Machine Learning I

Lecture 3: Agnostic PAC Learning

Nathaniel Bade

Northeastern University Department of Mathematics

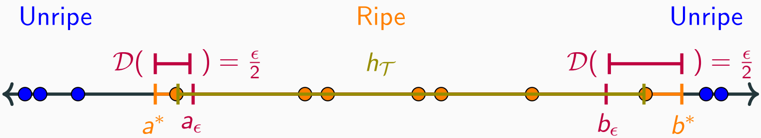
Table of contents

1. PAC Learning
2. Agnostic PAC Learning
3. Generalized Loss
4. Proving Bounds in APAC
5. Uniform Convergence
6. Hoeffding's Inequality
7. Deriving Hoeffding's Inequality
8. Finite Hypothesis Classes are APAC Learnable

PAC Learning

PAC Learnability

Mango Space


Unripe
Ripe
Unripe
$\mathcal{D}(\ \longmapsto\ ) = \frac{\epsilon}{2}$
$h_{\mathcal{T}}$
$\mathcal{D}(\ \longmapsto\ ) = \frac{\epsilon}{2}$
a^*
a_ϵ
b_ϵ
b^*

$$\mathcal{X} = \mathbb{R}$$

Last time, we saw that for a sufficiently large training set $\mathcal{T}$, the **ERM** rule on the hypothesis class of interval classifiers, $\mathcal{H}$, outputs a classifier $h_{\mathcal{T}}$ that is **probability approximately correct** or **PAC**.

That is, if the inputs have distribution $\mathcal{D}$ and f is the true labeling, then the error $L_{(\mathcal{D},f)}(h_{\mathcal{T}}) < \epsilon$ with probability $1 - \delta$, provided

$$N > 2\frac{\log(2/\delta)}{\epsilon} \,.$$

PAC Learnability

PAC Learnability: A hyptothesis class $\mathcal{H}$ is PAC learnable if there exists a function $N_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties:

For every $\epsilon, \delta \in (0,1)$, and for every distribution $\mathcal{D}$ on $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \mathcal{Y}$, if realizability holds, then with probability at least $1 - \delta$, $A(\mathcal{T}) = h$ has total error $L_{(\mathcal{D},f)}(h) < \epsilon$ provided $|\mathcal{T}| > N_{\mathcal{H}}(\epsilon, \delta)$.

PAC Learnability

PAC learnability has two parameters:

The **accuracy parameter** ϵ bounds the total error of the output classifier.

The **confidence parameter** δ indicates how likely the classifier is to make the accuracy requirement.

The randomness in these parameters reflects the fact that there is always some chance of choosing a highly non-representative data sample from $\mathcal{D}$.

Sample Complexity

The function $N_{\mathcal{H}} : (0,1)^2 \to \mathbb{R}$ is called the **sample complexity** of learning $\mathcal{H}$. It governs the number of examples needed to learn $\mathcal{H}$.

Note, that if $\mathcal{H}$ is PAC learnable there are actually many functions $N_{\mathcal{H}}$ that satisfy the requirements of PAC learnability. In general, we will use sample complexity to refer to the minimum such function in the class of all $N_{\mathcal{H}}$.

Sample Complexity for Interval Indicators

We can restate our result about the interval classifier in the PAC framework in terms of the sample complexity:

Our computation in Lecture 2 showed that the hypothesis class of integer indicator functions

$$\mathcal{H} = \{1_I \,:\, I = [a,b],\, a,b \in \mathbb{R}\}$$

is PAC learnable, with sample complexity

$$N_{\mathcal{H}}(\epsilon,\delta) \leq \mathrm{ceil}\left(2\frac{\log(2/\delta)}{\epsilon}\right).$$

Agnostic PAC Learning

Realizability Revisited

In many practical problems, the assumption of realizability does not hold. Recall that formally, realizability is the assumption that there exists an $h^* \in \mathcal{H}$ such that

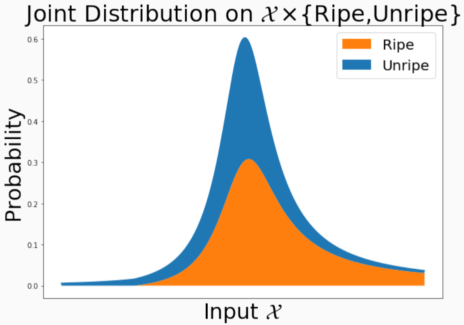
$$\mathbb{P}_{X \sim \mathcal{D}}(\, h^*(X) = f(X)\,) = 1\,.$$

Realizability can fail in two ways:

- The labels could be generated by a function f “far away” from $\mathcal{H}$.
- The labels may not be fully determined by the features.

We will now drop the notion of a labeling function f and instead assume the data labels are generated by a distribution.

Joint Distribution

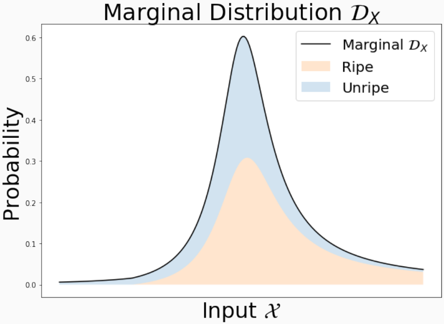


Let $\mathcal{D}$ be a joint distribution over the space of points and labels $\mathcal{X} \times \mathcal{Y}$. As we saw before, $\mathcal{D}$ can thought of as being composed of two parts:

The **marginal** distribution $\mathcal{D}_X$, which gives the likelihood of a point of $\mathcal{X}$ being sampled.

The **conditional** distribution $\mathcal{D}(Y|X = x)$, which give the likelyhood of a label given that x was sampled.

Joint Distribution

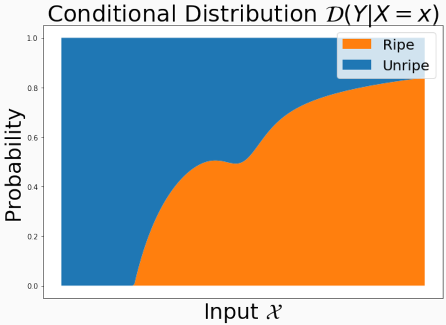


Let $\mathcal{D}$ be a joint distribution over the space of points and labels $\mathcal{X} \times \mathcal{Y}$. As we saw before, $\mathcal{D}$ can thought of as being composed of two parts:

The **marginal** distribution $\mathcal{D}_X$, which gives the likelihood of a point of $\mathcal{X}$ being sampled.

The **conditional** distribution $\mathcal{D}(Y|X = x)$, which give the likelyhood of a label given that x was sampled.

Joint Distribution



Let $\mathcal{D}$ be a joint distribution over the space of points and labels $\mathcal{X} \times \mathcal{Y}$. As we saw before, $\mathcal{D}$ can thought of as being composed of two parts:

The **marginal** distribution $\mathcal{D}_X$, which gives the likelihood of a point of $\mathcal{X}$ being sampled.

The **conditional** distribution $\mathcal{D}(Y|X = x)$, which give the likelyhood of a label given that x was sampled.

Error

Give a joint distribution on $\mathcal{X} \times \mathcal{Y}$, we redefine the **true error** (**risk**) of a predictor h to be

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[h(X) \neq Y\right],$$

or

$$L_{\mathcal{D}}(h) := \mathcal{D}\Big(\{(X, Y) \,:\, h(X) \neq Y\} \Big).$$

The **empirical risk** remains the same,

$$L_{\mathcal{T}}(h) := \frac{|\{i \,:\, h(x_i) \neq y_i\}|}{N}.$$

The goal now will be to find some classifier $h : \mathcal{X} \to \mathcal{Y}$ that (probably approximately) minimizes the **true risk** $L_{\mathcal{D}}(h)$.

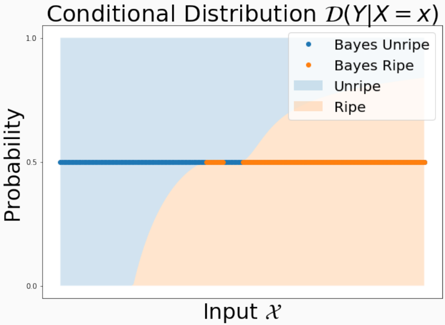
Bayes Optimal Predictor

Given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the "best" (measured by $L_{\mathcal{D}}(h)$ minimizing) possible label predicting function from $\mathcal{X}$ to $\{0,1\}$ will be

$$f_{\mathcal{D}} = \begin{cases} 1 & \text{if } \mathbb{P}[Y = 1 | X = x] \geq 1/2\,, \\ 0 & \text{otherwise.} \end{cases}$$

This function is known as the **Bayes optimal predictor**. It's a good exercise to show that for any other classifier g, $L_{\mathcal{D}}(f^*) \leq L_{\mathcal{D}}(g)$.

Bayes Optimal Predictor



For example, given the conditional probability for our simulated mango distribution, the Bayes optimal predictor simply selects the most likely labeling for each data point.

Of course, the Bayes optimal predictor “knows” the underlying distrbiution, while the PAC learner does not.

Agnostic PAC Learnability

Agnostic PAC Learnability: A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $N_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties:

For every $\epsilon, \delta \in (0,1)$, and for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, when running A on a training set of $N \geq N_{\mathcal{H}}(\delta, \epsilon)$ i.i.d. samples, then, with probability at least $1 - \delta$, the total error of $A(\mathcal{T}) = h$ is bounded by

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

(Of course, if realizability holds $\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') = 0$.)

Agnostic PAC Learnability

Agnostic PAC Learnability: If the realizability assumption does not hold, no learner can guarantee an arbitrarily small error. Instead, we declare the learner a success if the error is not must larger than the best achievable error within the hypothesis class $\mathcal{H}$.

Generalized Loss

Generalized Loss

There is one final piece we need to examine before we will have a full working definition of learnability: the loss function.

As discussed before, the loss function

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[h(X) \neq Y\right]$$

only makes sense for categorical labeling, since if Y is a continuous variable $\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[h(X) \neq Y\right] = 0$.

We will finish the PAC framework by generalizing the notion of loss.

Generalized Loss

Given a hypothesis class $\mathcal{H}$ and any domain $\mathcal{Z}$, a **loss function** will be any function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$. For labeling and regression problems, $\mathcal{Z}$ will be $\mathcal{X} \times \mathcal{Y}$.

We now define the **risk function** to be the expected loss of a classifier $h \in \mathcal{H}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{Z}$:

True risk : $\quad L_{\mathcal{D}}(h) := E_{Z \sim \mathcal{D}} \left[\ell(h, Z) \right].$

Similarly, the **empirical risk** is the expected loss over a given sample $\mathcal{T} = \{z_1, \ldots, z_N\}$

Empirical risk : $\quad L_{\mathcal{T}}(h) := \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \ell(h, z_i).$

Examples of Loss Functions

0-1 loss: For binary or multiclass classification, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y}$ is finite. We define

$$\ell_{0-1}(h,(X,Y)) := \begin{cases} 0 & \text{if } h(X) = Y\,, \\ 1 & \text{if } h(X) \neq Y\,. \end{cases}$$

We have been using **0-1 loss** in our mango example as well as in the definitions in PAC and agnostic PAC.

Examples of Loss Functions

Square Loss: For regression, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ but $\mathcal{Y}$ is continuous. The loss function

$$\ell_{\mathsf{sq}}(h, (X, Y)) := (h(X) - Y)^2$$

is the RSS function discussed in Lecture 1. We saw that minimizing over the class of all functions according to ℓ_{sq} leads to an optimal regression function

$$f(x) = E(Y|X = x)$$

Examples of Loss Functions

Absolute Loss: For regression, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ but $\mathcal{Y}$ is continuous. It was mentioned in Lecture 1 that minimizing with respect to the absolute loss

$$\ell_{\mathsf{abs}}(h, (X, Y)) := |h(X) - Y|$$

over the class of all functions leads to an optimal regression function

$$f(x) = \mathrm{Med}(Y|X = x)\,.$$

It should be noted that this median function is actually often preferable tot he expectation value (it's more stable in certain ways). However, the absolute value makes minimizing hard.

Agnostic PAC Learnability with Generalized Loss

Agnostic PAC With Loss: A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with resepct to a set $\mathcal{Z}$ and loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ if there exists a function $N_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties:

For every $\epsilon, \delta \in (0,1)$, and for every distribution $\mathcal{D}$ on $\mathcal{Z}$, when running A on a training set of $N \geq N_{\mathcal{H}}(\delta, \epsilon)$ i.i.d. samples, then, with probability at least $1 - \delta$, the total error of $A(\mathcal{T}) = h$ is bounded by

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

where $L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}\big[\ell(h, Z)\big]$.

We will this criteria **APAC learnable**.

Proving Bounds in APAC

Proving Bounds in APAC

Goal: Prove that every finite hypothesis class is APAC learnable.

To work in the APAC framework we need some tools:

Uniform convergence - gives criteria for APAC learnability.

Hoeffding's Inequality - computes sample complexity from the generalized loss.

Uniform Convergence

ϵ-Representative Training Sets

If a hypothesis class is APAC learnable, running $ERM(h)$ on a training set $\mathcal{T}$ will minimize the true error.

It suffices to ensure that the empirical risk $L_{\mathcal{T}}(h)$ is a good indicator of the true risk $L_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$.

A training set $\mathcal{T}$ is called ϵ-**representative** (w.r.t. $\mathcal{Z}$, $\mathcal{H}$, ℓ and $\mathcal{D}$) if

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

that is, if empirical risk on $\mathcal{T}$ is a good indicator of the true risk on $\mathcal{Z}$.

ϵ-Representative Training Sets

Lemma: If a training set is $\epsilon/2$-representative, then any output of $ERM_{\mathcal{H}}(\mathcal{T})$ (that is any $h_{\mathcal{T}} \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{T}}(h)$) satisfies

$$L_{\mathcal{D}}(h_{\mathcal{T}}) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\,.$$

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

$$L_{\mathcal{D}}(h_{\mathcal{T}}) \underset{\epsilon-rep}{\leq} L_{\mathcal{T}}(h_{\mathcal{T}}) + \frac{\epsilon}{2} \underset{ERM}{\leq} L_{\mathcal{T}}(h) + \frac{\epsilon}{2} \underset{\epsilon-rep}{\leq} L_{\mathcal{D}}(h) + \epsilon\,.$$

□

Uniform Convergence

Uniform Convergence: A hypothesis class has the uniform convergence property (w.r.t. Z and ℓ) if there exists a function $N_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$, if $\mathcal{T}$ is a training set of size $N \geq N_{\mathcal{H}}^{UC}(\epsilon, \delta)$ then, with a probability of at least $1 - \delta$ it is ϵ-representative.

Corollary: If a class has the uniform convergence property, then it is APAC learnable by ERM with sample complexity $N_{\mathcal{H}}^{UC}$.

Hoeffding’s Inequality

Hoeffding's Inequality

We now have a criteria for APAC learnability in terms of the sample sets $\mathcal{T}$, but we need some tools from probability to estimate expectation values.

Hoeffding's Inequality: Let $\theta_1, \ldots, \theta_N$ be i.i.d. random variables with $\mathbb{P}[a \leq \theta \leq b] = 1$. Define $\theta = \sum \theta_i$. Then for any $\epsilon > 0$,

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

Hoeffding's Inequality

Hoeffding's Inequality: Let $\theta_1, \ldots, \theta_N$ be i.i.d. random variables with $\mathbb{P}[a \le \theta \le b] = 1$. Define $\theta = \sum \theta_i$. Then for any $\epsilon > 0$,

$$\delta = \mathbb{P}\left[\left| \frac{1}{N} \sum_{i=1}^{N} (\theta_i - E[\theta_i]\,) \right| > \epsilon \right] \le 2 \exp\left(-2N\epsilon^2/(b-a)^2 \right) \,.$$

Relation to APAC: Setting $\theta_i = \ell(h, z_i)$ immediately gives

$$L_{\mathcal{T}}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell(h, z_i) = \frac{1}{N} \sum_{i=1}^{N} \theta_i \,,$$

and $L_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^{N} E[\theta_i] = E[\ell(h,z)]$. So

$$\delta = \mathcal{D}\Big(\left\{ \mathcal{T} : |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon \right\} \Big) \,.$$

Deriving Hoeffding's Inequality

Markov's Inequality

Hoeffdings inequality is a powerful, but complicated bound. We will start with **Markov** and **Chebyshev**'s inequalities as estimates for probabilties of the form

$$\mathbb{P}\big[f(\theta) > \epsilon\big]\,.$$

We will then define the exponential **moment generating function** to extend these results to sums of random variables

$$\mathbb{P}\big[\sum_i f(\theta_i) > \epsilon\big]\,.$$

We finally define Hoeffding's Lemma to bound exponential probabilities.

Markov's Inequality

Markov's inequality: Let $\theta \geq 0$ be a random variable on $\mathbb{R}_+$. Then, for all $\epsilon \geq 0$,

$$\mathbb{P}(\theta \geq \epsilon) \leq \frac{E[\theta]}{\epsilon}\,.$$

Proof:

$$\begin{aligned}
E[\theta] &= \int_{\mathbb{R}_+} \theta p(\theta) d\theta \\
&\geq \int_{\epsilon}^{\infty} \theta p(\theta) d\theta && \text{since } p(\theta) \geq 0 \\
&\geq \int_{\epsilon}^{\infty} \epsilon p(\theta) d\theta && \text{since } \theta > \epsilon \\
&= \epsilon \mathbb{P}(\theta > \epsilon)\,. && \square
\end{aligned}$$

Many probability estimates are just extensions are Markov's inequality.

Chebyshev's Inequality

Chebyshev's inequality: Let $\theta \geq 0$ be a random variable on $\mathbb{R}$ with $\mathsf{Var}(\theta) < \infty$. Then

$$\mathbb{P}(\,|\theta - E[\theta]|\, \geq \epsilon) \leq \frac{\mathsf{Var}[\theta]}{\epsilon^2}\,.$$

Proof: This is a direct appliction of Markov's inequality:

$$\begin{aligned}\mathbb{P}(\,|\theta - E[\theta]|\, \geq \epsilon) &= \mathbb{P}\Big((\theta - E[\theta])^2 \geq \epsilon^2\Big)\\ &\leq \frac{E\big[\,(\theta - E[\theta])^2\big]}{\epsilon^2}\,.\ \square\end{aligned}$$

As a side note, one nice consequence of Chebyshev's inequality is that the averages of random variables converge to their mean. This is one face of the **law of large numbers**.

Exercise: For θ_i i.i.d, with $E(\theta_i) = 0$, show that $\mathsf{Var}(\theta) = \frac{1}{N}\mathsf{Var}(\theta_i)$.

Moment generating functions

For a real random variable θ, the **moment generating function** is the exponential

$$M_\theta(\lambda) := E[e^{\lambda\theta}]\,.$$

Taking the expectation the exponential we can read the moments off of $M_\theta(\lambda)$:

$$M_\theta(\lambda) = 1 + \lambda E(X) + \frac{\lambda^2}{2} E(X^2) + \ldots$$

Moment generating functions

Moment generating functions play particularly nicely with sums of random variables, turning them into products: for θ_i i.i.d.,

$$M_{\theta_1+\ldots+\theta_N}(\lambda) = E\left(e^{\lambda(\theta_1+\ldots+\theta_N)}\right) = \prod_{i=1}^{N} M_{\theta_i}(\lambda)\,.$$

Crucially, by mononicity,

$$\mathbb{P}(\theta \geq \epsilon) = \mathbb{P}\left(e^{\theta} \geq e^{\epsilon}\right)\ ,$$

so we can recast probability inequalities in terms of exponential functions!

Hoeffding's Lemma

Finally, we want to bound $E(e^{\lambda\theta})$.

Hoeffding's Lemma: Let X be a random variable on $[a, b]$ with $E(X) = 0$. Then for all $\lambda > 0$,

$$E(e^{\lambda X}) \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

Proof: Since e^x is convex on $[a, b]$,

$$e^{\lambda X} \leq te^{\lambda a} + (1-t)e^{\lambda b}$$

for all $t \in (0, 1)$. Setting $t = \frac{b-x}{b-a}$ and taking expectation value we find

$$E(e^{\lambda X}) \leq \frac{b - E(X)}{b-a}e^{\lambda a} + \frac{E(X) - a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.$$

This last function can be bounded by Taylor approximation, giving the result. □

Hoeffding's Inequality

Now that we have the tools, lets recall the statement of Hoeffding's inequality:

Hoeffding's Inequality: Let $\theta_1, \ldots, \theta_N$ be i.i.d. random variables with $\mathbb{P}[a \leq \theta \leq b] = 1$. Define $\theta = \sum \theta_i$. Then for any $\epsilon > 0$,

$$\delta = \mathbb{P}\left[\left| \frac{1}{N} \sum_{i=1}^{N} (\theta_i - E[\theta_i]\,) \right| > \epsilon \right] \leq 2 \exp\left(-2N\epsilon^2/(b-a)^2 \right) .$$

Proof of Hoeffding's Inequality

Proof: Let $X_i = \theta_i - E[\theta_i]$ and $X = \frac{1}{N}\sum_{i=1}^N X_i$. By Markov's inequality,

$$\mathbb{P}[X \geq \epsilon] = \mathbb{P}\left[e^{\lambda X} \geq e^{\lambda\epsilon}\right] \leq E[e^{\lambda X}]e^{-\lambda\epsilon}$$

Since the X_i's are i.i.d.,

$$E[e^{\lambda X}] = \prod_{i=1}^N E\left[e^{\frac{\lambda X_i}{N}}\right] \leq \prod_{i=1}^N e^{\frac{\lambda^2(b-a)^2}{8N^2}} = e^{\frac{\lambda^2(b-a)^2}{8N}}$$

by Hoeffding's Lemma. Combining these results and setting $\lambda = 4N\epsilon/(b-a)^2$ gives

$$\mathbb{P}[X \geq \epsilon] \leq \exp\left(-2N\epsilon^2/(b-a)^2\right) .$$

Repeating the above for $\mathbb{P}[X \leq -\epsilon]$ and adding the results completes the proof. □

Hoeffding's Inequality

A little bit of algebra allows us to write the following:

Hoeffding's Inequality II: If we choose

$$N \geq \frac{(b-a)^2}{2\epsilon^2} \log \frac{2}{\delta}\,,$$

then, with probability $1-\delta$, the difference between the empirical mean $\frac{1}{N}\sum_{i=1}^{N}\theta_i$ and the true mean $E(\theta)$ is less than ϵ.

In a very general sense this is telling us about the relative "cost" of accuracy ϵ vs confidence δ. In short, accuracy is expensive, while confidence is cheap.

Hoeffding's Inequality

For example, according to

$$N \geq \frac{(b-a)^2}{2\epsilon^2} \log \frac{2}{\delta} \,,$$

if we want to increase the confidence 10-fold, $\delta \to \frac{\delta}{10}$,

$$N \geq \frac{(b-a)^2}{2\epsilon^2} \log \frac{2 \cdot 10}{\delta} = N + C_{\epsilon,a,b}$$

we have to add a fixed number of samples. If we want 100 times more confidence, just add $2C$ data points.

However, increase accuracy 10-fold, we need 100 times the number of samples.

Finite Hypothesis Classes are APAC Learnable

Finite Hypothesis Classes are APAC Learnable

We are in a position to prove our first big theorem. The fact that finite hypothesis classes are APAC learnable is an important yardstick in machine learning both as a theoretical comparison and because practically our hypothesis classes aren't infinite.

For example, $\mathcal{H}$ could be the set of predictors that can be implement in 10^{10} bits in Python. Our mango example is infinite, but if we discretize it (say by encoding the color and softness parameters as 64 bit floats) it becomes finite.

In fact, any time we discretize the data, the set of all functions $h : \mathcal{X} \to \mathcal{Y}$ be comes finite. However, this doesn't necessarily fix things for us; as we just saw accuracy isn't cheap.

Finite Hypothesis Classes are APAC Learnable

Finite Hypothesis Classes are APAC Learnable: Let $(\mathcal{Z}, \mathcal{D}, \ell)$ be a domain, distribution and loss function and let $\mathcal{H}$ be a finite hypothesis class. Fix $\epsilon, \delta \in (0, 1)$.

By uniform convergence, we need to find a sample size N the guarantees (with probability at least $1 - \delta$) that $h \in \mathcal{H}$,

$$|L_{\mathcal{T}}(h) - L_{\mathcal{D}}| \leq \epsilon.$$

This is equivalent to showing

$$\mathcal{D}^N\Big(\big\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\big\}\Big) < \delta\,.$$

Finite Hypothesis Classes are APAC Learnable

Finite Hypothesis Classes are APAC Learnable: Let $(\mathcal{Z}, \mathcal{D}, \ell)$ be a domain, distribution and loss function and let $\mathcal{H}$ be a finite hypothesis class. Fix $\epsilon, \delta \in (0, 1)$.

By uniform convergence, we need to find a sample size N the guarantees (with probability at least $1 - \delta$) that $h \in \mathcal{H}$,

$$|L_{\mathcal{T}}(h) - L_{\mathcal{D}}| \leq \epsilon.$$

This is equivalent to showing

$$\mathcal{D}^N\left(\boxed{\left\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\right\}}\right) < \delta\,.$$

Lets describe the set of bad training sets.

Finite Hypothesis Classes are APAC Learnable

The set of bad training sets can be written as the union of the bad sets for each *h*:

$$\left\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\right\} = \bigcup_{h\in\mathcal{H}} \left\{\mathcal{T} : |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\right\}.$$

We then use the union bound to write

$$\begin{aligned}\mathcal{D}^N\Big(\left\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\right\}\Big)\\ \leq \sum_{h\in\mathcal{H}} \mathcal{D}^N\Big(\left\{\mathcal{T} : \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\right\}\Big).\end{aligned}$$

Finite Hypothesis Classes are APAC Learnable

Let $\theta_i = \ell(h, z_i)$. Since h is fixed and z_i are i.i.d., θ_i are also i.i.d.. As before, $L_{\mathcal{T}}(h) = \frac{1}{N} \sum \theta_i$ and $L_{\mathcal{D}}(h) = E[\ell(h, z_i)] = \mu$. Finally, for simplicity, lets assume that the range of ℓ is $[0, 1]$. Hoeffding's inequality gives that for any $h \in \mathcal{H}$,

$$\mathcal{D}^N\Big(\big\{\mathcal{T} : \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\big\}\Big) = \mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} \theta_i - \mu\right| > \epsilon\right] \leq 2e^{-2N\epsilon^2}\,.$$

Summing over the contributions for all $h \in \mathcal{H}$ yields

$$\begin{aligned}\mathcal{D}^N\Big(\big\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\big\}\Big) &\leq \sum_{h \in \mathcal{H}} 2e^{-2N\epsilon^2} \\ &= 2|\mathcal{H}|e^{-2N\epsilon^2}\,.\end{aligned}$$

Finite Hypothesis Classes are APAC Learnable

Therefore, if we choose

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

then

$$\mathcal{D}^N\Big(\big\{\mathcal{T} : \exists h \in \mathcal{H}\,, |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| > \epsilon\big\}\Big) < \delta\,. \ \square$$

Finite Hypothesis Classes are APAC Learnable

Finite Hypothesis Classes are APAC Learnable: We have shown that $\mathcal{H}$ enjoys the uniform convergence property with

$$N_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \text{ceil} \left(\frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right) .$$

Therefore, the class is APAC learnable using ERM with sample complexity

$$N_{\mathcal{H}}(\epsilon, \delta) \leq N_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \text{ceil} \left(\frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right) .$$

Discretizing

As a final note, we can discretize a infinite hypothesis class to get a rough bound on the complexity. For example, each function in the class of interval indicators $I_{[a,b]}$ depends on two real parameters $a, b \in \mathbb{R}$ which we could represent as, say, 64 bit floats. In general, if we represent d parameters as b bit floats, there are at most

$$2^{d \times b}$$

such functions in $\mathcal{H}$. Applying the bound for finite classes, the sample complexity is at most

$$N \leq \frac{b \cdot d \log 4 + 2 \log(2/\delta)}{\epsilon^2} .$$

The sample complexity is linear in the number of parameters.

References

References Hoeffding's inequality can be found in Appendix B, or

https://people.cs.umass.edu/~domke/courses/sml2010/10theory.pdf

http://cs229.stanford.edu/extra-notes/hoeffding.pdf

http://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/notes/03_hoeffding.pdf

This lecture covers chapters 3 and 4 of Shalevl-Shwarts and Ben-David.