Bias vs Complexity $\rightarrow$ SRM $\rightarrow$ MDL

AW

Lecture Overview

Problem of Universal Learner

Estblished facts:

- Learning problem is defined by an unknown distribution D over $X \times Y$, where the goal of the learner is to find a predictor $h : X \to Y$, whose risk, $L_{D \sim X \times Y}(h)$ is "small enough".
- Training dataset S can mislead the learner, and result in choosing h such that $L_S(h)$ is small but $L_D(h)$ is not (aka *overfitting*)
- To beat overfitting we decided to limit class $\mathscr{H}$ of hypothesis. Choice of class reflects *prior knowledge* about the problem.
- Prior knowledge = a belief that one of the members of the class $\mathscr{H}$ is a low-error model for the problem

Question: is prior knowledge necessary to choose an h with small enough $L_{D \sim X \times Y}(h)$?

In other words, is it possible to construct a universal learner A such that for any given distribution D, if A receives enough i.i.d. examples from D (i.e. traing set of some standard size m), there is a high chance it outputs a predictor h that has a low risk?

No Universal Learner Exists

Theorem (No Free Lunch)

Let A be any learning algorithm for binary classification problem over a domain X. Let m be any number smaller than $|X|/2$, representing a training set size. Then, there exists a distribution D over $X \times \{0,1\}$ such that:

1. *There exists a function $f : X \to \{0,1\}$ with $L_{D\sim X\times\{0,1\}}(f) = 0$*
2. *With probability of at least $1/7$ over the choice of $S \sim D^m$ we have that $L_{D\sim X\times\{0,1\}}(A(S)) \geq 1/8$.*

If you are interested in proof see SSBD section 5.1.

What does No Free-Lunch Thorem Mean?

- Let X be any domain and the hypothesis class $\mathscr{H}$ be the class of all the functions $\{f|f : X \to \{0,1\}\}$. This class represents no prior knowledge
- Let X be so big that taking sample $M \geq |X|/2$ is unrealisric
- No-Free-Lunch theorem says that any learning algorithm that chooses its output from $\mathscr{H}$ (and in particular the ERM predictor), fails on some distribution D over $X \times \{0,1\}$ (i.e. fails on some instance of learning problem). So the class of all functions is not PAC learnable.

Corollary

Let X be an infinite domain set and let $\mathscr{H}$ be the set of all functions $\{f|f : X \to \{0,1\}\}$. Then, $\mathscr{H}$ is not PAC learnable.

Avoiding Learning Failures

- How to avoid learning failures?
  - Use our prior knowledge about a specific instance of learning problem:
    - Prohibit unrealistic distributions that causes failure .

    Such prior knowledge can be expressed by restricting our hypothesis class
- how to choose a good hypothesis class?
  - On the one hand, this class must contain a hypothesis with no error on our instance (then we'll find it by PAC learning) or at least contains hypothesis with small error (then we disover a good hypothesis using agnostic PAC learning).
  - On the other hand, we cannot simply choose the richest class (the class of all functions over the given domain).
- We need to find a tradeoff!

Lecture Overview

Types of Errors

- **The Approximation Error ϵ_{app}.** It is the minimum risk achievable by a predictor in the hypothesis class i.e. $\epsilon_{app} = \min_{h \in \mathscr{H}} L_D(h)$. It measures how much risk is due to restricting hypothesis class. In other words it measures how much inductive bias is built into learning.
  - It does not depend on the sample size
  - It is determined by the hypothesis class chosen.
  - Enlarging the hypothesis class can decrease the approximation error.

Types of Errors

- **The Approximation Error ϵ_{app}.** It is the minimum risk achievable by a predictor in the hypothesis class i.e. $\epsilon_{app} = \min_{h \in \mathscr{H}} L_D(h)$. It measures how much risk is due to restricting hypothesis class. In other words it measures how much inductive bias is built into learning.
- **The Estimation Error ϵ_{est}.** It is the difference between the approximation error and the error achieved by the ERM predictor, i.e. $\epsilon_{est}(h) = L_D(h) - \epsilon_{app}$.
  - It occurs because the training error is only an estimate of the true risk, and so the ERM predictor minimizing the empirical risk is only an estimate of the predictor minimizing the true risk.
  - How small it is depends on the training set size and on the size of the hypothesis class.
- The size of the hypothesis class is one of the measures of its complexity (we'll see more such measures later).

Types of Errors

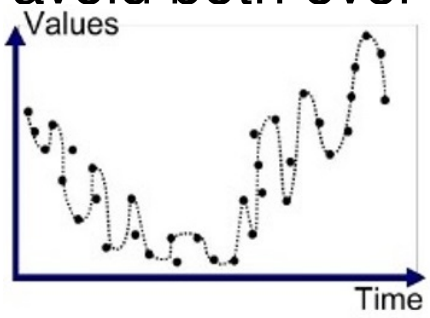
- **The Approximation Error ϵ_{app}.** It is the minimum risk achievable by a predictor in the hypothesis class i.e. $\epsilon_{app} = \min_{h \in \mathscr{H}} L_D(h)$. It measures how much risk is due to restricting hypothesis class. In other words it measures how much inductive bias is built into learning.
- **The Estimation Error ϵ_{est}.** It is the difference between the approximation error and the error achieved by the ERM predictor, i.e. $\epsilon_{est}(h) = L_D(h) - \epsilon_{app}$.
- Total risk of ERM learned hypothesis is $L_D(h_S) = \epsilon_{app} + \epsilon_{est}(h_s)$ where h_S is ERM learned hypothesis

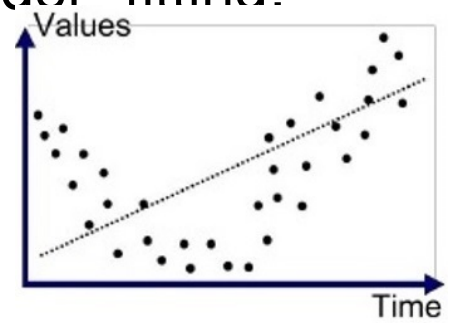
Types of Errors

- **The Approximation Error ϵ_{app}.** It is the minimum risk achievable by a predictor in the hypothesis class i.e. $\epsilon_{app} = \min_{h \in \mathscr{H}} L_D(h)$. It measures how much risk is due to restricting hypothesis class. In other words it measures how much inductive bias is built into learning.
- **The Estimation Error ϵ_{est}.** It is the difference between the approximation error and the error achieved by the ERM predictor, i.e. $\epsilon_{est}(h) = L_D(h) - \epsilon_{app}$.
- Total risk of ERM learned hypothesis is $L_D(h_S) = \epsilon_{app} + \epsilon_{est}(h_s)$ where h_S is ERM learned hypothesis
- The goal is to minimize the total risk, so there is a tradeoff called the bias-complexity tradeoff:
  - increasing size of $\mathscr{H}$ decreases ϵ_{app}, but may increase ϵ_{est} as large $\mathscr{H}$ may lead to overfitting.
  - limiting the size of $\mathscr{H}$ reduces the ϵ_{est}, but may lead to a model that is too simple increasesing ϵ_{app} (aka *underfitting*).

Types of Errors

- **The Approximation Error ϵ_{app}.** It is the minimum risk achievable by a predictor in the hypothesis class i.e. $\epsilon_{app} = \min_{h \in \mathscr{H}} L_D(h)$. It measures how much risk is due to restricting hypothesis class. In other words it measures how much inductive bias is built into learning.
- **The Estimation Error ϵ_{est}.** It is the difference between the approximation error and the error achieved by the ERM predictor, i.e. $\epsilon_{est}(h) = L_D(h) - \epsilon_{app}$.
- Total risk of ERM learned hypothesis is $L_D(h_S) = \epsilon_{app} + \epsilon_{est}(h_s)$ where h_S is ERM learned hypothesis
- The goal is to minimize the total risk, i.e. to find bias-complexity tradeoff to avoid both over- and under- fitting:





Lecture Overview

Agnostic PAC Unlearnable Classes

So far

- If class $\mathscr{H}$ is agnostic-PAC learnable then we can find the 'best' hypothesis in class that has error $\epsilon = \epsilon_{app}$ if the class $\mathscr{H}$ has uniform convergence property
- The sample size necessary to learn this hypothesis depends only on the class size, accuracy and confidence parameters

Often there are classes $\mathscr{G}$, $\mathscr{H}$ such that $\mathscr{G} \subset \mathscr{H}$ and $\mathscr{G}$ is agnostic PAC-learnable, but while $\mathscr{H}$ is not. Yet we have experimental evidence that some predictor $h \in \mathscr{H}$ for sample sizes $m_{\mathscr{G}}(\epsilon, \delta)$ (obtained for some pairs (ϵ, δ) for class $\mathscr{G}$) gives better generaliization error than ϵ. How can we construct predictors from $\mathscr{H}$ analytically?

What if we relax the requirement that sample size depends only on the accuracy, class size and confidence parameters? i.e. what if we allow the sample size to depend on the $h \in \mathscr{H}$ with which our learning algorithm A is competing?

Non-Uniform Learnability Model

Definition

A hypothesis class $\mathscr{H}$ is said to be non-uniformly learnable if there exist a learning algorithm, A, and a function $m^0_{\mathscr{H}} : \{0,1\}^2 \times \mathscr{H} \to N$ such that, for every $\epsilon, \delta \in \{0,1\}$ and for every $h \in \mathscr{H}$, if $m \geq m^0_{\mathscr{H}}(\epsilon, \delta, h)$ then for every distribution D, with probability of at least $1-\delta$ over the choice of $S \sim D^m$, it holds that $L_D(A(S)) \leq L_D(h) + \epsilon$.

Notice the difference: in agnostic PAC we have if $m \geq m_{\mathscr{H}}(\epsilon, \delta)$, then with probability of at least $1-\delta$ over the choice of $S \sim D^m$ it holds that

$$LD(A(S)) \leq \min_{h \in \mathscr{H}} L_D(h) + \epsilon$$

i.e. in agnostic PAC the comparison is with the best of $\mathscr{H}$ which entails that A is almost as good as any predictor in $\mathscr{H}$ up to ϵ on the same sample. In nonuniform case we can be almost as good as any $h \in \mathscr{H}$ up to ϵ but for some predictors we may need huge samples!

Union Classes $\mathscr{H}$

Consider known classes of predictors for binary classification problem:

- Class $\mathscr{H}_T$ of decision tree predictors is in fact union of classes of binary trees, ternary trees, quaternary trees, ... In other words it is union of classes of decision trees with maximum branching degree b varying over $b \in \mathbb{N}$, $b > 1$.
- Class $\mathscr{H}_P$ of polynomial classifiers that assign to an instance x its class $\mathrm{sgn}(p(x))$ for a given polynomial $p : \mathscr{R} \to \mathscr{R}$ where p is polynomial of degree n is in fact union $\bigcup_{i \in \mathscr{N}} P_i(x)$ where P_i is a vector space of all polynomials of degree i
- Class $\mathscr{H}_{NN}$ of nearest neighbor classifiers that classifies $\vec{x}$ by its nearest-neighbor class using distance (norm) $L^p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$. This class is a union of classes with $p = 1, 2, \ldots$.

What is in common for all these classes? In all these case hypothesis class $\mathscr{H}$ can be written as $\mathscr{H} = \bigcup_{i \in \mathbb{N}} \mathscr{H}_i$ for well defined hypothesis classes $\mathscr{H}_i$

Incorporating Prior Knowledge

- So far by specifying a hypothesis class $\mathscr{H}$ that includes a good predictor.
- What if all we believe in is that good predictor is
  - in one of the countable number of classes, i.e. in a class that is 'union' of hypothesis's classes,
  - more likely to be in some classes than in others?

  We could assign weights to component classes to formalize this knowledge

New Model:

- Hypothesis class $\mathscr{H}$ is such that $\mathscr{H} = \bigcup_{i \in \mathbb{N}} \mathscr{H}_i$ where each $\mathscr{H}_i$ is a class with uniform convergence property and sample complexity $m_{H_i}^{UC}(\epsilon, \delta)$
- Weight function $w : \mathbb{N} \to [0, 1]$ is such that $\sum_{n=1}^{\infty} w(n) \leq 1$ reflects preferences of hypothesis classes ($w(i)$ preference of class $\mathscr{H}_i$)

What are the Preferences?

What learning can we seek in this model? Which classes should we prefer?

- Before we only considered classes minimizing approximation error.
- But error consist of ϵ_{app} and ϵ_{est} - we can try trading some of ϵ_{app} to obtain better ϵ_{est}. So let us now prefer classes with low ϵ_{est}
- These preferences are assigned in context of non-uniform learning

Known facts in the model:

- Each $\mathscr{H}_n$ has uniform convergence property with sample complexity $m^{UC}_{\mathscr{H}_n}(\epsilon, \delta)$
- Fix sample size m and confidence level δ. Then minimum error that can be obtained with this sample size and confidence for class $\mathscr{H}_n$ is $\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) \,|\, m^{UC}_{\mathscr{H}_n}(\epsilon, \delta) < m\}$
- So for every m and δ, with probability of at least $1 - \delta$ over the choice of $S \sim D^m$ we have that for every $h \in \mathscr{H}_n$ holds $|L_D(h) - L_S(h)| \leq \epsilon_n(m, \delta)$

Base for Structural Risk Minimization

Theorem (Risk Bound for Union of Classes w/Uniform Convergence)

Let $w : N \to [0,1]$ *be a function such that* $\sum_{n=1}^{\infty} w(n) \leq 1$. *Let* $\mathscr{H}$ *be a hypothesis class that can be written as* $H = \bigcup_{n \in N} \mathscr{H}_n$, *where for each* n, H_n *satisfies the uniform convergence property with a sample complexity function* $m^{UC}_{\mathscr{H}_n}$. *Then, for every* $\delta \in (0,1)$ *and distribution* D, *with probability of at least* $1-\delta$ *over the choice of* $S \sim D^m$, *the following bound holds (simultaneously) for every* $n \in N$ *and* $h \in H_n$:

$$|L_D(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$

where $\epsilon_n(m,x) = \min\{\epsilon \in (0,1) \,|\, m^{UC}_{\mathscr{H}_n}(\epsilon, x) < m\}$. *Therefore, for every* $\delta \in (0,1)$ *and distribution* D, *with probability of at least* $1-\delta$ *for every* $h \in \mathscr{H}$ *it holds that*

$$L_D(h) \leq L_S(h) + \epsilon \text{ where } \epsilon \leq \min_{n:h \in \mathscr{H}_n} \epsilon_n(m, w(n) \cdot \delta)$$

Structural Risk Minimization Paradigm

- Denote $n(h) = \min\{n : h \in \mathscr{H}_n\}$
- The Risk Bound for Union of Classes w/Uniform Convergence implies that $L_D(h) \leq L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)$

Structural Risk Minimization (SRM) Paradigm

Suppose our prior knowledge includes the following:

- a hypothesis that has low error on all data must belong to a hypothesis class $\mathscr{H}$ that can be written as $\mathscr{H} = \bigcup_{n \in N} \mathscr{H}_n$, where for each n, $\mathscr{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathscr{H}_n}$
- this hypothesis is more likely to belong to some classes $\mathscr{H}_i$ that the others. The 'likelihood' can be expressed by a weight preference function $w : N \to [0, 1]$ where $\sum_{n=1}^{\infty} w(n) \leq 1$

In this setting we must find

$$h = \arg\min_{h \in \mathscr{H}} [L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)]$$

Lecture Overview

Occam's Razor

- SRM trades some of our bias toward low empirical risk for a bias toward classes for which $\epsilon_{n(h)}(m, w(n(h)) \cdot \delta)$ is smaller for the sake of a smaller estimation error.
- The estimation error is as good as our weight function. How do we choose it?

One way to choose weights is to make them a function of the *description length* of a hypothesis. Why? because of

> *Occam's Razor:* A short explanation (that is, a hypothesis that has a short length) tends to be more valid than a long explanation.

How to implement this idea?

Description languages

- Fix some finite alphabet Σ e.g $\Sigma = \{0, 1\}$.
- A string is a finite sequence from Σ, e.g. $l = 01110 \in \Sigma^5$ is a string of length $|l| = 5$. The set $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$.is a set of all finite strings in alphabet Σ
- A description language for $\mathscr{H}$ is a map $r : \mathscr{H} \to \Sigma^*$, that maps each member h of $\mathscr{H}$ to a string $d(h)$ that is the description (or *representation*) of h. We identify h with its representation and write $|h|$ for $|d(h)|$

Descriptions must be uniquely decodable and self-delimiting, so description languages must be prefix-free: for every distinct h and h', $d(h)$ is not a prefix of $d(h')$, i.e. string $d(h)$ is not exactly the first $|h|$ symbols of any longer string $d(h')$

Lemma (Kraft Inequality)

If $S \subseteq \Sigma^*$ is a prefix-free set of strings, then $\sum_{s \in S} 2^{-|s|} \leq 1$

Singleton Classes and Their Union

Consider a class of all decision trees of branching degrees up to b over n features each of which having finite domain of size at most m.

- We can compute a number of trees of depth d as a function of integers b, n and m
- Of course depth d is an integer, so the number of such trees countable
- Therefore, we can enumerate all trees

Now let $\mathscr{H}_i$ be a hypothesis class that consists of an individual decision tree h that has number i in the enumeration we designed. Then the class of all finite decision trees $\mathscr{H}$ is a countable union of singleton classes, i.e. $\mathscr{H} = \bigcup_{i=1}^{\infty} \mathscr{H}_i$

Description Length and Weights

Let $\mathscr{H} = \bigcup_{i=1}^{\infty} \mathscr{H}_i$ where each $\mathscr{H}_i$ is a singleton class containing one hypothesis h.

- Shown before: a finite hypothesis class $\mathscr{H}$ of binary classification hypothesis that has finite domain enjoys the uniform convergence property with sample complexity $m_{\mathscr{H}}^{UC}(\epsilon, \delta) \leq \lceil \frac{\log(2|H|/\delta)}{2\epsilon^2} \rceil$
- For singleton classes this sample complexity becomes $m_{\{h\}}^{UC}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{2\epsilon^2} \rceil$
- For a given m and δ, algebraic manipulation gives $\epsilon_n(m, \delta) \leq \sqrt{\frac{\log(2/\delta)}{2m}}$ where n is the number of hypothesis h in a given enumeration
- To use SRM we need $\epsilon_{n(h)}(m, w(n(h)) \cdot \delta) \leq \sqrt{\frac{\log(2/(w(n(h)) \cdot \delta))}{2m}}$. Substituting it into SRM rule $h = \arg\min_{h \in \mathscr{H}} [L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)]$ we get new rule $h = \arg\min_{h \in \mathscr{H}} \left[L_S(h) + \sqrt{\frac{\log(2/\delta) - \log w(n(h))}{2m}} \right]$

Minimum Description Length Paradigm

- Suppose $\mathscr{H} = \bigcup_{i=1}^{\infty} \mathscr{H}_i$ where each $\mathscr{H}_i$ is a singleton class containing hypothesis h and $L : \mathscr{H} \to \{0,1\}^*$ a prefix-free description language
- By Kraft inequality $\sum_{h \in L} 2^{-|h|} \leq 1$, so we can assign weight of

$$w(h) = w(n(h)) \stackrel{def}{=} \frac{1}{2^{|h|}}$$

- Then for every sample size, m, every confidence parameter, $\delta > 0$, and every probability distribution, D, with probability greater than $1 - \delta$ over the choice of $S \sim D^m$ we have that,

$$L_D(h) \leq L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}}$$

Minimum Description Length Paradigm

Given a countable hypothesis class $\mathscr{H}$ described by a prefix-free language L, a training set $S \sim D^m$ and a confidence parameter $0 < \delta < 1$ find

$$h \in \arg\min_{h \in \mathscr{H}} \left[L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}} \right]$$