Indian Institute of Technology, Kanpur
Statistical and Algorithmic Learning Theory, 18–19

Instructors: Purushottam Kar
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE

4

PAC Learning and Agnostic Setting

1. PAC Learning

In the previous scribe, we have shown that for a finite hypothesis space, the ERM model for a sample with a sufficiently large size, independent of the distribution will be *probably approximately correct.* We can now formally define Probably Approximately Correct (PAC) model.

Definition 4.1 (PAC Learnability, [?]). A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every prediction function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}$, $\mathcal{D}$, f, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis $\hat{f}_S$ for a training sample S and loss function l such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{\mathcal{D},f}(\hat{f}) \leq \epsilon$.

In the above definition, ϵ is defined as the accuracy parameter, which allows a margin for error in prediction. This is attributed to the fact that S is a finite sample, and there is a chance that it does not very faithfully represent the real distribution. δ is known as the confidence paramter.

$m_{\mathcal{H}}$ is the *sampling complexity* of the hypothesis class $\mathcal{H}$ *i.e.* the number of minimum samples required to guarantee a probably approximately correct algorithm. From the previous scribe, we know that if

$$m \geq \frac{\log\left(\left|\,\mathcal{H}\,\right| / \delta\right)}{\epsilon} \tag{1}$$

then we can have a PAC solution for the $\mathcal{H}$. Therefore, the minimum sampling complexity must be less than this value. Hence, we can write

$$m_{\mathcal{H}} \leq \left\lceil \frac{\log\left(\left|\,\mathcal{H}\,\right| / \delta\right)}{\epsilon} \right\rceil \tag{2}$$

2. Agnostic Setting — Releasing the Realizability assumption

For now, we have assumed that the distribution $\mathcal{D}$ is realizable *i.e.* $\exists\, f \in \mathcal{F}$ such that $\mathrm{er}^{l}_{D}\left[\, f \,\right] = 0$ for some loss function. However this is not a realistic setting, considering there can be noise in the observed data points $\mathcal{X}, \mathcal{Y}$. It might be unwise to assume that the labels are completely determined by the features.

For example, consider a binary classification problem, where for two feature vectors $\mathbf{x}^1$ and $\mathbf{x}^2$ have the same values of the features however $y^1 \neq y^2$. In this case, it is never possible to find a prediction function that gives zero error on the distribution.

Therefore, we must relax the realizability assumption. In this case, we redefine the distribution $\mathcal{D}$ to be a joint probability distribution over the feature space and the output space.

Our goal is the same. We wish to find a prediction function f^* that minimizes the l-risk. It should be clear that since we generally do not know the distribution $\mathcal{D}$, we cannot find the optimal prediction function f^*, however we can only find a predictor that is probably approximately close to f^*. Hence, we define PAC learning, however, now for agnostic setting.

Definition 4.2 (Agnostic PAC Learnability, [?]). A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis f_S for a training sample S and a loss function l such that, with probability of at least $1 - \delta$, (over the choice of the m training examples),

$$\mathrm{er}^l_S\,[\, f_S \,] \quad \leq \quad \min_{f' \in \mathcal{H}} \mathrm{er}^l_D\,[\, f' \,] + \epsilon$$

Note. The definition of Agnostic PAC Learnability boils down to the PAC Learning under Realizable Assumption if the Assumption indeed holds true, as $\min_{f \in \mathcal{H}} \mathrm{er}^l_D\,[\, f \,] = 0$

Remark. We assume the loss function to be measurable for all functions and at all points *i.e.* $\forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}$, $\forall\, f \in \mathcal{H}$, $l(f(x), y)$ is measurable. Note that if the loss function is not measurable, then it cannot be a random variable.

References

[1] Shai Shalev-Shwartz, Shai Ben-David *Understanding Machine Learning from Theory to Algorithms.* http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning