STAT–S 782 | 15 — Introduction to learning - 2 | 12 October 2017

Lecturer: Prof. McDonald | Scribe: Kiran Kumar

Other types of learning

- Agnostic learning (model-free)
- Adversarial learning

In this scenario the target $C^*(or\ f^*)$ may not be in $C(or\mathfrak{F})$

The setup for learning involves the following components

- Sets x,y,u
- The distribution $\mathcal{P}$ on $x \times y$
- The function class $\mathfrak{F}(or\ C) = \{f : x \to u\}$
- loss $l : y \times y \to [0, 1]$

The process for learning generally involves the following steps

1. Get sample iid $Z^n = (z_1, \ldots, z_n)$; $\quad z_i \overset{iid}{\sim} p \in \mathcal{P}$
2. Let Algorithm $\mathcal{A} = \{A_n\}_{n=1}^{\infty}, \quad A_n : Z^n \to \mathfrak{F}$
3. Form a hypothesis $\widehat{f}_n = A_n(Z^n) = A_n\big((x_1, y_1), \ldots, (x_n, y_n)\big)$
4. Find expected loss of the hypothesis $L_p(\widehat{f}_n) = \mathbb{E}\Big[l(y^{n+1}, \widehat{f}(x^{n+1}) \mid Z^n\Big]$
5. Then the "best prediction" is given by $L_p^*(\mathfrak{F}) = \inf_{f \in \mathfrak{F}} L_p(f); \quad 0 \leq L_p^*(\mathfrak{F} \leq L_p(\widehat{f}_n \leq 1$

We can then define risk as follow

$$r_A(n, \epsilon) = \sup_{p \in \mathcal{P}} p^n(Z^n : L_p(\widehat{f}_n \geq L_p^*(\mathfrak{F} + \epsilon)$$

This is the worst loss over data set.

Example: Noisy Classification

We have $\mathcal{X}$, $p_x \in \mathcal{P}_{\mathcal{X}}$, C^*, C

- $X^n = (x_1, , x_n) \overset{iid}{\sim} p_x$
- The classification happens with a probability, $y_i = \begin{cases} I(x_i \in C^*), & w.p. \quad 1-\eta \\ 1 - I(x_i \in C^*), & w.p. \quad \eta \end{cases}$
- $\eta < \frac{1}{2}$ is the noise rate
- $y = u = \{0,1\}, \quad \mathfrak{F} = \{I_c : c \in \mathcal{C}\}$
- A loss function $l(y,u) = |y-u|^2$
- Set of probabilites $\{p_{x,c} : p_x \in \mathcal{P}_x, c \in \mathcal{C}\}$

$$\begin{aligned} p_{y \mid x,c}(1 \mid X = x, c) &= (1-\eta)I(x \in c) + \eta I(x \in c^{\complement}) \\ &= (1-\eta)I(x \in c) + \eta(1 - I(x \in c)) \end{aligned}$$

$$\begin{aligned} p_{y \mid x,c}(0 \mid X = x, c) &= 1 - P_{y \mid x,c}(1 \mid X = x, c) \\ &= \eta I(x \in c) + (1-\eta)(1 - I(x \in c)) \end{aligned}$$

So, for any $A \subseteq x$

$$\begin{aligned} p_{x,c}(A \times 1) &= \int_A p_{y \mid x,c}(1 \mid X = x) \quad p_x(dx) \\ &= \int_A (1-\eta)I(x \in C) + \eta(1 - I(x \in c)) \quad p_x(dx) \\ \ldots &= \eta p_x(A) + (1-2\eta)p_x(A \cap c) \end{aligned}$$

Similarly,

$$p_{x,c}(A \times 0) = (1-\eta)p_x(A) - (1-2\eta)p_x(A \cap c)$$

The loss of some hypothesis c' for any c can be written as

$$
\begin{aligned}
L_{p_{x,c}}(I_{c'}) &= \int_{x\times\{0,1\}} |y - I(x \in c')|^2 \quad p_{x,c}(dx\ dy)\\
&= \int_X |0 - I(x \in c')|^2 \quad p_x(dx \times \{0\}) + \int_X |1 - I(x \in c')|^2 \quad p_x(dx \times \{1\})\\
&= \int_X I(x \in c') \quad p_x(dx \times \{0\}) + \int_X I(x \in (c')^{\complement}) \quad p_x(dx \times \{1\})\\
&= p_{x,c}(c' \times \{0\}) + p_{x,c}((c')^{\complement} \times \{1\})\\
&= (1 - eta)p_x(c') + (1 - 2\eta)p_x(c' \cap c) + \eta p_x((c')^{\complement}) + (1 - 2\eta)p_x((c')^{\complement} \cap c)\\
&= \ldots = \ldots = \ldots\\
&= \eta + (1 - 2\eta)L_{p_x}(c', c)\\
&\text{The loss can be approximated using} \quad p_{x,c}(c\Delta c')
\end{aligned}
$$

The best hypothesis can then be given by

$$
\begin{aligned}
L^*_{p\times c}(c) &= \inf_{c'\in c} L_{p\times c}(c')\\
&= \eta + (1 - 2\eta) \inf_{c'\in c} p_{x,c}(c\Delta c')\\
&= \eta \quad \text{if (c'=c)}
\end{aligned}
$$

Thus, an infimum is achieved at $c' = C$.

$$
\begin{aligned}
L_{p\times c}(c') &\geq L^*_{p\times c} + \epsilon\\
p_{x,c}(c\Delta c') &\geq \frac{\epsilon}{1 - 2\eta}
\end{aligned}
$$

A noisy classification to some accuracy ϵ is like noise free classification to accuracy $\frac{\epsilon}{1-2\eta}$. Since $(1 - 2\eta)$ is very small so the bound is larger.

Noisy classification is more difficult than noiseless classification.