10-301/601: Introduction to Machine Learning Lecture 15 – Learning Theory (Finite Case)

Henry Chai

10/24/22

Front Matter

- Announcements
  - HW5 released 10/13, due 10/27 at 11:59 PM
  - Exam 3 scheduled
    - Thursday, December 15th from 9:30 AM to 11:30 AM
  - Sign up for peer tutoring! See Piazza for more details.

Q & A:

Where have you been???

- Sorry, I've been training a fresh neural network…

Q & A:

My HW5 code isn't working, what should I do???

- Review the recitation material!
  - Specifically, test your implementation against the numerical examples our TAs worked through and make sure you're getting the same values

ML Big Picture

Learning Paradigms:
What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation:
What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Application Areas
Key challenges?
NLP, Speech, Computer Vision, Robotics, Medicine, Search

Theoretical Foundations:
What principles guide learning?

- ❑ probabilistic
- ❑ information theoretic
- ❑ evolutionary search
- ❑ ML as optimization

Facets of Building ML Systems:
How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

Big Ideas in ML:
Which are the ideas driving development of the field?

- *inductive bias*
- *generalization / overfitting*
- *bias-variance decomposition*
- *generative vs. discriminative*
- *deep nets, graphical models*
- *PAC learning*
- *distant rewards*

ML Big Picture

Learning Paradigms:

What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation:

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Application Areas

Key challenges?

NLP, Speech, Computer Vision, Robotics, Medicine, Search

Theoretical Foundations:

What principles guide learning?

- ❑ probabilistic
- ❑ information theoretic
- ❑ evolutionary search
- ❑ ML as optimization

Facets of Building ML Systems:

How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

Big Ideas in ML:

Which are the ideas driving development of the field?

- *inductive bias*
- *generalization / overfitting*
- *bias-variance decomposition*
- *generative vs. discriminative*
- *deep nets, graphical models*
- ***PAC learning***
- *distant rewards*

Statistical Learning Theory Model

1. Data points are generated iid from some *unknown* distribution
$$\boldsymbol{x}^{(i)} \sim p^*(\boldsymbol{x})$$
2. Labels are generated from some *unknown* function
$$y^{(i)} = c^*\left(\boldsymbol{x}^{(i)}\right)$$
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

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
- Test error rate
  - Used to estimate hypothesis performance
  - Good estimate of your hypothesis's true error
- Validation error rate
  - Used to set hypothesis hyperparameters
  - Slightly "optimistic" estimate of your hypothesis's true error
- Training error rate
  - Used to set model parameters
  - Very "optimistic" estimate of your hypothesis's true error

Types of Risk

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

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

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

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

where $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ is the training data set and $\boldsymbol{x} \sim \mathcal{D}$ denotes a point sampled uniformly at random from $\mathcal{D}$

Three Functions of Interest

- The *true function*, c^*

- The *expected risk minimizer*,

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

- The *empirical risk minimizer*,

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

Poll Question 1: Which of the following are *always* true?

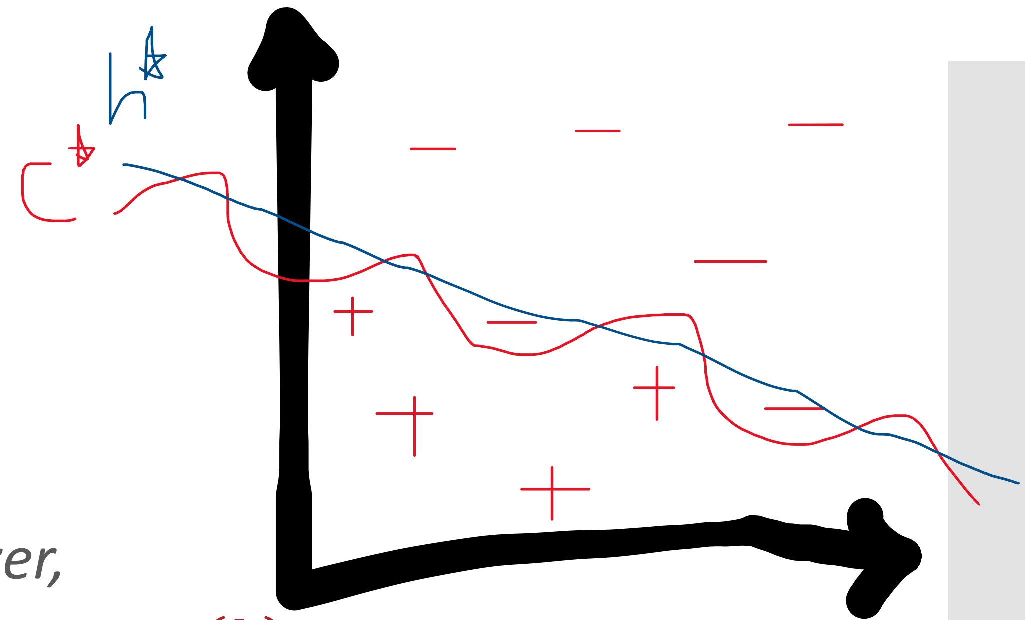
A. $c^* = h^*$
B. $c^* = \hat{h}$
C. $h^* = \hat{h}$
D. $c^* = h^* = \hat{h}$
E. None of the above
F. **TOXIC**

- The *true function*, c^*
- The *expected risk minimizer*,

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

- The *empirical risk minimizer*,

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



If H is all linear separators,
Then…

Key Question

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

PAC Learning

- The sample complexity of an algorithm/hypothesis set is the number of labelled training data points needed to satisfy the PAC criterion for some δ and ϵ
- PAC = **<u>P</u>**robably **<u>A</u>**pproximately **<u>C</u>**orrect
- PAC Criterion:

$$P\left(\left|R(h) - \hat{R}(h)\right| \leq \epsilon\right) \geq 1 - \delta \; \forall \, h \in \mathcal{H}$$

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 an algorithm/hypothesis set is the number of labelled training data points needed to satisfy the PAC criterion for some δ and ϵ
- 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

- 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

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

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

Proof of Theorem 1: Finite, Realizable Case

1. Assume there are K "bad" hypotheses in $\mathcal{H}$, i.e., $h_1, h_2, \dots, h_K$ that all have $R(h_k) > \epsilon$
2. Pick one bad hypothesis, h_k
   A. Probability that h_k correctly classifies the first training data point $\leq 1 - \epsilon$
   B. Probability that h_k correctly classifies all M training data points $\leq (1 - \epsilon)^M$
3. Probability that at least one bad hypothesis correctly classifies all M training data points =

$$P(h_1 \text{ correctly classifies all } M \text{ training data points} \cup \\ h_2 \text{ correctly classifies all } M \text{ training data points} \cup \\ \vdots \\ \cup\, h_K \text{ correctly classifies all } M \text{ training data points})$$

Proof of Theorem 1: Finite, Realizable Case

$$P(h_1 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$h_2 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$\vdots$$
$$\cup\, h_K \text{ correctly classifies all } M \text{ training data points})$$

$$\leq \sum_{k=1}^{K} P(h_k \text{ correctly classifies all } M \text{ training data points})$$

by the union bound: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$\leq P(A) + P(B)$

Proof of Theorem 1: Finite, Realizable Case

$$\sum_{k=1}^{K} P(h_k \text{ correctly classifies all } M \text{ training data points})$$

$$\leq k(1-\epsilon)^M \leq |\mathcal{H}|(1-\epsilon)^M$$

because $k \leq |\mathcal{H}|$

3. Probability that at least one bad hypothesis correctly classifies all M training data points $\leq |\mathcal{H}|(1-\epsilon)^M$
4. Using the fact that $1 - x \leq \exp(-x)\ \forall\, x$,
   $$|\mathcal{H}|(1-\epsilon)^M \leq |\mathcal{H}| \exp(-\epsilon)^M = |\mathcal{H}| \exp(-M\epsilon)$$
5. Probability that at least one bad hypothesis correctly classifies all M training data points $\leq |\mathcal{H}| \exp(-M\epsilon)$, which we want to be *low*, i.e., $|\mathcal{H}| \exp(-M\epsilon) \leq \delta$

Proof of Theorem 1: Finite, Realizable Case

$$|\mathcal{H}| \exp(-M\epsilon) \leq \delta \rightarrow \exp(-M\epsilon) \leq \frac{\delta}{|\mathcal{H}|}$$

$$\rightarrow -M\epsilon \leq \log\left(\frac{\delta}{|\mathcal{H}|}\right)$$

$$\rightarrow M \geq \frac{1}{\epsilon}\left(-\log\left(\frac{\delta}{|\mathcal{H}|}\right)\right)$$

$$\rightarrow M \geq \frac{1}{\epsilon}\left(\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right)$$

$$\rightarrow M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$$

Proof of Theorem 1: Finite, Realizable Case

6. Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that ∃ a bad hypothesis $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ and $\hat{R}(h_k) = 0$ is $\leq \delta$

$$\Updownarrow$$

Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ have $\hat{R}(h_k) > 0$ is $\geq 1 - \delta$

Proof of Theorem 1: Finite, Realizable Case

6. Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ have $\hat{R}(h_k) > 0$ is $\geq 1 - \delta$

$$\Updownarrow$$

Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $\hat{R}(h_k) = 0$ have $R(h_k) \leq \epsilon$ is $\geq 1 - \delta$

(proof by contrapositive)

Aside: Proof by Contrapositive

- The contrapositive of a statement $A \Rightarrow B$ is $\neg B \Rightarrow \neg A$
- A statement and its contrapositive are logically equivalent, i.e., $A \Rightarrow B$ means that $\neg B \Rightarrow \neg A$
- Example: "it's raining $\Rightarrow$ Henry brings an umbrella"

  is the same as saying

  "Henry didn't bring an umbrella $\Rightarrow$ it's not raining "

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

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

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

Poll Question 2:

- Let $\mathcal{H}$ be the set of all conjunctions over M Boolean variables, $\boldsymbol{x} \in \{0,1\}^M$; examples of conjunctions are
  - $h(\boldsymbol{x}) = x_1(1 - x_2)x_4x_{10}$
  - $h(\boldsymbol{x}) = (1 - x_3)(1 - x_4)x_8$
- Assuming $c^* \in \mathcal{H}$, if $M = 10$, $\epsilon = 0.1$, and $\delta = 0.01$, at least how many labelled examples do we need to satisfy the PAC criterion using Theorem 1?

A. 1 **(TOXIC)**
B. $10(2\ln 10 + \ln 100) \approx 92$
C. $10(3\ln 10 + \ln 100) \approx 116$
D. $10(10\ln 2 + \ln 100) \approx 116$
E. $10(10\ln 3 + \ln 100) \approx 156$
F. $100(2\ln 10 + \ln 10) \approx 691$
G. $100(3\ln 10 + \ln 10) \approx 922$
H. $100(10\ln 2 + \ln 10) \approx 924$
I. $100(10\ln 3 + \ln 10) \approx 1329$

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

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

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

- Solving for ϵ gives...

Statistical Learning Theory Corollary 1

- 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. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

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

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

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

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
- Solving for ϵ gives…

Statistical Learning Theory Corollary 2

- 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 $1 - \delta$.

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 $1 - \delta$.

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

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

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

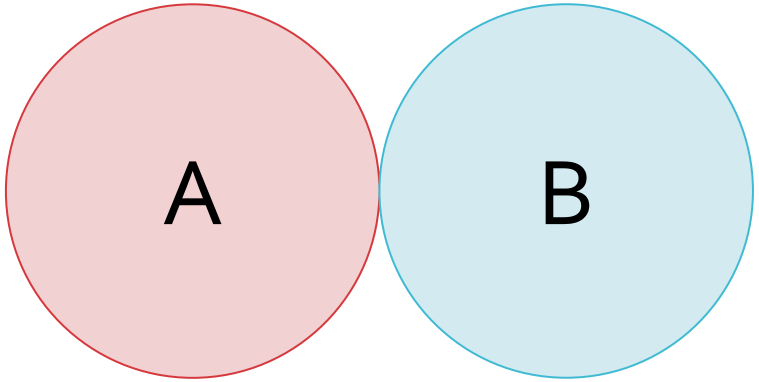
$$R(h) \leq \hat{R}(h) + \infty \quad \text{(not a very meaningful result…)}$$

with probability at least $1 - \delta$.

- Insight: $|\mathcal{H}|$ measures how complex our hypothesis set is
- Idea: define a different measure of hypothesis set complexity

The Union Bound…

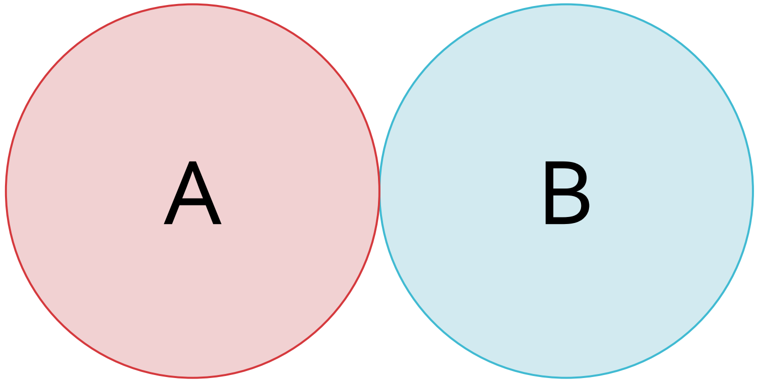
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

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

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

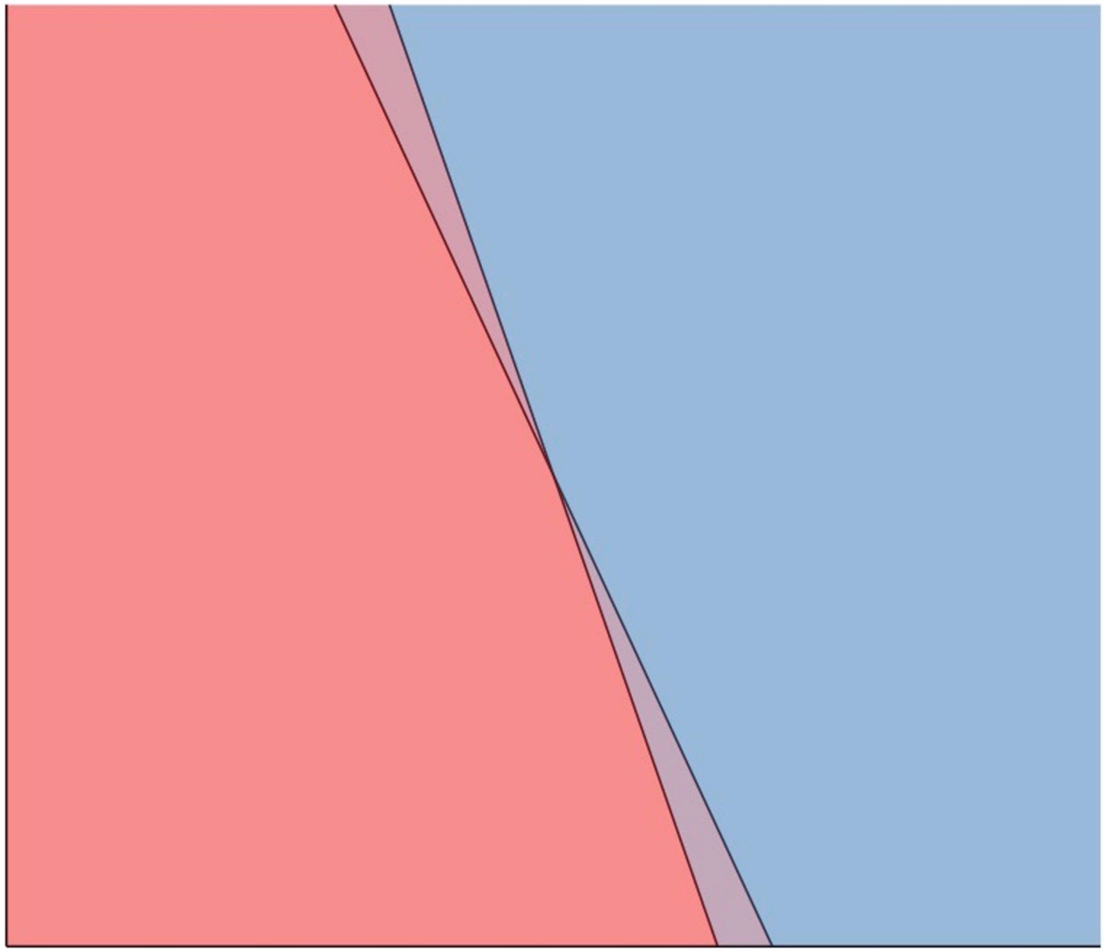


Intuition

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

- "h_1 is consistent with all M training data points"
- "h_2 is consistent with all M training data points"

will overlap a lot!

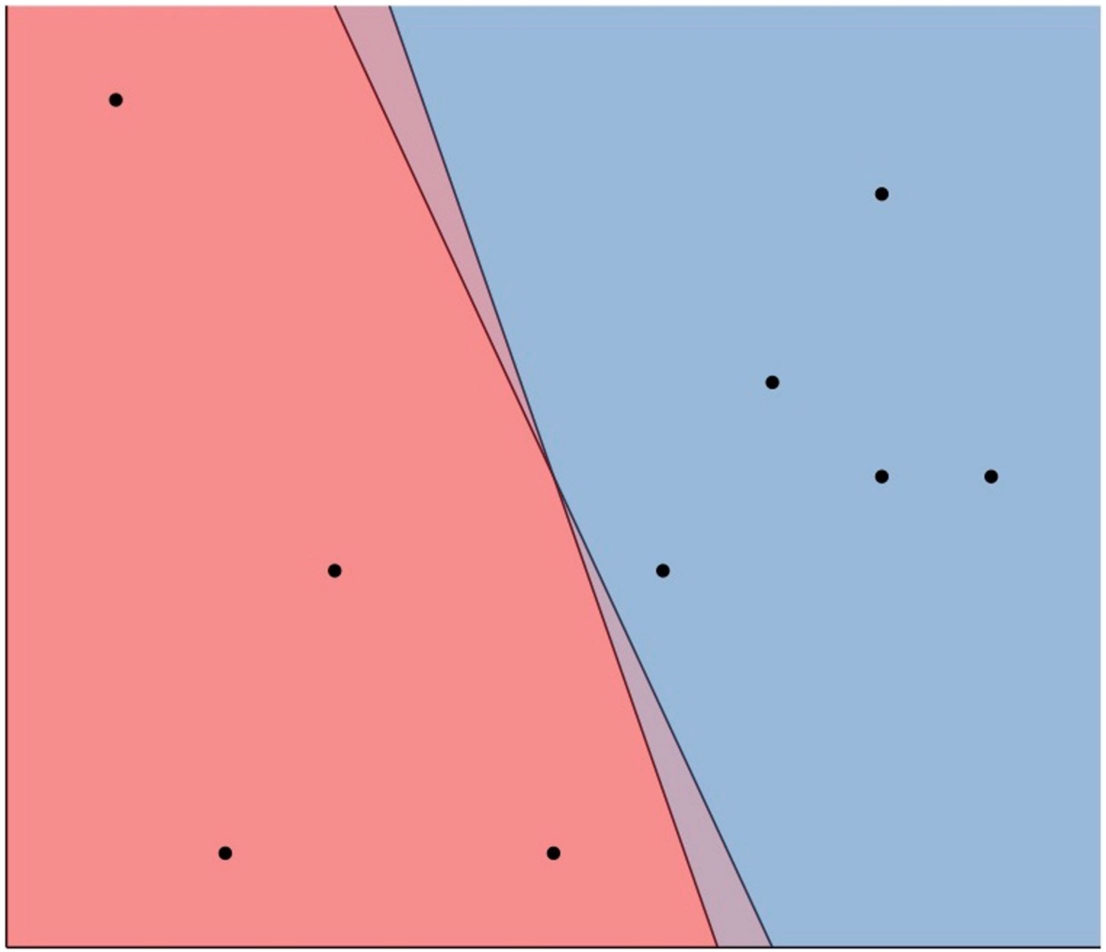


Intuition

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

- "h_1 is consistent with all M training data points"
- "h_2 is consistent with all M training data points"

will overlap a lot!



Theorem 3: Vapnik-Chervonenkis (VC)-Bound

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution p^*, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon}\left(d_{VC}(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

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

- $d_{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$, a measure of how complex our hypothesis set is, suitable when $|\mathcal{H}| = \infty$