CS6781 - Theoretical Foundations of Machine Learning

Lecture 7: Introduction to Agnostic Learning

February 11, 2020

Lecturer: Nika Haghtalab *Readings: None*
Scribe: Rodrigo Delgado

Up to this point in the course, we have been investigating PAC learning under the realizability assumption. Namely, we have been assuming that our instances are labeled by a concept c^* in our concept class $\mathcal{C}$. This has allowed us to design algorithms that, given sufficiently many labeled instances, can, with arbitrarily high probability, find a concept $c \in \mathcal{C}$ with an arbitrarily small error. We will now explore a learning model that does not assume the existence of any such $c^* \in \mathcal{C}$ that generates the labels!

Abandoning the realizability assumption is crucial, because in many application it is unreasonable to assume that it is satisfied. Consider, for example, the task of learning how to tag social media posts as "appropriate" or "innappropriate". In this case, our domain $\mathcal{X}$ is a set of representations of posts — such as the author, bag of words, and hashtags — on a social media platform. $\mathcal{Y}$ is the set {"appropriate", "innapropriate"}, and S is a subset of $\mathcal{X} \times \mathcal{Y}$, where each labeled instance $(x, y) \in S$ was labeled by someone (a user, an employee, an investigator, etc). In this environment, the realizability assumption might fail for any of the following reasons:

1. It may be impossible to find a function that accurately labels all of the instances in S. For example, one appropriate post and one inappropriate post may map to the same feature vector in $\mathcal{X}$, i.e., have the same tags, bag of words representation, etc. It is also possible that a labeler may be making mistakes when deciding whether a post is appropriate for publication.

2. Even if it is possible to find a function that accurately labels all of the instance in S, it is not necessarily the case that this function will lie in the relevant concept class $\mathcal{C}$. Expanding the concept class to include a perfectly consistent function increases the complexity of the concept class and increases its VC dimension. The larger the VC dimension, the larger the sample set S that we need to guarantee PAC learnability.

1 Notations

To meaningfully address the world beyond realizability, we need to reframe a few of the definitions we have been using so far.

So far, we have been speaking of $\mathcal{D}$ as a distribution over $\mathcal{X}$. But moving forward we will be thinking of $\mathcal{D}$ as a joint distribution over $\mathcal{X} \times \mathcal{Y}$. This is just to say that instead of first picking $x \in \mathcal{X}$ according to a distribution and having $y = c^*(x)$, we will pick $(x, y) \in \mathcal{X} \times \mathcal{Y}$ at once according to a distribution. This formality opens our model to the possibility outlined in 1. Namely, it will be possible to pick $(x, y) \sim \mathcal{D}$ and $(x, y') \sim \mathcal{D}$ with $y \neq y'$. That said, a joint distribution

can be decoupled. After all, picking an element of $\mathcal{X} \times \mathcal{Y}$ should be the same as first picking an element in $\mathcal{X}$ and then picking an element in $\mathcal{Y}$. That is for any joint probability distribution $P(X, Y) = P(X)P(Y|X)$. More specifically, define $\mathcal{D}_{\mathcal{X}}$, the marginal distribution of $\mathcal{D}$ with respect to $\mathcal{X}$, by

$$\Pr_{\mathcal{D}_{\mathcal{X}}}[X] = \int_{y \in \mathcal{Y}} \Pr_{\mathcal{D}}[X, Y = y].$$

The conditional label distribution of $\mathcal{D}$ with respect to $x \in \mathcal{X}$ is $\Pr[Y] = \Pr_{\mathcal{D}}[(X, Y) \mid X = x]$. That is, the conditional label distribution of $\mathcal{D}$ with respect to $x \in \mathcal{X}$ is the probability of seeing label y i conditional on having picked x. Then we can write the joint distribution $\mathcal{D}$ as a product of the marginal distribution $\mathcal{D}_{\mathcal{X}}$ with the conditional label distribution $\mathcal{D}_{\mathcal{Y}|x}$:

$$\Pr_{\mathcal{D}}[(x, y)] = \Pr_{\mathcal{D}_{\mathcal{X}}}[x] \cdot \Pr[y|x].$$

Similarly, our notions of error must be adapted to fit the use of a joint distribution. In particular, after fixing an $h \in \mathcal{C}$, we define the true error of h by

$$\text{err}_{\mathcal{D}}(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]. \tag{1}$$

And for any set of m labeled instances $S = \{(x_1, y_1), \ldots (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$, we define the empirical error of h by

$$\text{err}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(h(x_i) \neq y_i). \tag{2}$$

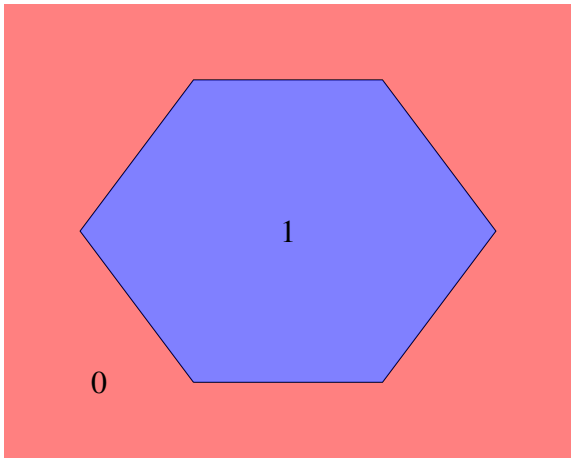
This is a generalization of the errors we saw in the realizable setting. That is, in the case of a joint distribution on $\mathcal{X} \times \mathcal{Y}$ that is obtained from a distribution on $\mathcal{X}$ and allowing $y = c^*(x)$ according to a chosen concept $c^* \in \mathcal{C}$, then (1) and (2) errors in this case are equal to the true and empirical errors we have been using under the realizability assumption.

2 Learnability without Realizability

Let us discuss multiple natural goals that one may have in absence of realizability and then define a new model of learnability for this setting.

Suppose that, as in the second explanation of noise, there is a deterministic way of labelling $\mathcal{D}$, but any concept h such that $\text{err}_{\mathcal{D}}(h) = 0$ lies outside of the concept class $\mathcal{C}$. For example, suppose $\mathcal{X}$ is $\mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and the concept class $\mathcal{C}$ is the set of linear thresholds in $\mathbb{R}^2$, but the

distribution $\mathcal{D}$ that is generated by the characteristic function of a polygon as below.



That is, if x is in the blue region of the plane, $y = 1$, and if x is in the red region of the plane $y = 0$. It should be clear that there is no linear threshold on the plane that accurately labels $\mathcal{D}$. But, by construction, there is a polygon whose characteristic function accurately labels $\mathcal{D}$. So, the unavoidable question is, is it reasonable to expand our concept class $\mathcal{C}$ to incorporate the characteristic functions of all polygons? Unfortunately, if we were to do this, we would sacrifice the prospect of PAC learnability, for, by expanding our concept class to include all polygons, we would increase the VC Dimension of it to "infinity". In turn, we would need infinitely many samples in order to learn $\mathcal{D}$. More generally expanding $\mathcal{C}$ will come at the cost of the sample complexity required for PAC learnability. Of course, this isn't always undesirable. If, for example, we were trying to learn the characteristic function of a rectangle, it *would* be reasonable to expand the class of linear thresholds to include the characteristic functions of rectangles. After all, such an expansion of the concept class would not increase the VC Dimension to "infinity".

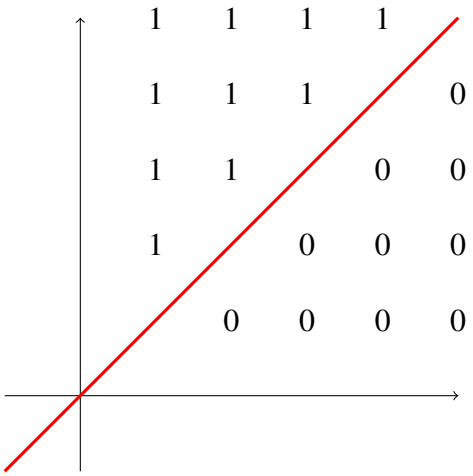
Suppose, instead, that, as in the first explanation of noise, there is no deterministic way of labelling S. For a simple example, suppose $\mathcal{X} = \{0\}$ and $\mathcal{Y} = \{0, 1\}$. But S contains 100 labeled instances: 51 of them are labeled 0 and the other 49 are labeled 1. In this case, the most accurate function we can cook up maps 0 to 0. It *will* be wrong almost half of the time, but, on the flip side, it will be right more than half of the time. And the only other possible option, the function that maps 0 to 1 will be wrong more often. This "best we can do" choice is an example of a Bayes Optimal Classifier. Slightly more generally, if $\mathcal{Y} = \{0, 1\}$ is any set of binary labels and $\mathcal{D}$ is a joint distribution of $\mathcal{X} \times \mathcal{Y}$, the Bayes Optimal Classifier $h_{\text{Bayes}} : \mathcal{X} \to \mathcal{Y}$ for $\mathcal{D}$ is defined by

$$h_{\text{Bayes}}(x) = \begin{cases} 0 & \text{if } \Pr[Y = 0|x] \geq 1/2 \\ 1 & \text{otherwise} \end{cases} .$$

The Bayes Optimal Classifier is *optimal* in the sense that it achieve minimum error that *any* function can achieve.

Of course, we can cook up distributions on which the Bayes Optimal Classifier is wrong with

probability greater than $1/2 - \epsilon$ for arbitrarily small $\epsilon \in \mathbb{R}$. For instance,



where every point above the red line is labeled 1 with probability $1/2 + \epsilon$ and labeled 0 with probability $1/2 - \epsilon$, and every point below the red line is labeled 0 with probability $1/2 + \epsilon$ and labeled 1 with probability $1/2 - \epsilon$. Then the Bayes Optimal Classifier will label every point above the red line by 1 and every point below the red line by 0. However, by construction, the Bayes Optimal Classifier will be wrong with probability $1/2 - \epsilon$. Of course, the catch is that, no matter what, the Bayes Optimal Classifier will be right with probability greater than $1/2$. Even more generally, if $\mathcal{Y}$ is any set of binary labels and $\mathcal{D}$ is a joint distribution of $\mathcal{X} \times \mathcal{Y}$, the Bayes Optimal Classifier $h_{\text{Bayes}} : \mathcal{X} \to \mathcal{Y}$ for $\mathcal{D}$ is defined by

$$h_{\text{Bayes}}(x) = \arg\max_y \Pr[Y = y|x].$$

That is h_{Bayes} maps x to the element y that is most likely to be paired with x under $\mathcal{D}$. All of this suggests that an algorithmic approach to learning a joint distribution $\mathcal{D}$ for which there is no deterministic labelling. Namely, find the Bayes Optimal Classifier for $\mathcal{D}$. For in this case we will have found the function that is the most accurate on $\mathcal{D}$. The limitation of this approach is that the conditional label distribution $\Pr[Y = y|x]$ is unknown. In fact, in the worst case, we need to observe every x many times in order to form a good estimate of $\Pr[Y = y|x]$. This could require and infinitely large sample set. So unless h_{Bayes} is known to come from a concept class with a small VC-dimension, it is possible that approximating h_{Bayes} with high enough probability requires "infinitely" many samples.

With all this said, we will now present a model for PAC learning outside of the realizable setting that is not susceptible to the pitfalls outlined above.

Definition 2.1. (Agnostic PAC Learning) An algorithm $\mathcal{A}$ agnostically PAC learns a class of hypothesis $\mathcal{C}$ if there is a function $m_{\mathcal{C}}(\epsilon, \delta)$ such that the following is true: for any joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, any $\epsilon > 0$, and any $\delta > 0$, and an i.i.d. sample set S of at least $m_{\mathcal{C}}(\epsilon, \delta)$ samples, $\mathcal{A}(S)$ returns $h_{\mathcal{A}}$ such that, with probability greater than $1 - \delta$,

$$\text{err}_{\mathcal{D}}(h_{\mathcal{A}}) \leq \min_{c \in \mathcal{C}} \text{err}_{\mathcal{D}}(c) + \epsilon.$$

In the next lectures we will study the sample complexity of agnostic learning.