Indian Institute of Technology, Kanpur
Probability Theory, 18–19

Instructors: Purushottam Kar
Authors: Gurpreet Singh
Date: January 17, 2018

SCRIBE

2

Parametric Learning and Convergence

Parametric Learning

Before we understand parametric learning theory, it is best to know the standard notation we will be using in this text.

Notations

1. **Instance/Feature Space** $(\mathcal{X})$ — This is the space of the random variable that defines the data distribution.
2. **Output Space** $(\mathcal{Y})$ — This defines the space of the label or value mapped to each data point. There can be different Output Spaces depending on the problem
   - Labelling: $\{-1, 1\}, \mathbb{R}^p$
   - Permutation: S_n
   - Hierarchy / Tree
   - Alternate Representation: $\mathbb{R}^D \to \mathbb{R}^d$ (*e.g.* Dimensionality Reduction)
3. **Hypothesis/Model Space** $(\mathcal{F})$ — This defines the set of all functions $\{f : \mathcal{X} \to \mathcal{Y}\}$ belonging to a certain function class, for example, all linear functions or all neural networks. Generally equal to $\mathcal{Y}^{\mathcal{X}}$.
4. **Distribution** $(\mathcal{D})$ — In case of noisy setting or agnostic setting, this is the probability distribution over $\mathcal{X} \times \mathcal{Y}$, typically unknown.

   However, if the setting is realizable (non-agnostic), then the distribution is only over $\mathcal{X}$, and we map $f^* : x \mapsto y$.
5. **Training Sample** (S) — This is a tuple of n data points and their mapped label/parameter sampled from the distribution or randomly obtained. $S = \left(\left(x^1, y^1\right), \left(x^2, y^2\right) \ldots \left(x^n, y^n\right)\right) \in (\mathcal{X} \times \mathcal{Y})^n$

   If the setting is agnostic, then $S \overset{iid}{\sim} \mathcal{D}$
6. **Loss Function** (l) — Loss function defines the similarity / dissimilarity in the estimated label value and the actual / observed label value. We say $l : (f(\mathcal{X}), \mathcal{Y}) \to \mathbb{R}$. There are different types of loss functions, for example the squared loss function l^{sq} and the $0-1$ loss function l^{0-1}.
7. **Algorithm** (A) — We say that the algorithm maps the generated / sampled training sample to a function belonging in the hypothesis space *i.e.*

$$A : S \mapsto \hat{f}_s \in \mathcal{F}$$
$$A : \bigcup_{n=1}^{\infty} (\mathcal{X}, \mathcal{Y})^n \mapsto \mathcal{F}$$

8. **l-risk** $(\text{er}_D^l\left[\,f\,\right])$ — This is defined as the expected loss/error that is obtained for a function $f \in \mathcal{F}$ given a loss function $l : (f(\mathcal{X}), \mathcal{Y}) \to \mathbb{R}$.

$$\text{er}_D^l\left[\,f\,\right] \quad \triangleq \quad \underset{(x,y)\sim D}{\mathbb{E}}\left[\,l(f(x), y)\,\right]$$

9. **emperical risk** $(\mathrm{er}^l_D[\,f\,])$ — This is defined as the weighted loss/error on the training sample that is obtained for a function $f \in \mathcal{F}$ given a loss function $l : (f(\mathcal{X}), \mathcal{Y}) \to \mathbb{R}$

$$\mathrm{er}^l_S[\,f\,] \quad \triangleq \quad \frac{1}{n}\sum_{i=1}^{n} l(f(x^i), y^i)$$

Exercise 2.1. Show that $\mathrm{er}^l_D[\,f\,] = \mathbb{E}\left[\mathrm{er}^l_S[\,f\,]\right]$

Exercise 2.2. Show that $\mathrm{er}^{l^{0-1}}_D[\,f\,] = \mathrm{er}^{0-1}_S[\,f\,]$ where

$$l^{0-1}(\hat{y}, y) \quad = \quad 1 - \mathbb{I}[\,y = \hat{y}\,]$$

Exercise 2.3. Find out $\mathrm{er}^{l^{0-1}_\alpha}_D[\,f\,]$ in terms of the emerical risk where

$$l^{0-1}_\alpha \quad = \quad \begin{cases} \alpha & \hat{y} \neq y, y = 1 \\ 1-\alpha & \hat{y} \neq y, y = 0 \\ 0 & \hat{y} = y \end{cases}$$

Toy Binary Classification Example

Assume we have a binary classification problem, with a finite hypothesis space $\mathcal{F} = \{f_1, f_2 \ldots f_m\}$, where $\forall\, f \in \mathcal{F}, f : \mathcal{X} \to \mathcal{Y}$

We first sample a training sample, $S \overset{iid}{\sim} D^n$. Then, we find the best model in $\mathcal{F}$ *i.e.*

$$f^* = \underset{f\in\mathcal{F}}{\arg\min}\ \mathrm{er}^{0-1}_D[\,f\,]$$

One possible solution is to estimate the l-risk using emperical risk. Hence, we now minimize emperical risk (Emperical Risk Minimization)

$$f^* \quad \approx \quad \hat{f} \quad = \quad \underset{f\in\mathcal{F}}{\arg\min}\ \mathrm{er}_S[\,f\,]$$

Pointwise and Uniform Convergence

We continue with our analysis of the toy binary classification example. As stated earlier, we had defined the best function f^* and $\hat{f}$ as

$$\begin{aligned} f^* &= \underset{f\in\mathcal{F}}{\arg\min}\ \mathrm{er}^l_D[\,f\,] \\ \hat{f} &= \underset{f\in\mathcal{F}}{\arg\min}\ \mathrm{er}^l_S[\,f\,] \\ f^* &\approx \hat{f} \end{aligned}$$

Since our hypothesis space is finite, we can, without loss of generality, say that $f^* = f_1$. Then,

(i) we do not want $f_2, f_3 \ldots f_m$ to perform well on S *i.e.* have high(er) emperical risk or training error.

(ii) we do not want f_1 to perform poorly on S *i.e.* have low(er) emperical risk.

(iii) we try to ensure all f_i give faithful and honest performance on S *i.e.* $\forall\, f \in \mathcal{F}, \mathrm{er}^l_D[\,f\,] \approx \mathrm{er}^l_S[\,f\,]$

We say that we need f_1 to give good performance, however we need to define what is "good". We say that S is good with respect to a function $f \in \mathcal{F}$ ($S \in \text{good}_f(\epsilon)$) if

$$\left| \text{er}^l_D[\,f\,] - \text{er}^l_S[\,f\,] \right| \quad \leq \quad \epsilon$$

This is known as pointwise convergence. We define it more formally below.

Definition 2.1. We say that a hypothesis class, $\mathcal{F}$ has pointwise convergence for a given ϵ and a loss function l, if for all functions $f \in \mathcal{F}$, and a sample S of size n, the property

$$\left| \text{er}^l_D[\,f\,] - \text{er}^l_S[\,f\,] \right| \quad \leq \quad \epsilon$$

holds true.

Theorem 2.1. If $S \overset{iid}{\sim} D^n$, then $\forall f \in \mathcal{F}$,

$$\mathbb{P}\left[\, S \in \text{good}_f(\epsilon) \,\right] \quad \geq \quad 1 - 2\exp\left(\frac{-n\epsilon^2}{3}\right)$$

or equivalently

$$\mathbb{P}\left[\, \left| \text{er}^l_D[\,f\,] - \text{er}^l_S[\,f\,] \right| > \epsilon \,\right] \quad \leq \quad 2\exp\left(\frac{-n\epsilon^2}{3}\right)$$

Exercise 2.4. Prove Theorem 2.1

Hint: Take a Bernoulli random variable and use Chernoff's Bound

Pointwise convergence is a good property for individual functions, however, we want such a property to hold for all functions collectively. Hence, we define a much stronger convergence property, called uniform convergence.

In order to define Uniform Convergence, we must first define a desired situation for the defined hypothesis class.

$$\mathbb{P}\left[\text{er}^l_D[\,\hat{f}\,] > \text{er}^l_D[\,f^*\,] \right] \quad < \quad \delta$$

First, let us define the "goodness" of a sample in terms of uniform convergence. We say a sample S is good *i.e.* $S \in \text{good}(\epsilon)$ if for all functions $f \in \mathcal{F}$, $S \in \text{good}_f(\epsilon)$

Since we define $\hat{f} = \arg\min_{f \in \mathcal{F}} \text{er}^l_S[\,f\,]$, we can say

$$\text{er}^l_S[\,\hat{f}\,] \quad \leq \quad \text{er}^l_S[\,f^*\,]$$

Suppose $S \in \text{good}(\epsilon)$, then

$$\begin{aligned} \text{er}^l_D[\,\hat{f}\,] \quad &\leq \quad \text{er}^l_S[\,\hat{f}\,] + \epsilon \\ &\leq \quad \text{er}^l_S[\,f^*\,] + \epsilon \\ &\leq \quad \text{er}^l_D[\,f^*\,] + 2\epsilon \end{aligned}$$

Hence, we can say that if

$$\begin{aligned} S \in \text{good}(\epsilon) \quad &\Longrightarrow \quad \text{er}^l_D[\,\hat{f}\,] \leq \text{er}^l_D[\,f^*\,] + \frac{\epsilon}{2} \\ \therefore \quad \text{er}^l_D[\,\hat{f}\,] > \text{er}^l_D[\,f^*\,] + \frac{\epsilon}{2} \quad &\Longrightarrow \quad S \notin \text{good}(\epsilon) \end{aligned}$$

Here, we use an identity, the proof of which is left as an exercise.

$$\text{If } A \implies B \text{ then } \mathbb{P}\,[\,A\,] \leq \mathbb{P}\,[\,B\,]$$

Therefore, we can say

$$\mathbb{P}\left[\mathrm{er}^{l}_{D}[\,\hat{f}\,] > \mathrm{er}^{l}_{D}\,[\,f^{*}\,]\right] \quad \leq \quad \mathbb{P}\left[\,S \notin \text{ good}\left(\frac{\epsilon}{2}\right)\right]$$

We can further reduce the RHS of the above inequality in terms of the size of the sample S, ϵ and the size of the hypothesis set, and hence state the condition for uniform convergence.

Theorem 2.2. A hypothesis function, $\mathcal{F}$ is said to have the uniform convergence property if for some sample S of size n, and some ϵ

$$\mathbb{P}\left[\left|\mathrm{er}^{l}_{D}[\,\hat{f}\,] - \mathrm{er}^{l}_{D}\,[\,f^{*}\,]\right| > 2\epsilon\right] \quad \leq \quad \mathbb{P}\,[\,S \notin \text{ good}\,(\epsilon)\,] \quad \leq \quad 2\exp\left(\frac{-n\epsilon^{2}}{3}\right)\cdot m$$

Proof. From earlier, we know that if pointwise convergence holds, then uniform convergence also holds

$$\begin{aligned}
\forall\, f \in \mathcal{F}, S \in \text{ good}_{f}\,(\epsilon) \quad &\implies \quad S \in \text{ good}\,(\epsilon) \\
\therefore \quad \mathbb{P}\,[\,S \in \text{ good}\,(\epsilon)\,] \quad &\geq \quad \mathbb{P}\,[\,\forall\, f \in \mathcal{F}, S \in \text{ good}_{f}\,(\epsilon)\,] \\
\implies \mathbb{P}\,[\,S \notin \text{ good}\,(\epsilon)\,] \quad &\leq \quad \mathbb{P}\,[\,\exists\, f \in \mathcal{F}, S \notin \text{ good}_{f}\,(\epsilon)\,]
\end{aligned}$$

Since on the RHS, we just have a union, using basic probability theory inequalities, we can write

$$\begin{aligned}
\mathbb{P}\,[\,S \notin \text{ good}\,(\epsilon)\,] \quad &\leq \quad \sum_{k=1}^{m} \mathbb{P}\,[\,S \notin \text{ good}_{f_k}\,(\epsilon)\,] \\
&\leq \quad \sum_{k=1}^{m} 2\exp\left(\frac{-n\epsilon^{2}}{3}\right) \\
&= \quad 2\exp\left(\frac{-n\epsilon^{2}}{3}\right)\cdot m
\end{aligned}$$

Hence, we can say that the theorem is valid. □

<++>