Understanding Machine Learning Theory: Part I

Seyed Iman Mirzadeh

Washington State University

June 2021

Before We Start

- There will be lots of new definitions. However, you are already familiar with most of them. But maybe here the notations/names are different.

Before We Start

- There will be lots of new definitions. However, you are already familiar with most of them. But maybe here the notations/names are different.
- So, *please*, stop me if you don't remember the definition of a variable!

Before We Start

- There will be lots of new definitions. However, you are already familiar with most of them. But maybe here the notations/names are different.
- So, *please*, stop me if you don't remember the definition of a variable!
- Also obviously, please stop me if you don't understand an explanation. This is a rather long presentation. The ambiguity is only going build up! So you might suffer from boredom if you don't ask your questions and ask for clarification!

Understanding Machine Learning Theory

So, I decided to learn more about ML Theory. In my opinion, this book [at least so far!] explains the basics of theory very well.

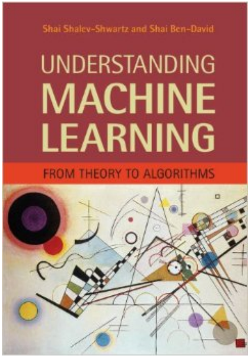

Shai Shalev-Shwartz and Shai Ben-David
UNDERSTANDING
MACHINE
LEARNING
FROM THEORY TO ALGORITHMS

Figure: from
`https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/`

Outline

Introduction

- Statistical Learning Framework
- Measures of Success

Empirical Risk Minimization (ERM)

- Overfitting
- Inductive Bias
- Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias

- Mathematical Setup
- ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

Outline

Introduction
- Statistical Learning Framework
- Measures of Success

Empirical Risk Minimization (ERM)
- Overfitting
- Inductive Bias
- Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias
- Mathematical Setup
- ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

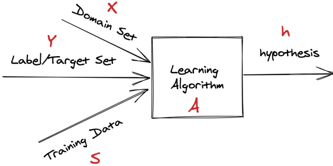
Statistical Learning Framework

Learner's Input:

- **Domain Set** (Input Space): Set of all possible examples/instances we wish to label, shown by X.
- **Label Set** (Target Space): Set of all possible labels, shown by Y.
- **Sample** (Training Data): A finite sequence of pairs in $X \times Y$ shown by $S = ((x_1, y_1), \cdots, (x_m, y_m))$.

Lerner's Output:

- **Hypothesis**: The learner outputs a mapping function $h : X \to Y$ that can assign a value to all $x \in X$. Another notation for the hypothesis can be $A(S)$ which means the output of the learning algorithm A, upon receiving the training sequence S. Also, we might show the hypothesis learned on training data S by $h_S : X \to Y$.



Statistical Learning Framework (2)

Assumption about data generation model:

Statistical Learning Framework (2)

Assumption about data generation model:

1. The instances of training data, S , is generated using a probability distribution $\mathcal{D}$ over X.

Statistical Learning Framework (2)

Assumption about data generation model:

1. The instances of training data, S , is generated using a probability distribution $\mathcal{D}$ over X.
2. The labels are generated using a target function $f : X \rightarrow Y$, that is $f(x_i) = y_i,\ \forall x_i \in S$

Statistical Learning Framework (2)

Assumption about data generation model:

1. The instances of training data, S , is generated using a probability distribution $\mathcal{D}$ over X.
2. The labels are generated using a target function $f : X \rightarrow Y$, that is $f(x_i) = y_i, \ \forall x_i \in S$
3. The learner doesn't know anything about $\mathcal{D}$ and only observes sample S.

Measures of Success

Measures of Success

Definition (True Risk/Error, or Generalization Error)

The probability to draw a random instance $x \sim \mathcal{D}$, such that $h(x) \neq f(x)$:

$$\boxed{L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}[\{x : h(x) \neq f(x)\}]} \qquad (1)$$

Measures of Success

Definition (True Risk/Error, or Generalization Error)

The probability to draw a random instance $x \sim \mathcal{D}$, such that $h(x) \neq f(x)$:

$$\boxed{L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}[\{x : h(x) \neq f(x)\}]} \tag{1}$$

Definition (Empirical Risk/Error, or Training Error)

A measure for the risk/error of the learner's hypothesis on the sample S.

$$\boxed{L_S(h) = \frac{|i \in [m] : h(x_i) \neq y_i|}{m}} \tag{2}$$

Measures of Success

Definition (True Risk/Error, or Generalization Error)

The probability to draw a random instance $x \sim \mathcal{D}$, such that $h(x) \neq f(x)$:

$$\boxed{L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}[\{x : h(x) \neq f(x)\}]} \tag{1}$$

Definition (Empirical Risk/Error, or Training Error)

A measure for the risk/error of the learner's hypothesis on the sample S.

$$\boxed{L_S(h) = \frac{|i \in [m] : h(x_i) \neq y_i|}{m}} \tag{2}$$

Note that the the learner doesn't have access to $\mathcal{D}$ and can only see sample S.

Outline

Introduction

- Statistical Learning Framework
- Measures of Success

Empirical Risk Minimization (ERM)

- Overfitting
- Inductive Bias
- Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias

- Mathematical Setup
- ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

Empirical Risk Minimization (ERM)

Definition

Since the training sample is the snapshot of the world that is available to the learner, it makes sense to search for a solution that works well on that data.
This learning paradigm – coming up with a hypothesis h that minimizes $L_S(h)$ – is called *Empirical Risk Minimization*.

Papayas Example

This is a running example throughout the first few chapters:

Example

Imagine you have just arrived in some small Pacific island. You soon become familiar with a new fruit that you have never tasted before, called *Papaya*! You have to learn how to predict whether a papaya you see in the market is tasty or not

Papayas Example

This is a running example throughout the first few chapters:

Example

Imagine you have just arrived in some small Pacific island. You soon become familiar with a new fruit that you have never tasted before, called *Papaya*! You have to learn how to predict whether a papaya you see in the market is tasty or not

Overfitting

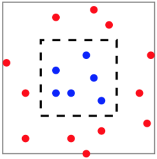
Assume in the Papayas Example, we come up with the idea of classifying papayas into two categories ($1 =$ tasty, $0 =$ not tasty) using two features: softness and color.

Overfitting

Assume in the Papayas Example, we come up with the idea of classifying papayas into two categories ($1 =$ tasty, $0 =$ not tasty) using two features: softness and color.

Now, assume that the samples are coming from distribution $\mathcal{D}$ such that the instances are distributed uniformly within the gray square below.

Also, assume the true labeling function f is such that it assigns 1 if an instance is within the inner dashed square, and 0 otherwise. We assume the area of the inner circle equals 1 and the area of the gray square is 2.



Overfitting(2)

Now, let's say we are feeling too smart and come up with this hypothesis:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] : x_i = x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Overfitting(2)

Now, let's say we are feeling too smart and come up with this hypothesis:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] : x_i = x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

i.e., I memorize everything that I have seen and output the same label as in my memory, otherwise, I will output 0.

Overfitting(2)

Now, let's say we are feeling too smart and come up with this hypothesis:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] : x_i = x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

i.e., I memorize everything that I have seen and output the same label as in my memory, otherwise, I will output 0.

Clearly I have minimized the empirical risk ($L_S(h) = 0$). But what about the true risk?

Overfitting(2)

Now, let's say we are feeling too smart and come up with this hypothesis:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] : x_i = x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

i.e., I memorize everything that I have seen and output the same label as in my memory, otherwise, I will output 0.
Clearly I have minimized the empirical risk ($L_S(h) = 0$). But what about the true risk?

$$\begin{aligned} L_{\mathcal{D},f}(h_S) &= \mathcal{D}[\{x : h_S(x) \neq f(x)\}] \\ &= \mathcal{D}[\{x : h_S(x) = 0, f(x) = 1\}] \\ &= \frac{\text{Area of inner circle}}{\text{Total area}} = \frac{1}{2} \end{aligned}$$

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!
- A common way, is to **restrict** the learner choose in advance (before seeing the data) a set of predictors. This set of predictors called a *hypothesis class* and denoted by $\mathcal{H}$.

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!
- A common way, is to **restrict** the learner choose in advance (before seeing the data) a set of predictors. This set of predictors called a *hypothesis class* and denoted by $\mathcal{H}$.
- Each hypothesis $h \in \mathcal{H}$ is a function of form $h : X \mapsto Y$. Then for a given class $\mathcal{H}$, and a training sample S we define:

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!
- A common way, is to **restrict** the learner choose in advance (before seeing the data) a set of predictors. This set of predictors called a *hypothesis class* and denoted by $\mathcal{H}$.
- Each hypothesis $h \in \mathcal{H}$ is a function of form $h : X \mapsto Y$. Then for a given class $\mathcal{H}$, and a training sample S we define:

$$\boxed{\mathrm{ERM}_{\mathcal{H}} \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)} \tag{4}$$

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!
- A common way, is to **restrict** the learner choose in advance (before seeing the data) a set of predictors. This set of predictors called a *hypothesis class* and denoted by $\mathcal{H}$.
- Each hypothesis $h \in \mathcal{H}$ is a function of form $h : X \mapsto Y$. Then for a given class $\mathcal{H}$, and a training sample S we define:

$$\boxed{\mathrm{ERM}_{\mathcal{H}} \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)} \tag{4}$$

- By restricting the learner to choosing a predictor from $\mathcal{H}$, we *bias* it toward a particular set of predictors. Such restrictions are often called an *inductive bias*.

Inductive Bias

- As we saw, the ERM rule might lead to overfitting. How to fix it?
- We should look for conditions that guarantees ERM doesn't overfit!
- A common way, is to **restrict** the learner choose in advance (before seeing the data) a set of predictors. This set of predictors called a *hypothesis class* and denoted by $\mathcal{H}$.
- Each hypothesis $h \in \mathcal{H}$ is a function of form $h : X \mapsto Y$. Then for a given class $\mathcal{H}$, and a training sample S we define:

$$\boxed{\mathrm{ERM}_{\mathcal{H}} \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)} \tag{4}$$

- By restricting the learner to choosing a predictor from $\mathcal{H}$, we *bias* it toward a particular set of predictors. Such restrictions are often called an *inductive bias*.

Finite Hypothesis Class

- The simplest type of restriction on a class is imposing an upper bound on its size (i.e, the number of predictors $h \in \mathcal{H}$). Such a class called a *finite hypothesis class*.

Finite Hypothesis Class

- The simplest type of restriction on a class is imposing an upper bound on its size (i.e, the number of predictors $h \in \mathcal{H}$). Such a class called a *finite hypothesis class*.
- Example: let $\mathcal{H}$ be the class of all single neuron networks with 2 parameters (one weight and one bias). what is $|\mathcal{H}|$?

Finite Hypothesis Class

- The simplest type of restriction on a class is imposing an upper bound on its size (i.e, the number of predictors $h \in \mathcal{H}$). Such a class called a *finite hypothesis class*.
- Example: let $\mathcal{H}$ be the class of all single neuron networks with 2 parameters (one weight and one bias). what is $|\mathcal{H}|$?
- The finite class restriction seems to be a strong assumption. But in practice it's not. Why?

Finite Hypothesis Class

- The simplest type of restriction on a class is imposing an upper bound on its size (i.e, the number of predictors $h \in \mathcal{H}$). Such a class called a *finite hypothesis class*.
- Example: let $\mathcal{H}$ be the class of all single neuron networks with 2 parameters (one weight and one bias). what is $|\mathcal{H}|$?
- The finite class restriction seems to be a strong assumption. But in practice it's not. Why?
- If we assume that we are using a computer to implement our algorithm, then each parameter/variable will have finite bits.

Outline

Introduction

Statistical Learning Framework

Measures of Success

Empirical Risk Minimization (ERM)

Overfitting

Inductive Bias

Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias

Mathematical Setup

ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

Mathematical Setup: Assumptions

Before we start, we need to have two assumptions for our anlysis:

Definition (The Realizability Assumption)

We assume that there exists a hypothesis $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$.

Mathematical Setup: Assumptions

Before we start, we need to have two assumptions for our anlysis:

Definition (The Realizability Assumption)

We assume that there exists a hypothesis $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$.

i.e., there exists a perfect hypothesis in our hypothesis class. We may not find this but at least we assume such a hypothesis exists.

Mathematical Setup: Assumptions

Before we start, we need to have two assumptions for our anlysis:

Definition (The Realizability Assumption)

We assume that there exists a hypothesis $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$.

i.e., there exists a perfect hypothesis in our hypothesis class. We may not find this but at least we assume such a hypothesis exists.

Definition (The i.i.d Assumption)

The examples in the training set (Sample S) are independently and identically distributed (i.i.d.) according to the distribution $\mathcal{D}$. (notation: $S \sim \mathcal{D}^m$)

Mathematical Setup: Assumptions

Before we start, we need to have two assumptions for our anlysis:

Definition (The Realizability Assumption)

We assume that there exists a hypothesis $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$.

i.e., there exists a perfect hypothesis in our hypothesis class. We may not find this but at least we assume such a hypothesis exists.

Definition (The i.i.d Assumption)

The examples in the training set (Sample S) are independently and identically distributed (i.i.d.) according to the distribution $\mathcal{D}$. (notation: $S \sim \mathcal{D}^m$)

i.e., every $x_i \in S$ is freshly sampled according to $\mathcal{D}$ and then labeled according to the labeling function, f.

Mathematical Setup: Analysis Parameters

- S is sampled randomly from $\mathcal{D}$. So, when the ERM tries to minimize the error on S, its output h_S is also a random variable. Since h_S is a random variable, $L_{\mathcal{D},f}(h_S)$ is also a random variable!

Mathematical Setup: Analysis Parameters

- S is sampled randomly from $\mathcal{D}$. So, when the ERM tries to minimize the error on S, its output h_S is also a random variable. Since h_S is a random variable, $L_{\mathcal{D},f}(h_S)$ is also a random variable!
- Example, if by chance, our sample S is biased and don't represent $\mathcal{D}$ well, we might get high error. We can't guarantee this won't happen. So we have to account for this.

Mathematical Setup: Analysis Parameters

- S is sampled randomly from $\mathcal{D}$. So, when the ERM tries to minimize the error on S, its output h_S is also a random variable. Since h_S is a random variable, $L_{\mathcal{D},f}(h_S)$ is also a random variable!
- Example, if by chance, our sample S is biased and don't represent $\mathcal{D}$ well, we might get high error. We can't guarantee this won't happen. So we have to account for this.

Definition (Confidence parameter $(1-\delta)$)

The probability of getting a non-representative sample $S \sim \mathcal{D}^m$ is denoted by δ, and $1-\delta$ is called the *confidence parameter*.

Mathematical Setup: Analysis Parameters

- Not all hypotheses $h \in \mathcal{H}$ is good and we can't guarantee perfect label prediction.

Wrap-Up (Review)

So many definitions and notations:

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.
3. **ERM hypothesis** $h_S \in \text{argmin}_{h \in \mathcal{H}}$ where $\mathcal{H}$ is our hypothesis class and we assume it has *finite* size.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.
3. **ERM hypothesis** $h_S \in \text{argmin}_{h \in \mathcal{H}}$ where $\mathcal{H}$ is our hypothesis class and we assume it has *finite* size.
4. **Confidence Parameter** $(1 - \delta)$: The probability of not getting a bad sample $S \sim \mathcal{D}^m$.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.
3. **ERM hypothesis** $h_S \in \text{argmin}_{h \in \mathcal{H}}$ where $\mathcal{H}$ is our hypothesis class and we assume it has *finite* size.
4. **Confidence Parameter** $(1 - \delta)$: The probability of not getting a bad sample $S \sim \mathcal{D}^m$.
5. **Accuracy Parameter** (ϵ): Our failure/success threshold. A learner is successful if $L_{\mathcal{D},f}(h_S) \leq \epsilon$.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.
3. **ERM hypothesis** $h_S \in \text{argmin}_{h \in \mathcal{H}}$ where $\mathcal{H}$ is our hypothesis class and we assume it has *finite* size.
4. **Confidence Parameter** $(1 - \delta)$: The probability of not getting a bad sample $S \sim \mathcal{D}^m$.
5. **Accuracy Parameter** (ϵ): Our failure/success threshold. A learner is successful if $L_{\mathcal{D},f}(h_S) \leq \epsilon$.
6. **Realizability Assumption**: $\exists h \in \mathcal{H}, L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$.

Wrap-Up (Review)

So many definitions and notations:

1. **Risks**: $L_s(h)$ is the *empirical risk*, and $L_{\mathcal{D}}(h)$ is the *true risk*.
2. Our **sample**: S with size m, sampled i.i.d from the distribution $\mathcal{D}$.
3. **ERM hypothesis** $h_S \in \text{argmin}_{h \in \mathcal{H}}$ where $\mathcal{H}$ is our hypothesis class and we assume it has *finite* size.
4. **Confidence Parameter** $(1 - \delta)$: The probability of not getting a bad sample $S \sim \mathcal{D}^m$.
5. **Accuracy Parameter** (ϵ): Our failure/success threshold. A learner is successful if $L_{\mathcal{D},f}(h_S) \leq \epsilon$.
6. **Realizability Assumption**: $\exists h \in \mathcal{H}, L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$.

Any questions on the notations/definitions?

Mathematical Analysis

- What do we want to show?

Mathematical Analysis

- What do we want to show?
- We want to show given our setup, the probability of ERM failing to learn a good hypothesis is bounded.

Mathematical Analysis

- What do we want to show?
- We want to show given our setup, the probability of ERM failing to learn a good hypothesis is bounded.
- i.e., we want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$

Mathematical Analysis

- What do we want to show?
- We want to show given our setup, the probability of ERM failing to learn a good hypothesis is bounded.
- i.e., we want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Another notation is: $\boxed{\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon\}]}$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$:
  $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples:
  $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- i.e., For every "misleading" sample $S \in M$, there exist a "bad" hypothesis $h \in \mathcal{H}_B$ such that looks "good" as far as h is concerned (since $L_S(h) = 0$.

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$:
  $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples:
  $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- i.e., For every "misleading" sample $S \in M$, there exist a "bad" hypothesis $h \in \mathcal{H}_B$ such that looks "good" as far as h is concerned (since $L_S(h) = 0$.
- Now recall the **realizability assumption**:
  $\exists h \in \mathcal{H} : L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$:
  $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples:
  $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- i.e., For every "misleading" sample $S \in M$, there exist a "bad" hypothesis $h \in \mathcal{H}_B$ such that looks "good" as far as h is concerned (since $L_S(h) = 0$.
- Now recall the **realizability assumption**:
  $\exists h \in \mathcal{H} : L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$
- By this assumption know that $\boxed{L_S(h_S) = 0}$

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- i.e., For every "misleading" sample $S \in M$, there exist a "bad" hypothesis $h \in \mathcal{H}_B$ such that looks "good" as far as h is concerned (since $L_S(h) = 0$.
- Now recall the **realizability assumption**: $\exists h \in \mathcal{H} : L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$
- By this assumption know that $\boxed{L_S(h_S) = 0}$
- Hence, the event $L_{\mathcal{D},f}(h_S) > \epsilon$ can happen if for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. i.e., the output of ERM will have 0 empirical loss.

Mathematical Analysis (2)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- Let's separate the set of "bad" hypotheses in $\mathcal{H}$:
  $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- Also, let's separate "misleading" or (non-representative) samples:
  $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- i.e., For every "misleading" sample $S \in M$, there exist a "bad" hypothesis $h \in \mathcal{H}_B$ such that looks "good" as far as h is concerned (since $L_S(h) = 0$.
- Now recall the **realizability assumption**:
  $\exists h \in \mathcal{H} : L_{\mathcal{D},f}(h) = 0, L_S(h) = 0$
- By this assumption know that $\boxed{L_S(h_S) = 0}$
- Hence, the event $L_{\mathcal{D},f}(h_S) > \epsilon$ can happen if for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. i.e., the output of ERM will have 0 empirical loss.
- Hence, $\boxed{\{S : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M}$

Mathematical Analysis (3)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- “bad” hypotheses: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- “misleading” samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- $\boxed{\{S : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M}$

Mathematical Analysis (3)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- "bad" hypotheses: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- "misleading" samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- $\boxed{\{S : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M}$
- Hence, $\boxed{M = \bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}}$

Mathematical Analysis (3)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- "bad" hypotheses: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- "misleading" samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- $\boxed{\{S : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M}$
- Hence, $\boxed{M = \bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}}$
- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}]}$

Mathematical Analysis (3)

- We want to upper bound: $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}]}$
- "bad" hypotheses: $\boxed{\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}}$
- "misleading" samples: $\boxed{M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}}$
- $\boxed{\{S : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M}$
- Hence, $\boxed{M = \displaystyle\bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}}$
- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h \in \mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- So the R.H.S is an upper bound for what we wanted. Can we make it simpler?

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h \in \mathcal{H}_B}\{S : L_S(h) = 0\}]}$

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h \in \mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- Union bound trick: $P(A \cup B) \leq P(A) + P(B)$

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- Union bound trick: $P(A \cup B) \leq P(A) + P(B)$
- $\boxed{D^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}\ ] \leq \sum_{h\in\mathcal{H}_B} D^m[S : L_S(h) = 0]}$

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- Union bound trick: $P(A \cup B) \leq P(A) + P(B)$
- $\boxed{D^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}\ ] \leq \displaystyle\sum_{h\in\mathcal{H}_B} D^m[S : L_S(h) = 0]}$
- The probability of each R.H.S summand?

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h \in \mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- Union bound trick: $P(A \cup B) \leq P(A) + P(B)$
- $$\boxed{D^m[\cup_{h \in \mathcal{H}_B}\{S : L_S(h) = 0\}\ ] \leq \sum_{h \in \mathcal{H}_B} D^m[S : L_S(h) = 0]}$$
- The probability of each R.H.S summand?

$$\begin{aligned} D^m[S : L_S(h) = 0] &= \mathcal{D}^m[\{S : \forall i : h(x_i) = f(x_i)\}] \\ &= \prod_{i=1}^{m} \mathcal{D}[\{x_i : h(x_i) = f(x_i)\}] \quad \text{i.i.d assumption} \\ &= \prod_{i=1}^{m} 1 - L_{\mathcal{D},f}(h) \leq \prod_{i=1}^{m} 1 - \epsilon \\ &= (1 - \epsilon)^m \leq e^{-\epsilon m} \end{aligned}$$

Mathematical Analysis (4)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq \mathcal{D}^m(M) = \mathcal{D}^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}]}$
- Union bound trick: $P(A \cup B) \leq P(A) + P(B)$
- $\boxed{D^m[\cup_{h\in\mathcal{H}_B}\{S : L_S(h) = 0\}\ ] \leq \sum_{h\in\mathcal{H}_B} D^m[S : L_S(h) = 0]}$
- The probability of each R.H.S summand?

$$
\begin{aligned}
D^m[S : L_S(h) = 0] &= \mathcal{D}^m[\{S : \forall i : h(x_i) = f(x_i)\}] \\
&= \prod_{i=1}^{m} \mathcal{D}[\{x_i : h(x_i) = f(x_i)\}] \ \text{ i.i.d assumption} \\
&= \prod_{i=1}^{m} 1 - L_{\mathcal{D},f}(h) \leq \prod_{i=1}^{m} 1 - \epsilon \\
&= (1-\epsilon)^m \leq e^{-\epsilon m}
\end{aligned}
$$

- We are done! $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}}$

Mathematical Analysis (∞)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}}$

Mathematical Analysis (∞)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}}$
- The above bound holds for **any** ϵ, δ. So, if we want to to make sure that (ϵ, δ), our learner succeeds, how many examples do we need?

Mathematical Analysis (∞)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}}$
- The above bound holds for **any** ϵ, δ. So, if we want to to make sure that (ϵ, δ), our learner succeeds, how many examples do we need?
- $\mathcal{H}e^{-\epsilon m} \leq \delta$, solve for m. We get: $\boxed{m \geq \frac{\ln(|H|/\delta)}{\epsilon}}$

Mathematical Analysis (∞)

- $\boxed{\mathcal{D}^m[\{S : L_{\mathcal{D},f}(h_S) > \epsilon\}] \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}}$
- The above bound holds for **any** ϵ, δ. So, if we want to to make sure that (ϵ, δ), our learner succeeds, how many examples do we need?
- $\mathcal{H}e^{-\epsilon m} \leq \delta$, solve for m. We get: $\boxed{m \geq \frac{\ln(|H|/\delta)}{\epsilon}}$

Corollary

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies: $m \geq \frac{\ln(|H|/\delta)}{\epsilon}$.
Then for **any labeling function** f, and for **any distribution** $\mathcal{D}$, for which the realizability assumption holds, with probability of at least $1 - \delta$, over the choices of an i.i.d sample S of size m, **every ERM hypothesis** h_S satisfies $L_{\mathcal{D},f}(h_S) \leq \epsilon$

Outline

Introduction

- Statistical Learning Framework
- Measures of Success

Empirical Risk Minimization (ERM)

- Overfitting
- Inductive Bias
- Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias

- Mathematical Setup
- ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

A Prelude to PAC Learning

Corollary

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies: $m \geq \frac{\ln(|H|/\delta)}{\epsilon}$.
Then for **any labeling function** f, and for **any distribution** $\mathcal{D}$, for which the realizability assumption holds, with probability of at least $1-\delta$, over the choices of an i.i.d sample S of size m, **every ERM hypothesis** h_S satisfies $L_{\mathcal{D},f}(h_S) \leq \epsilon$

Definition (PAC Learning)

A hypothesis class $\mathcal{H}$ is *PAC learnable* if there exists a function $m_{\mathcal{H}} : (0,1)^2 \mapsto \mathbb{N}$, and a leraning algorithm A with the the following property: For every, $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over X, and for every labeling function $f : X \mapsto \{0,1\}$, if the realizability assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d samples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that with probability at least $1-\delta$ over the choice of examples, $L_{(\mathcal{D},f)}(h) \leq \epsilon$

Outline

Introduction

- Statistical Learning Framework
- Measures of Success

Empirical Risk Minimization (ERM)

- Overfitting
- Inductive Bias
- Finite Hypothesis Class

Mathematical Analysis of ERM with Inductive Bias

- Mathematical Setup
- ERM with Finite Class Hypothesis

A Prelude to PAC Learning

Conclusion

Conclusion

1. Basics of Statistical Learning Framework

Conclusion

1. Basics of Statistical Learning Framework
2. Empirical Risk Minimization

Conclusion

1. Basics of Statistical Learning Framework
2. Empirical Risk Minimization
3. Mathematical Analysis of Learnability

Conclusion

1. Basics of Statistical Learning Framework
2. Empirical Risk Minimization
3. Mathematical Analysis of Learnability
4. Introduction to PAC Learning.

Future Lecture

1. Given a problem, how do we know that the realizability assumption holds? What if there is no h^* with $L_{\mathcal{D}}(h^*) = 0$?

Future Lecture

1. Given a problem, how do we know that the realizability assumption holds? What if there is no h^* with $L_{\mathcal{D}}(h^*) = 0$?
2. We'll see the Agnostic PAC Learning where we relax the realizability assumption

Future Lecture

1. Given a problem, how do we know that the realizability assumption holds? What if there is no h^* with $L_{\mathcal{D}}(h^*) = 0$?
2. We'll see the Agnostic PAC Learning where we relax the realizability assumption
3. What if our problem is not binary classification?

Future Lecture

1. Given a problem, how do we know that the realizability assumption holds? What if there is no h^* with $L_{\mathcal{D}}(h^*) = 0$?
2. We'll see the Agnostic PAC Learning where we relax the realizability assumption
3. What if our problem is not binary classification?
4. We'll extend the PAC Learning definition by introducing the generalized loss functions into our risk definitions.

Thank you!

seyediman.mirzadeh@wsu.edu