Machine Learning

Understanding Machine Learning and its Applications

Akash Cherukuri

The reference for these notes have been taken from “Understanding Machine Learning: From Theory to Algorithms ©2014 by Shai Shalev-Shwartz and Shai Ben-David"

Contents

1 Formal Learning Models 4

1.1 PAC Learning 4

1.2 Agnostic PAC Learning 4

1.2.1 Redefining Data Generating Distribution 5

1.2.2 Redefining Risk Calculations 5

1.2.3 Bayes' Optimal Predictor 5

1.3 Extending the Scope of PAC Learning 6

1.4 Learning via Uniform Convergence 7

1.5 Finite Hypothesis Classes are Agnostic PAC Learnable 7

2 VC Dimension 9

2.1 No Free Lunch Theorem 9

2.2 Error Decomposition 10

2.3 VC Dimensions 11

1. Formal Learning Models

1.1 PAC Learning

We have seen previously that for a finite hypothesis class $\mathcal{H}$, assuming that the *Realizability Assumption* holds, for a sufficiently large sample S, the ERM algorithm yields an estimator which is *Probably Approximately Correct*. We will now define PAC learning formally.

Definition 1.1.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is said to be *PAC Learnable* iff there exists a function $m_{\mathcal{H}}(0,1)^2 \to \mathbb{R}$ and a learning algorithm for every distribution $\mathcal{D}$ over χ and every binary "True Label" f, for a sample size larger than $m_{\mathcal{H}}(\varepsilon,\delta)$ under the *Realizability Assumption*, yields an estimator which is *Accurately Correct* to a margin of ε with a *Probability* of $(1-\delta)$.

Notice that the minimum number of samples needed depends on the parameters (ε,δ). The function $m_{\mathcal{H}}$ is said to be the *Sample Complexity* of learning $\mathcal{H}$, as it gives an estimate of the size of the training sample needed. It is clear that there might exist infinitely many such functions, therefore we define the sample complexity to be the minimal of all such possible functions.

We have already derived a bound for the sample complexity on case that $\mathcal{H}$ was finite:

$$m_{\mathcal{H}} \leq \left\lceil \frac{\log(|\mathcal{H}|)/\delta}{\varepsilon} \right\rceil$$

1.2 Agnostic PAC Learning

We now try to relax the assumptions so far. The two main assumptions that we've been using are:

1. Realizability Assumption
2. The labelling function is binary in nature

We will be focusing on the Realizability assumption for now. It may be too strong in some cases, as the Data set need not have a subset which depicts the entire data set properly. Also, two data sets

being equal need not necessarily imply that their labels are equal as well. For example, we have two papayas with the same color and hardness, but one of the papayas is sweet and the other is bitter. We will be taking this into consideration as well.

1.2.1 Redefining Data Generating Distribution

Previously, we've taken that the distribution $\mathcal{D}$ is over $\mathcal{X}$, but this fails to account for the papayas example previously shown. We thus redefine the distribution to be over $(\mathcal{X}, \mathcal{Y})$, from which the marginal distribution $\mathcal{D}_{\mathcal{X}}$ can be calculated and the data points generated. This distribution is also renamed as *Data Labels Generating Distribution* to take into account that even $\mathcal{Y}$ is distributed according to $\mathcal{D}$ now.

1.2.2 Redefining Risk Calculations

By redefining the data generating distribution, we have also rendered the true error calculation obsolete. However, using the new definition of $\mathcal{D}$ we can redefine the true risk to be the *Probability of picking a pair* (x,y) *such that* $h(x) \neq y$. More formally,

$$\mathcal{L}_{(x,y)\sim D} = P[\{(x,y) : h(x) \neq y\}]$$

However, the distribution is unavailable to the agent during its training. Thus the definition of empirical risk is unaffected by these changes, and it remains as follows. Notice that the empirical loss is the same as assuming that the distribution $\mathcal{D}$ is *Uniform* in the sample set S.

$$\mathcal{L}_S(h_S) = \frac{|\{x : x \in S, h(x) \neq y\}|}{m}$$

1.2.3 Bayes' Optimal Predictor

For any probability distribution $\mathcal{D}$ over a binary label $\{0,1\}$, the estimator which gives the least error is the following:

$$h^*(x) = \begin{cases} 1; & \text{if } P(y=1|x) \geq 0.5 \\ 0; & \text{otherwise} \end{cases}$$

However, because the distribution is unknown, we cannot use this predictor. However we can use the idea that there is a minimum probable risk that any estimator can attain to define when a hypothesis class is said to be *learnable*.

Definition 1.2.1 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is said to be *Agnostic PAC Learnable* iff there exists a function $m_{\mathcal{H}}(0,1)^2 \to \mathbb{R}$ and a learning algorithm for every distribution $\mathcal{D}$ over $\mathcal{X}\{0,1\}$, for a sample size larger than $m_{\mathcal{H}}(\varepsilon, \delta)$, which yields an estimator which is *Accurately Correct* to a margin of $\varepsilon + \min_{h\in\mathcal{H}}(\mathcal{L}_{\mathcal{D}}(h))$ with a *Probability* of $(1-\delta)$.

Notice that the error now corresponds to the relative best that the hypothesis class can produce, which would become the absolute error is we took that the Realizability Assumption is valid. This is therefore, a more general way of defining the term.

1.3 Extending the Scope of PAC Learning

So far, we've taken that the label set is *binary* in nature. This was the second assumption which we wished to generalize in section 1.2. Extending the label set to be finite, we can understand that the redefined terms of data generating distribution and the rick functions still hold good. For example, let the label set $\mathcal{Y}$ be the classification of emails, and our agent is supposed to learn how to do this properly. We can see that, in this case, the risk function would become *"Probability that the assigned label by the estimator is incorrect"*, which can be found using the data-labels generating distribution.

However, if the label set was not finite we would have to make some changes. For example, let's say we wish to find the approximate dimensions of an object by looking at its picture alone. The label set, in this case is referred to as the *Target Set*, would be the set of all positive real numbers. Notice how we cannot use the earlier definition of loss functions; hence we iterate them.

Definition 1.3.1 (Generalized Loss Functions) Let $\mathcal{H}$ be the Estimator Class, and $\mathcal{Z}$ be some domain. Any function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ is said to be a loss function.

In the case of estimators, the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ but this need not always be the case. We can see that the distribution $\mathcal{D}$ would now be over the set $\mathcal{Z}$. Therefore, using this new definition of a loss function, we can see that the True Risk would be the expected value of the loss function. Formally,

$$\mathcal{L}_{\mathcal{D}}(h) = \underset{z\sim\mathcal{D}}{E}[l(h,z)]$$

The definition of empirical risk shall be similar:

$$\mathcal{L}_S(h) = \frac{1}{m}\sum_{z\in\mathcal{Z}} l(h,z)$$

The definition of *Agnostic Learnability* doesn't change much, as there is only a significant change in the way that the loss functions are defined, and it would imply that a change would occur in the way that we calculate the minimum possible loss for a given class of estimators.

1.4 Learning via Uniform Convergence

Till now, we've discussed with and without using the Realizability Assumption that we could obtain an estimator for a finite hypothesis class which models the sample properly. However, this doesn't imply that the other estimators, with a similar empirical risk, would model the data set properly. For instance, assume we obtained another estimator whose $\mathcal{L}_S(h)$ was nearly the same, but was much easier to compute. We have no way of saying whether this estimator would model the data set properly. This is the inspiration behind *Uniform Convergence*. We essentially try to ensure that the risks of all the estimators are uniformly convergent to their true values.

Definition 1.4.1 (ε-Representative Sample) A sample set S is said to be ε-representative (w.r.t given $\mathcal{Z}, \mathcal{D}, f, \mathcal{H}$) if for all $h \in \mathcal{H}$ the modulus of the difference between true risk and the empirical risk is lesser than ε.

$$|\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \leq \varepsilon$$

There is a very interesting relation between the two ε that have been used here, and in the definition of PAC Learnability. This has been stated in the theorem below, following its proof.

Theorem 1.4.1 Assume that the sample S is $\frac{\varepsilon}{2}$-representative (w.r.t given $\mathcal{Z}, \mathcal{D}, f, \mathcal{H}$), then any estimator $h_S = \text{ERM}_{\mathcal{H}}$ would satisfy the following inequality.

$$\mathcal{L}_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) + \varepsilon$$

That is, if we could show that the probability of obtaining such a sample S was at least $(1-\delta)$, then the ERM-Hypothesis would be Agnostic Learnable (from definition 1.2.1). This forms the basis of defining what *Uniform Convergence* means.

Definition 1.4.2 (Uniform Convergence) A hypothesis class $\mathcal{H}$ is said to be *Uniformly Convergent* wrt the label function f and the domain $\mathcal{Z}$ iff for every distribution $\mathcal{D}$ over $\mathcal{Z}$ and every $(\varepsilon, \delta) \in (0,1)^2$, there exists a function $m^{UC} : (0,1)^2 \to \mathbb{N}$ such that a sample S with size greater than $m^{UC}(\varepsilon, \delta)$ would be ε-representative with a probability of $(1-\delta)$.

Similar to how we had defined the sample complexity to be the value of $m_{\mathcal{H}}(\varepsilon, \delta)$, we define the sample complexity for the hypothesis class to be uniformly convergent as $m^{UC}_{\mathcal{H}}(\varepsilon, \delta)$. It is clearly evident that $m_{\mathcal{H}}(\varepsilon, \delta) \leq m^{UC}_{\mathcal{H}}(\varepsilon/2, \delta)$. The proof can be arrived at via contradiction fairly quickly, because if the inequality were the other way around, it would imply that there would be cases such that the set is representative but not Agnostic PAC Learnable.

We have previously shown that PAC Learning occurs for all finite hypothesis classes, as long as they satisfy the realizability assumption. We now use the concept of uniform convergence to show that the realizability assumption is not necessary, and that there would analogously exist a minimum sample size.

1.5 Finite Hypothesis Classes are Agnostic PAC Learnable

Similar to what we had done previously, we with to arrive upon a result for a function which would state the minimum possible size of the sample needed. Let us define the problem statement first.

Given the values of ε, δ we need to find an expression for the m, such that;

$$|S| > m \implies \mathrm{P}\left[\forall h \in \mathcal{H} : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \leq \varepsilon\right] \geq 1 - \delta$$

Equivalently, we can write that

$$\mathrm{P}\left[\forall h \in \mathcal{H} : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\right] < \delta$$
$$\implies \mathrm{P}\left[\bigcup_{h \in \mathcal{H}} \{S : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\}\right] < \delta$$

We now try to maximize the LHS of the inequality to get minimum value of m.

$$\mathrm{P}\left[\bigcup_{h \in \mathcal{H}} \{S : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\}\right] \leq \sum_{h \in \mathcal{H}} \mathrm{P}\left[|\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\right]$$

If we look at the definitions of the loss functions, the True loss function is the expected value of the generalized loss function where as the Empirical loss function is the empirical mean of the values of the loss function over all the data points in the sample. Although we can say from *The Law of Large Numbers* that they both would be nearly equal asymptotically, this isn't true for finite values. We therefore use an inequality called the *Hoeffding's Inequality* which tells us how close they both can get.

Theorem 1.5.1 (Hoeffding's Inequality) Given m i.i.d random variables $x_1, x_2 \cdots x_m$ where $\mathrm{E}[x_i] = \mu$ and $\mathrm{P}[a \leq x_i \leq b] = 1$, then for any $\varepsilon > 0$ we have that:

$$\mathrm{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \mu\right| > \varepsilon\right] < 2\exp\left[\frac{-2m\varepsilon^2}{(b-a)^2}\right]$$

Assuming that the loss function has a range of $(0, 1)$, we can use this theorem in the inequality which we obtained earlier;

$$\begin{aligned}\mathrm{P}\left[\bigcup_{h \in \mathcal{H}} \{S : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\}\right] &\leq \sum_{h \in \mathcal{H}} \mathrm{P}\left[|\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| > \varepsilon\right] \\ &< 2|\mathcal{H}|\exp\left[\frac{-2m\varepsilon^2}{(b-a)^2}\right] \\ &< 2|\mathcal{H}|\exp\left[-2m\varepsilon^2\right]\end{aligned}$$

From this we obtain that;

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \implies \mathrm{P}\left[\forall h \in \mathcal{H} : |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \leq \varepsilon\right] \geq 1 - \delta$$

We have successfully shown that *Every finite hypothesis class is Uniformly Convergent, and thus by extension, Agnostic PAC Learnable*. This can be improperly extended to infinite hypothesis classes as well, because all numbers in practice are represented by, say, 64 bits. This would imply that the set $\mathbb{N}$ would be finite and have a size of 2^{64}. A more formal proof will be learnt further down the line.

2. VC Dimension

Let us review what we've done so far. We first came up with an algorithm to find an estimator which tries to minimize the true risk by assuming that the empirical risk is related to it. We then realised that this need not always be the case, and that *Overfitting* can occur. Inorder to combat this problem we induced a bias in the system by restricting the Hypothesis Class to a finite set of functions, using some prior knowledge about the problem.

> *However, is it not possible to have a learning algorithm which requires no prior knowledge, and can come up with an estimator for a binary label set with reasonable accuracy and a high enough probability?*

This is the question which we wish to answer here. The *No-Free-Lunch-Theorem* defined below states that such a learning algorithm is impossible for binary labelling. This means that some prior knowledge about the system is necessary to have a proper leaning algorithm, such as about the distribution $\mathcal{D}$ or whether the system satisfies the Realizability Assumption and the such.

2.1 No Free Lunch Theorem

We shall state and prove the No Free Lunch Theorem, and understand why having a universal learner is not possible.

Theorem 2.1.1 (No Free Lunch) Let A be a learning algorithm for the task of binary classification to the label set $\{0,1\}$ over the domain set $\mathcal{X}$. Let m represent the size of training data, and $m < |\mathcal{X}/2|$. Then, for any distribution D over the set $\mathcal{X} \times \{0,1\}$ we can state that:

1. There exists a function f such that $\mathcal{L}_D(f) = 0$
2. With a probability of at least 1/7 we have that $\mathcal{L}_D(A(S)) > 1/8$

Therefore from this theorem we can clearly see that no "One-Size-Fits-All" solution exists for our problem. Now we can establish *why* prior knowledge is so important by using this theorem.

Without any prior knowledge, we would have to consider the entire hypothesis set, $\mathcal{H}$, which would contain all the possible hypotheses. By contradiction, we can prove that no algorithm exists which can reliably output a predictor for such a case, i.e., this is not PAC Learnable.

This can proved by contradiction, assume that there existed an algorithm, A, with respect to which the problem is PAC Learnable. If we were to take some values of $\varepsilon < 1/7, \delta < 1/8$, we can quite clearly see that the assumption breaks down. *Therefore, to be able to reliably produce an estimator, prior information is mandatory.*

Now the problem boils down to choosing which Hypothesis class should be chosen, and how. On the one hand, we would be needed to have a class which either has the true estimator(which is unknown) with the realizability assumption, or if it doesn't, have the minimum value of the true error of the estimator to be low. On the other hand, if we were to try and increase the size of the Hypothesis Class, that would cause the decrease in the standards of the learning. We breakdown the error that is generated by the estimator to better understand the cause of this problem.

2.2 Error Decomposition

We know that the true error for an estimator h is represented as $\mathcal{L}_D(h)$. We now decompose the error as follows:

$$\mathcal{L}_D(h) = \varepsilon_{app} + \varepsilon_{est} = \min_{h' \in \mathcal{H}} \mathcal{L}_D(h') + \{\mathcal{L}_D(h) - \min_{h' \in \mathcal{H}} \mathcal{L}_D(h')\}$$

Here, ε_{app} stands for the *Approximation Error*, and ε_{est} stands for the *Estimation Error*. We shall look at what each of them represents subsequently.

1. *Approximation Error* quantifies the amount of error that is obtained by approximating that the estimator belongs to the estimator class. In other words, it quantifies the amount of inductive bias that has been introduced in the system. Under the realizability assumption, we can see that the approximation error is zero. *Also note that Bayes' Estimator gives a lower bound for what the approximation error can be.*
2. *Estimation Error* results because the emperical risk is only an estimate of what the true risk is, and it might be misleading as well. As we've seen already, *Estimation error increases logarithmically with $|\mathcal{H}|$ and decreases with m.* That is, this error is dependant on the complexity of the sample.

We strive to minimize both the errors as much as possible, but we can see that they are related in polar opposite manners to the size of the Hypothesis Class. Thus, we enter a trade-off, which is called as the *Bias-Complexity Tradeoff*. Causing the estimator class to be rich causes overfitting, wheras having a small hypothesis class causes underfitting. Most of empirical research strives on obtaining the perfect estimator class that excels in minimizing both to the maximum possible extent.

However, all this analysis has been done assuming that the hypothesis class, $\mathcal{H}$, is finite. Let us now shift our goals towards the analysis of PAC Learnability when the hypothesis class is infinite in size. Surely, there must be some characteristics common to all the infinite classes, and that is what we wish to uncover.

2.3 VC Dimensions

In order to show that infinite hypothesis classes do exist, we first give an example of such a class, and prove that it is in fact PAC Learnable.

Example 2.1 Define a hypothesis class, $\mathcal{H}$ as;

$$\mathcal{H} = \{h_a : \forall a \in \mathbb{R}\} \text{ where } h_a(x) = \begin{cases} 1, & x > a \\ 0, & \text{otherwise} \end{cases}$$

Clearly the hypothesis class is infinite. To prove that this is PAC learnable, with a sample complexity of $m_{\mathcal{H}} \leq \left\lceil \frac{\log(2/\delta)}{\varepsilon} \right\rceil$, we first label the true value of a as a_t, such that $\mathcal{L}_D(h_{a_t}) = 0$. Now assume that probability of the error which we can allow in the measurement of a_t is ε. That means;

$$\exists (a_o, a_1) \in \mathbb{R}^2 \text{ such that } \mathrm{P}[x \in (a_o, a_t)] = \mathrm{P}[x \in (a_t, a_1)] = \varepsilon$$

Let the sample-label set be S, which is obviously finite. Ideally, we know that all the values of $\{x, 1\} \in S$ would have $x < a_t$, and the rest would be $\{x, 0\}$, $x > a_t$. Therefore, for our estimator to be able to give a value in between (a_o, a_1), we can see that;

$$\max_{\{x,1\} \in S}(x) > a_o \text{ AND } \min_{\{x,0\} \in S}(x) < a_1$$

For simplicity, we shall represent $\max_{\{x,1\} \in S}(x)$ as b_o and $\min_{\{x,0\} \in S}(x)$ as b_1. Let the parameter that is obtained by the ERM estimator be a_{ERM}. We can see that $a_{ERM} \in (b_o, b_1)$. Therefore, the condition mentioned above is *SUFFICIENT* for the learning to be successful. This implies;

$$\begin{aligned} \mathrm{P}[\mathcal{L}_D(h_{ERM}) < \varepsilon] &\geq \mathrm{P}[b_o > a_o \wedge b_1 < a_1] \\ \mathrm{P}[\mathcal{L}_D(h_{ERM}) > \varepsilon] &\leq \mathrm{P}[b_o < a_o \vee b_1 > a_1] \\ &\leq \mathrm{P}[b_o < a_o] + \mathrm{P}[b_1 > a_1] \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} \qquad \text{(Using sample complexity)} \\ &\leq \delta \end{aligned}$$

Therefore, the hypothesis class is PAC Learnable as shown using the ERM hypothesis. ■

Therefore, we have seen that finiteness of a hypothesis class is a sufficient condition for learnability, and it is not necessary. We shall now use the definition of VC Dimensions to create a classification which would enable us to differentiate which classes are learnable and which are not.