Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Prannay Khosla
Instructor: Purushottam Kar
Date: January 10, 2018

LECTURE
13

Consistency and Calibration

1 Introduction

In the previous lectures we have covered uniform convergence, VC dimension and other concepts which help us formally study the proceses of Learning. We have also studied Algorithmic Stability of learning and algorithmic techniques in learning. We have covered Emipirical risk minimization, structural risk minimization and regularized risk minimization and analysed them well. We are now ready to define the notion of **Learnability** of a problem, in the ϵ, δ we have been studying it before. That is to say, till now we have seen results of the following form :

$$\mathbb{P}\left[|er^l_{\mathcal{D}} - er^l_{\mathcal{S}}| > \epsilon\right] \leq \delta$$

where we claim that an event happens with high probability $(1-\delta)$ for some given bound ϵ, where $\exists\delta$ for every ϵ chosen from some interval (generally $[0,1]$). Now our aim is to study why such bounds help us in learning, and make generalized claims, where if we know such bounds hold, then what that implies.

We also have studied the l^{0-1} loss function, but also realized that we do not actually ever use the l^{0-1} loss in general but actually use the sigmoid loss, or the hinge loss. We want to study why it suffices to study that and in what settings.

2 Bayes Risk

We start by defining some quantities that will help us analyse the give problem:

Definition 13.1. Bayes Risk : It is defined with respect to a loss function for every distribution $\mathcal{D}$ as follows :

$$er^{l,*}_{\mathcal{D}} \triangleq \min_{f:measurable} er^l_{\mathcal{D}}[f]$$

It is the true optimal minima that can be acheived by for a given loss function and distribution $\mathcal{D}$ from which the samples are being drawn. The function in context can by any measurable function, ensuring that the error term is well defined, since we know that function l is Borel Measurable. This in-turn is also a measure in the noise of the sample's labels. This value can not be smaller than the inherent noise in the system, and is information theoretically the optimum we can ever hope to reach, with any learning system.

Definition 13.2. Bayes Predictor For the above expression, the minimum would be achieved by a function (call it f^*), then function f^* is Bayes predictor.

$$er^{l,*}_{\mathcal{D}} \triangleq \arg\min_{f:measurable} er^l_{\mathcal{D}}[f]$$

Definition 13.3. Regret / Excess Risk It is a measure of how far away we are from the optimal. In some sense it can be understood as the mis-classification rate of the function under consideration, or a measure of the how the different the values predicted by us are from the true values.

$$\mathcal{R}^l_{\mathcal{D}}[f] \triangleq er^l_{\mathcal{D}}[f] - er^{l,*}_{\mathcal{D}}$$

Definition 13.4. Regret for a function class The same way we defined the excress risk for a given function, we can define it for a function class. It would act as a measure of how well the function class can learn the said function or parameter that is being learnt. It measures the best we can do with a said function class.

$$er^{l,\mathcal{F}}_{\mathcal{D}} \triangleq \min_{f \in \mathcal{F}} er^l_{\mathcal{D}}[f]$$

2.1 Approximation vs Estimation Error

We can decompose the regret for a given function, while working with a said function class $\mathcal{F}$ as follows :

$$\mathcal{R}^l_{\mathcal{D}}[f] = \underbrace{(er^l_{\mathcal{D}}[f] - er^{l,\mathcal{F}}_{\mathcal{D}})}_{estimation} + \overbrace{(er^{l,\mathcal{F}}_{\mathcal{D}} - er^{l,*}_{\mathcal{D}})}^{approximaion}$$

The **approximation** error refers to that part of the error which can not be optimized by better estimation techniques, algorithmic or otherwise. This is because that is the minimum noise or error the given learnt function will have because of the hypothesis function class we have chosen. It can be seen that this does not depend on the algorithm, or the function that has been learnt and would be constant across all the functions in the said function class.

The **estimation** error refers to that part of the error which measures how well / poorly the given function performs with respect to the optimal function in the function class we are working with. This is something that we can minimize and is what has to be optimized by algorithmic techniques.

We see that this is closely related to the Bias Variance trade-off, since if you choose a very small function class, even though the estimation error might be small, approximation error might be very large since the function class will not have enough flexibility. On the other hand, a very large function class might have approximation error as very small, but due to increased flexibility it might be very hard to minimize the estimation error.

3 PAC Learning

Definition 13.5. A function class $\mathcal{F}$ is said to be agnostically Probably Approximately Correct (PAC) learnable with respect to a loss function l if there exists an algorithm

$$\mathcal{A} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}'$$

such that $\mathcal{A} : \mathcal{S} \to f_{\mathcal{S}}$

$$\mathbb{P}\left[er^l_{\mathcal{D}}[f_{\S}] - er^{l,\mathcal{F}}_{\mathcal{D}} > \epsilon\right] < \delta$$

where $n \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta})$

Definition 13.6. A function class $\mathcal{F}$ is said to be Probably Approximately Correct (PAC) learnable with respect to a loss function l if there exists an algorithm

$$\mathcal{A} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}'$$

such that $\mathcal{A} : \mathcal{S} \to f_{\mathcal{S}}$

$$\mathbb{P}\left[er^l_{\mathcal{D}}[f_{\S}] > \epsilon \right] < \delta$$

or rather $er^l_{\mathcal{D}}[f] < \epsilon$ where $n \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta})$

3.1 Agnostic PAC vs PAC

The key difference between the two definitions is that the first one talks about reaching the optimal that is theoretically possible for any given model or learning algorithm, while the second talks about minimizing the error. This makes the first one capture much more larger set of problems than can be captured in the second one. There are a lot of problems, where even humans do not reach consensus, and there is a lot of noise, which can not be covered in this class. But if we have a learning algorithm that is able to mimic exactly identical behaviour that humans show, i.e. assign probabilities to possible outputs that a set of humans would do in when chosen randomly for the experiment, then it gets captured in agnostic PAC learning, making it a much more powerful and practical class for learning experiments. Therefore we can conclude that

$$AgnosticPAC^l_{\mathcal{D}} \subset PAC^l_{\mathcal{D}}$$

3.2 PAC Learnnig and Uniform Convergence

From the structure of the definition it is easy to see some general results about PAC Learning.

Corollary 13.1. $\mathcal{F}$ is PAC Learnable if $\mathcal{F}$ exhibits Uniform Convergence

Proof. We can see from uniform convergence that if the function $g(S)$ is defined as

$$g(S) = \sup_{f \in \mathcal{F}} \{ er^l_{\mathcal{D}}[f] - er^l_{\mathcal{S}}[f] \}$$

$$\mathbb{P}\left[g(S) \geq \mathbb{E}\left[g(S) \right] + \epsilon \right] \leq \delta$$

Then with probability $1 - 2\delta$ we can show that

$$er^l_{\mathcal{D}}[f] - er^{l,\mathcal{F}}_{\mathcal{D}} \leq 2\epsilon$$

Therefore $\mathcal{F}$ is PAC learnable with respect to l. □

Now since we have the above corollary, it is natural to ask the converse of it. That is, is it the case that the every PAc Learnable function class exhibits uniform convergence. It can be shown under mild conditions that it is the case

Theorem 13.2. Under some mild conditions (to be specified later), $\mathcal{F}$ is PAC Learnable with respect to loss l
$\Rightarrow \mathcal{F}$ shows uniform convergence.

Proof. To be covered later □

4 Regret Transfer Bounds

Informally it has been stated before that **regret transfer** bounds is the study of how solve easy problems because we can not solve harder problems and still get results that are worthwhile. This basically means that since we can not always solve the exact problem of classification we try to solve it by using alternate loss functions. But that does not give us a guarantee that what we are doing is *consistent* with the actual aim, i.e. that solving a weaker problem with a different loss function gives us the predictions or results we required. In order to do this we have to ensure that the following :

For every alternate loss function we are using (call it $\tilde{l}$),

$$\mathcal{R}^{l}_{\mathcal{D}}[f] \leq \psi(\mathcal{R}^{\tilde{l}}_{\mathcal{D}})$$

This bound would be uniform, over all function classes and all functions in those function classes.

4.1 Posterior distributions

Posterior distributions are generally distirbutions of the form $\eta(X) \triangleq \mathbb{P}[Y = 1|X]$. In the noiseless case we would have the true posterior distribution to be $\eta(X) \in \{0, 1\}$, while in the noisy case $\eta(X) \in [0, 1]$.

Given a class conditional distribution $\mathbb{P}[X|Y = i]$ where $i \in \{-1, 1\}$ we can use the distribution to get a posterior distribution of the form $\eta(X)$ as follows :

$$\eta(X) = \mathbb{P}[Y = 1|X] = \frac{\mathbb{P}[X|Y = 1]\,\mathbb{P}[Y = 1]}{\mathbb{P}[X]}$$

$\mathbb{P}[X] = \sum_{i \in \mathcal{Y}} \mathbb{P}[X|Y = i]\,P(Y = i)$ From that we can obtain the following analysis :

$$er^{l}_{\mathcal{D}}[f] = \mathbb{E}[l(f(X), Y)] = \mathop{\mathbb{E}}_{X,Y}[\mathbb{I}\{f(X) \neq Y\}] = \mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{Y|X}[\mathbb{I}\{f(X) \neq Y\}\,|X]\right]$$

$$\Rightarrow er^{l}_{\mathcal{D}}[f] = \mathop{\mathbb{E}}_{X}\left[\mathbb{I}\{f(X) = 1\}\mathop{\mathbb{E}}_{Y|X}[\mathbb{I}\{Y = -1\}] + \mathbb{I}\{f(X) = -1\}\mathop{\mathbb{E}}_{Y|X}[\mathbb{I}\{Y = 1\}]\right]$$

$$\Rightarrow er^{l}_{\mathcal{D}}[f] = \mathop{\mathbb{E}}_{X}[\mathbb{I}\{f(X) = 1\}(1 - \eta(X)) + \mathbb{I}\{f(X) = -1\}\eta(X)]$$

$$er^{l}_{\mathcal{D}}[f] = \begin{cases} 1 - \eta(X) & f(X) = 1 \\ \eta(X) & f(X) = -1 \end{cases}$$

We can see that the optimal f^* can be characterized as follows :

$$f^* = \begin{cases} 1 & \eta(X) \geq 1/2 \\ -1 & otherwise \end{cases}$$

4.2 Plug In classifiers

The posterior probability mentioned in the previous section is a generally unknown quantity and can not be estimiated in most situations. But there are methods to approximate them. Such distributions, represented as $\hat{\eta}(X)$, can be plugged instead of $\eta(X)$ in order to get $\hat{f}$:

$$\hat{f} = \begin{cases} 1 & \hat{\eta}(X) \geq 1/2 \\ -1 & otherwise \end{cases}$$

These estimates are called class probability estimates since estimate the class conditional distributions.

Lemma 13.3. $\mathcal{R}_D^{0-1}[\hat{\eta}] \leq 2\mathop{\mathbb{E}}_X[|\hat{\eta}(X) - \eta(X)|]$

Proof.

$$er_{\mathcal{D}}^{0-1} \triangleq er_{\mathcal{D}}^{l^{0-1}} \Rightarrow er_{\mathcal{D}}^{0-1,*} = \min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{I}\{f(X) \neq Y\}]$$

From this we can derive bounds for the regret as follows :

$$\mathcal{R}_{\mathcal{D}}^{0-1} = er_{\mathcal{D}}^{l}[\hat{f}] - er_{\mathcal{D}}^{l}[f^*]$$

By symmetrization we can get the following :

$$\Rightarrow \mathcal{R}_{\mathcal{D}}^{0-1} = \mathop{\mathbb{E}}_X\left[\mathop{\mathbb{E}}_Y\left[l^{0-1}(\hat{f}(X), Y) - l^{0-1}(f^*(X), Y)\right]|X\right]$$

Then we can show pointwise,

$$\mathop{\mathbb{E}}_Y\left[l^{0-1}(\hat{f}(X), Y) - l^{0-1}(f^*(X), Y)\right] = \eta(X)(\mathbb{I}\left\{\hat{f}(X) = 0\right\}) + (1 - \eta(X))(\mathbb{I}\left\{\hat{f}(X) = 1\right\} - 1)$$

$$\mathop{\mathbb{E}}_Y\left[l^{0-1}(\hat{f}(X), Y) - l^{0-1}(f^*(X), Y)\right] \leq 2|\eta(X) - \hat{\eta}(X)|$$

which can be proved by a case by case analysis. □

Lemma 13.4.

$$er_{\mathcal{D}}^{0-1} \triangleq \min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{I}\{f(X) \neq Y\}] \geq \mathop{\mathbb{E}}_X[\min\{\eta(X), 1 - \eta(X)\}]$$

Proof. We prove the pointwise bound, and then claim that there would exist a Borel measurable function f that would satisfy the said bound at every point, and therefore acheive the limit in the theorem.

$$\min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{Y}}[f(X) \neq Y|X] = \min_{f \in \mathcal{F}} \eta\mathbb{I}\{f(X) \neq 1\} + (1 - \eta)\mathbb{I}\{f(X) \neq -1\} \geq \min\{\eta(X), 1 - \eta(X)\}$$

The second inequality holds because of the following claim :

Claim 13.5. $a + b \geq c, a \geq 0, b \geq 0$
$\forall \eta \in [0, 1], \eta a + (1 - \eta)b \geq c \min\{\eta, 1 - \eta\}$

Proof. $a \geq c - b \Rightarrow \eta a + (1 - \eta)b \geq c\eta + b(1 - 2\eta)$
This expression is $\geq c\eta$ if $1 - 2\eta < 0$. Similarly we get that $\eta a + (1 - \eta)b \geq c(1 - \eta) + a(2\eta - 1)$ which is larger than $c(1 - \eta)$ if $2\eta - 1 \leq 0$.

Since $1 - 2\eta < 0$ and $2\eta - 1 < 0$ are mutually exclusive events
we know that $\eta a + (1 - \eta)b \geq c \cdot \min\{\eta, 1 - \eta\}$ □

Using the above claim, and extending the pointwise claim to the uniform claim we complete the proof. □

4.3 Hinge Loss

We consider the hinge loss for classification $l(f(X), Y) = \max(0, 1 - Y \cdot f(X))$. Much like the $0-1$ classification loss we extend the pointwise limit to the hinge loss as follows :

$$\mathcal{C}(f, X) = \min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{Y} [l(f(X), Y)|X] = \min_{f \in \mathcal{F}} \eta(X) l(f(X), +1) + (1 - \eta(X)) l(f(X), -1)$$

$$\mathcal{C}(f, X) = \min_{f \in \mathcal{F}} \eta(X)[1 - f(X)]_+ + (1 - \eta(X))[1 + f(X)]_+ \geq 2 \min(\eta(X), 1 - \eta(X))$$

Extending it further we get that $er_{\mathcal{D}}^{h,*} = 2 \mathop{\mathbb{E}}_{X} [C(f, X)] \geq 2 \mathop{\mathbb{E}}_{X} [\min(\eta(X), 1 - \eta(X))]$

Theorem 13.6.

$$\mathcal{R}_{\mathcal{D}}^{0-1} \leq \mathcal{R}_{\mathcal{D}}^{h}$$

Proof. It suffices to analyse the following result

$$\eta(X)[1 - f(X)]_+ + (1 - \eta(X))[1 + f(X)]_+ - \eta(X)\mathbb{I}\{f(X) \neq 1\} - (1 - \eta(X))\mathbb{I}\{f(X) \neq -1\}$$

$$= \eta(X)([1 - f(X)]_+ - \mathbb{I}\{f(X) \neq 1\}) + (1 - \eta(X))([1 + f(X)]_+ - \mathbb{I}\{f(X) \neq -1\})$$

We can see that

$$[1 - f(X)]_+[1 + f(X)]_+ - (\mathbb{I}\{f(X) \neq -1\} + \mathbb{I}\{f(X) \neq 1\}) \geq 2 + f(X) - f(X) - 1 \geq 1$$

$$[1 - f(X)]_+ - \mathbb{I}\{f(X) \neq 1\} > 2 - 1 > 0$$

$$[1 + f(X)]_+ - \mathbb{I}\{f(X) \neq -1\} > 2 - 1 > 0$$

$$\Rightarrow \eta(X)([1-f(X)]_+ - \mathbb{I}\{f(X) \neq 1\}) + (1-\eta(X))([1+f(X)]_+ - \mathbb{I}\{f(X) \neq -1\}) \geq \min(\eta(X), 1-\eta(X))$$

$$\Rightarrow \mathcal{R}_{\mathcal{D}}^{0-1} \leq \mathcal{R}_{\mathcal{D}}^{h}$$

□