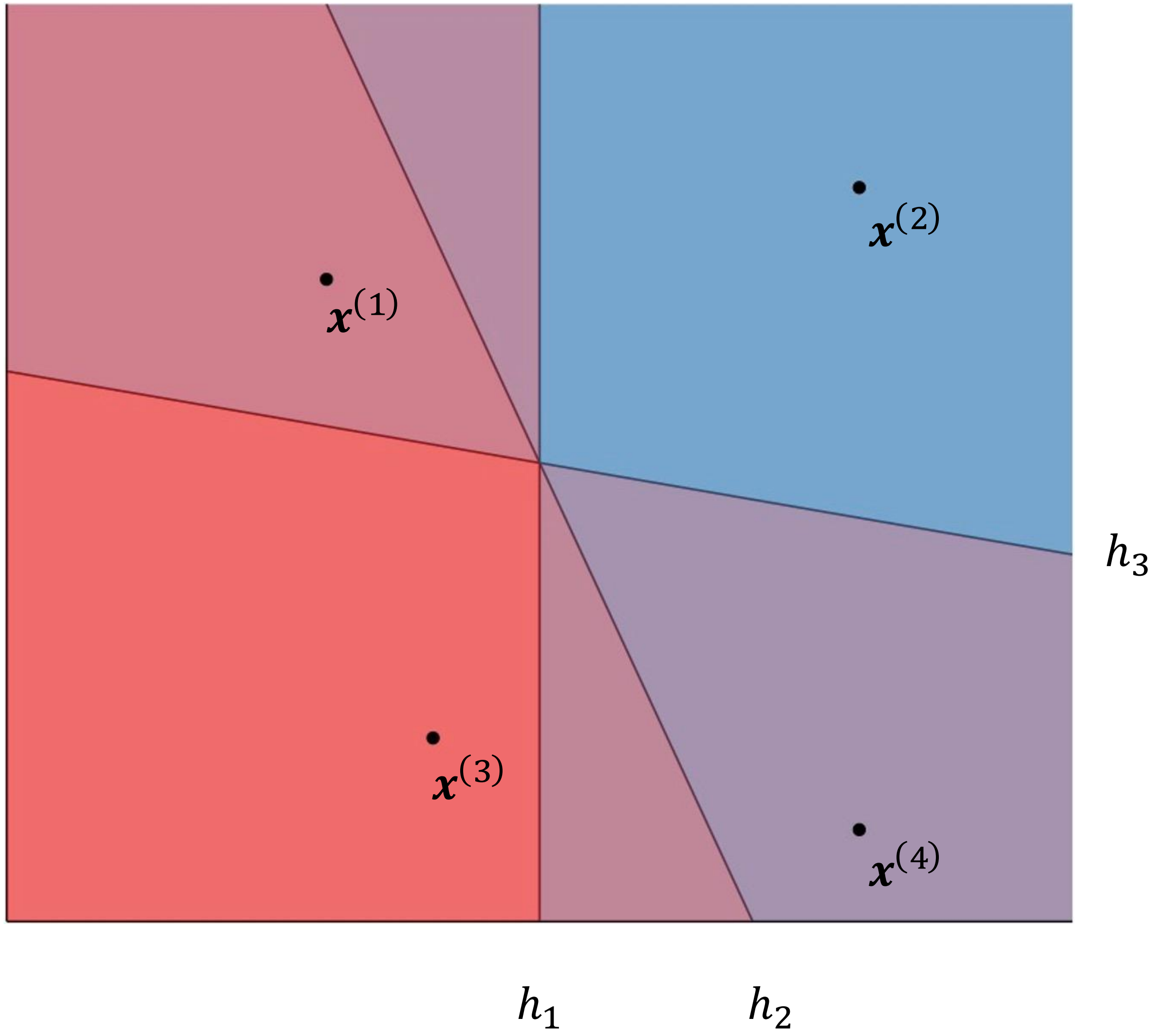
Labellings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right)$ is a vector of M +1's and -1's

- Insight: given $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$, each hypothesis in $\mathcal{H}$ induces a labelling *but not necessarily a unique labelling*
  - The set of labellings induced by $\mathcal{H}$ on S is $\mathcal{H}(S) = \left\{\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right) \middle| h \in \mathcal{H}\right\}$

Example: Labellings

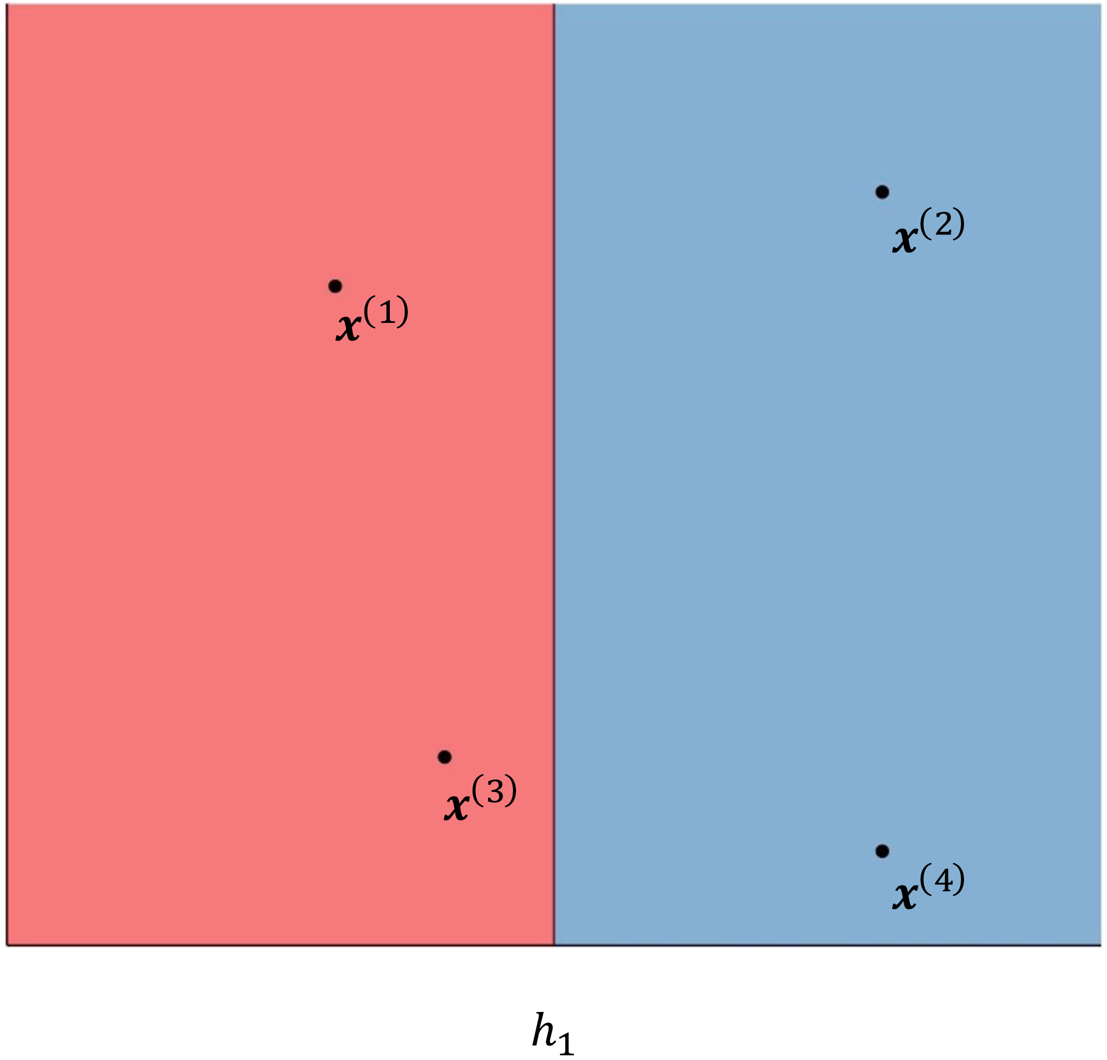
$\mathcal{H} = \{h_1, h_2, h_3\}$



Example: Labellings

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

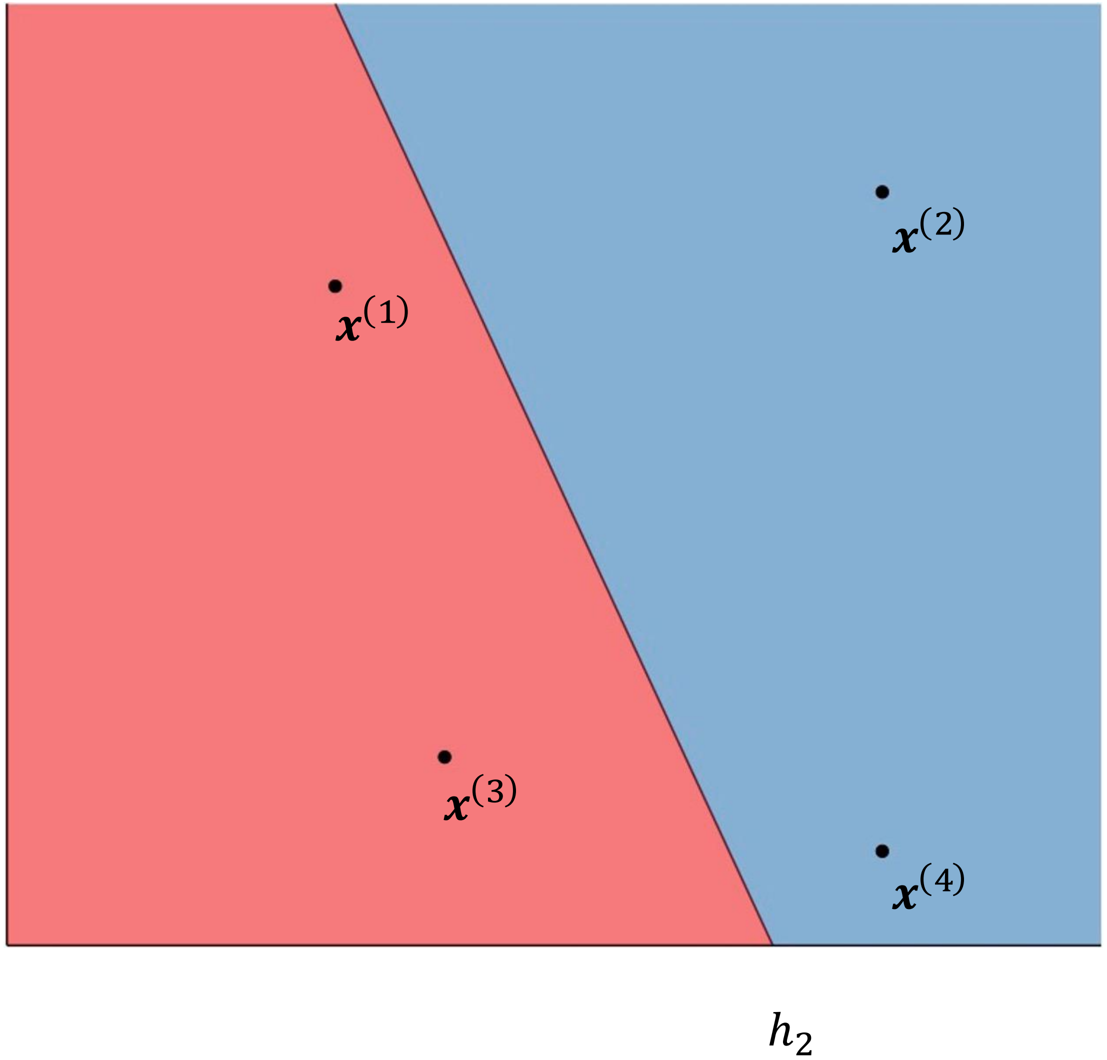
$$\left(h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})\right)$$
$$= (-1, +1, -1, +1)$$



Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

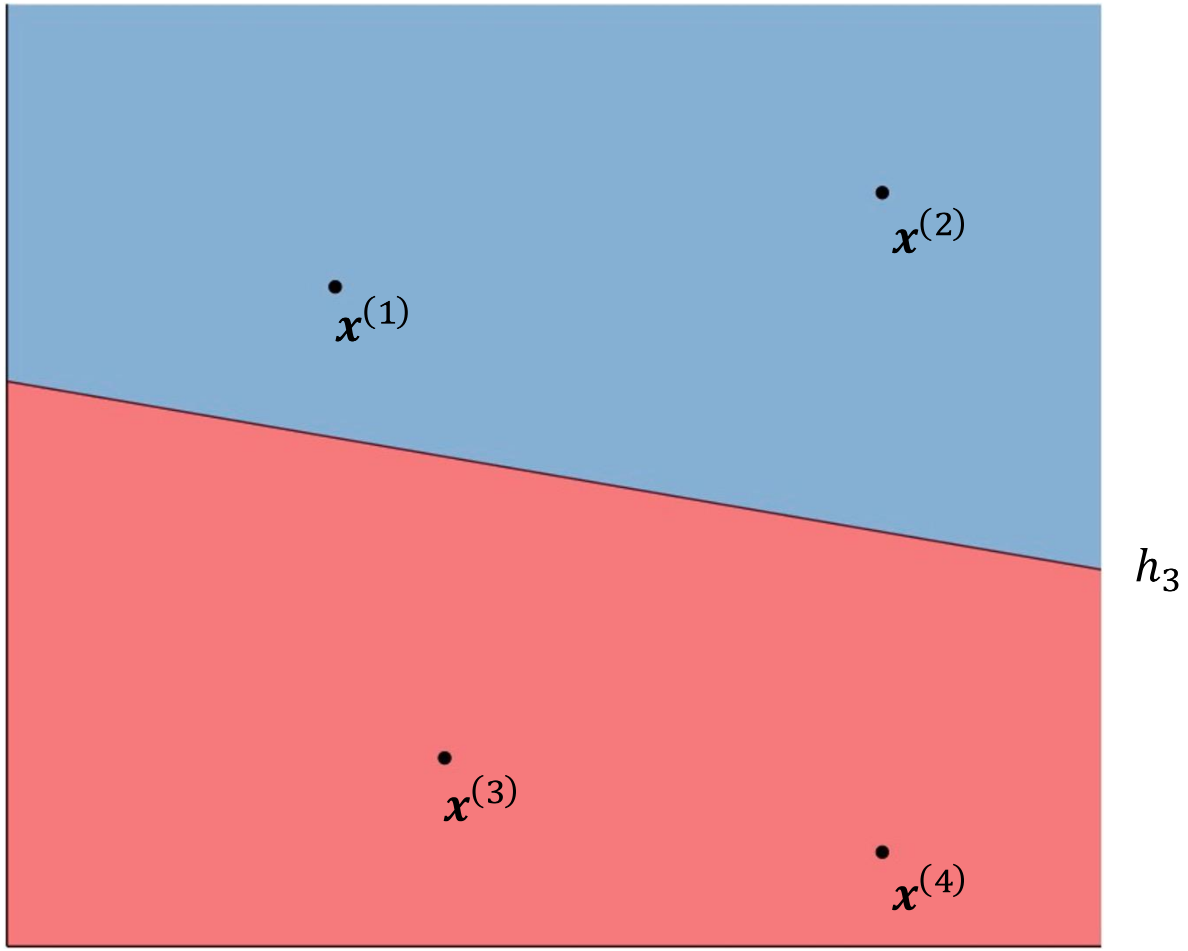
$$\left(h_2\left(\boldsymbol{x}^{(1)}\right), h_2\left(\boldsymbol{x}^{(2)}\right), h_2\left(\boldsymbol{x}^{(3)}\right), h_2\left(\boldsymbol{x}^{(4)}\right)\right) = (-1, +1, -1, +1)$$



Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_3\left(\boldsymbol{x}^{(1)}\right), h_3\left(\boldsymbol{x}^{(2)}\right), h_3\left(\boldsymbol{x}^{(3)}\right), h_3\left(\boldsymbol{x}^{(4)}\right)\right)$$
$$= (+1, +1, -1, -1)$$

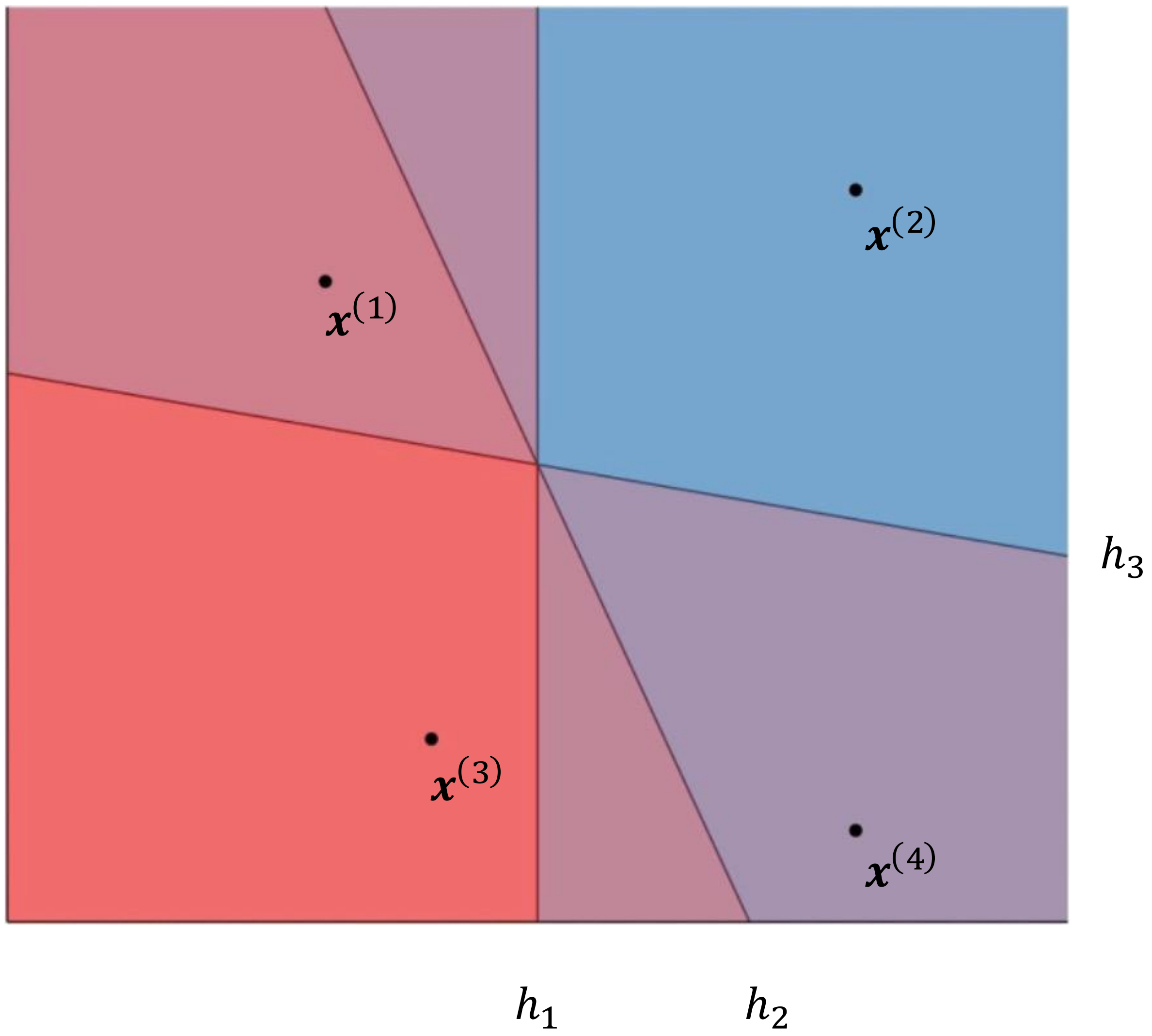


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S)$
$= \{(+1, +1, -1, -1), (-1, +1, -1, +1)\}$

$|\mathcal{H}(S)| = 2$

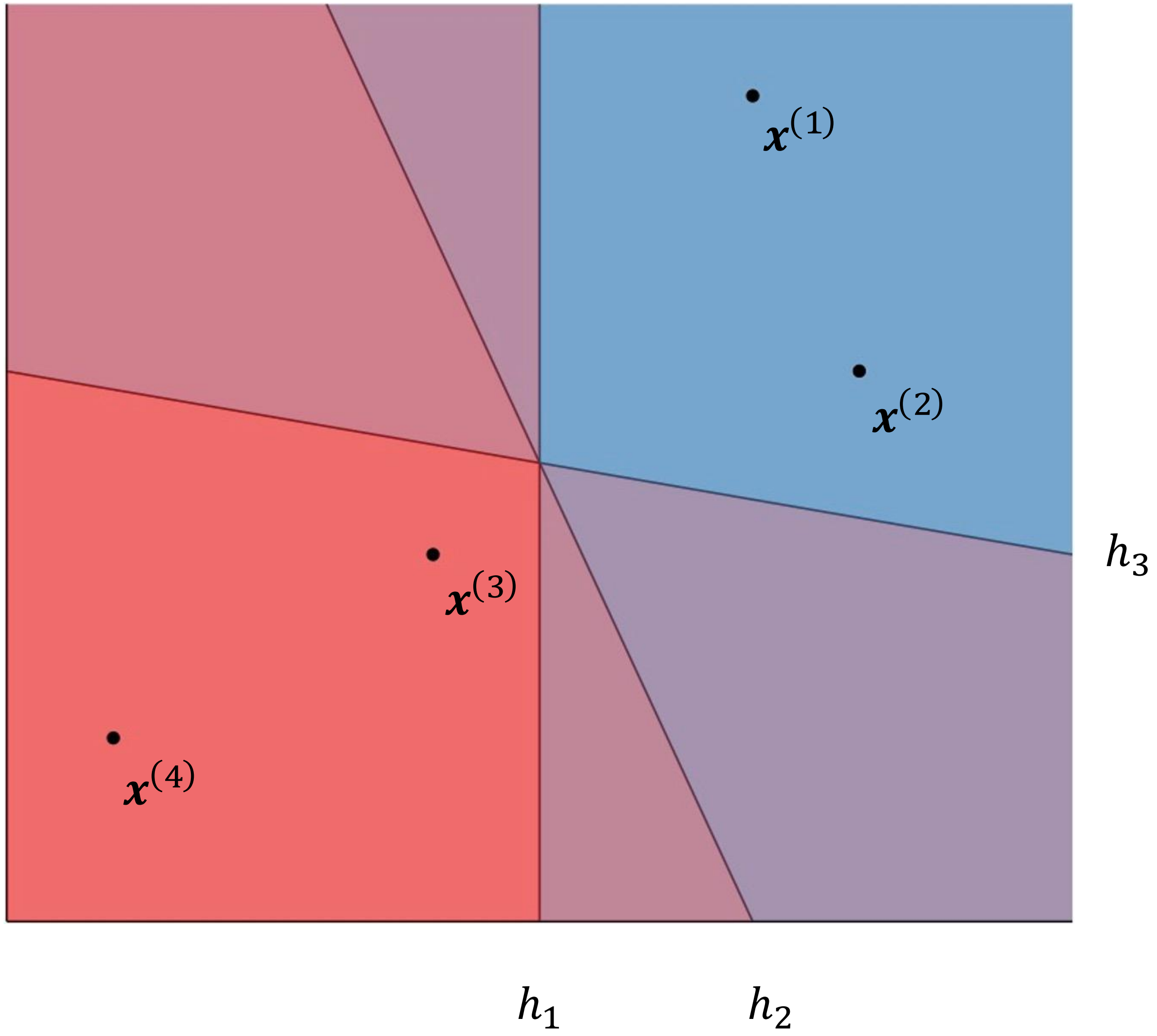


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S) =$
$\{(+1, +1, -1, -1)\}$

$|\mathcal{H}(S)| = 1$



Key Takeaways

- Statistical learning theory model
- Expected vs. empirical risk of a hypothesis
- Four possible cases of interest
  - realizable vs. agnostic
  - finite vs. infinite
- Sample complexity bounds and statistical learning theory corollaries for finite hypothesis sets