Bayes optimal rule and No Free Lunch

Sivan Sabato
The Department of Computer Science
Ben-Gurion University of the Negev

March 16, 2016

Some notation conventions: For an integer n, $[n] := \{1, \ldots, n\}$. The function $\mathbb{I}[P]$ is 1 if P holds, and 0 otherwise.

1 Supervised learning for binary classification

Let the example space be $\mathcal{X}$, and assume labels are either 1 or 0. This is our label space $\mathcal{Y} = \{0, 1\}$. We assume a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. In these notes we will usually assume for convenience that $\mathcal{X}, \mathcal{Y}$ are discrete.
A passive learning algorithm $\mathcal{A}$ gets a sample $S \sim \mathcal{D}^m$ as input, and outputs a function from $\mathcal{X}$ to $\mathcal{Y}$. The algorithm $\mathcal{A}$ is simply a function: $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{S}_m \to \mathcal{Y}^{\mathcal{X}}$. We will usually assume for convenience that $\mathcal{A}$ is deterministic. Denote the output of $\mathcal{A}$ given sample S by $\mathcal{A}[S]$. The goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$, such that the error of f on $\mathcal{D}$ is low. This error is defined as

$$\mathrm{err}(f, \mathcal{D}) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}[f(X) \neq Y].$$

2 The Bayes optimal rule

We want to minimize $\mathrm{err}(f, \mathcal{D})$. We have

$$\begin{aligned}\mathrm{err}(f, \mathcal{D}) = \sum_{x\in\mathcal{X}} \mathbb{P}[X = x \wedge Y \neq f(x)] = \\ \sum_{x\in\mathcal{X}} \mathbb{P}[X = x]\mathbb{P}[Y \neq f(x) \mid X = x] = \\ \sum_{x\in\mathcal{X}} \mathbb{P}[X = x](1 - \mathbb{P}[Y = f(x) \mid X = x]).\end{aligned}$$

Hence, the optimal solution is f_{Bayes}, defined as:

$$f_{\text{Bayes}}(x) := \operatorname*{argmax}_{y\in\{0,1\}} \mathbb{P}_{\mathcal{D}}[Y = y \mid X = x].$$

f_{Bayes} is called the *Bayes optimal* predictor.

3 No Free Lunch

The *No Free Lunch* theorem shows that it's impossible to have a learning algorithm that gets a low error for every possible distribution using a fixed sample size (assuming that the sample comes from the same distribution).

Theorem 3.1 (No free lunch)**.** *Let $m \leq |\mathcal{X}|/2$. Let $\mathcal{Y} = \{0, 1\}$.*
For passive learning algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ such that

1. *There is some $f \in \mathcal{Y}^{\mathcal{X}}$ with $\mathrm{err}(f, \mathcal{D}) = 0$, but*
2. $\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D})] \geq 1/4$.

Proof. Let $C \subseteq \mathcal{X}$ be some subset such that $|C| = 2m$. Then $|\mathcal{Y}^C| = 2^{2m} := T$. Denote the functions in $\mathcal{Y}^C$ by $f_1, \ldots, f_T$. For $i \in [T]$, Let $\mathcal{D}_i$ be the following distribution over $C \times \{0, 1\}$:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_i}[X = x \wedge Y = y] = \begin{cases} \frac{1}{2m} & x \in C, y = f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

Then $\mathrm{err}(f_i, \mathcal{D}_i) = 0$. Let $\mathcal{A}$ be some learning algorithm. We will show that

$$\exists i, \mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D})] \geq 1/4.$$

Define $\mathcal{D}_X$ over $\mathcal{X}$:

$$\mathbb{P}_{X\sim\mathcal{D}_X}[X = x] = \begin{cases} \frac{1}{2m} & x \in C \\ 0 & \text{otherwise} \end{cases}$$

Look at drawing a sample from $\mathcal{D}_i$: it is equivalent to drawing only the x's from $\mathcal{D}_X$, and then setting $y = f_i(x)$. Denote: For $S_X = (x_1, \ldots, x_m)$,

$$S_X^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))).$$

So we get

$$\mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D}_i)] = \mathbb{E}_{S_X\sim\mathcal{D}_X^m}[\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i)].$$

Now

$$\begin{aligned} \max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D}_i)] &\geq \frac{1}{T}\sum_{i\in T} \mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D}_i)] \\ &= \frac{1}{T}\sum_{i\in T} \mathbb{E}_{S_X\sim\mathcal{D}_X^m}[\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i)] \\ &= \mathbb{E}_{S_X\sim\mathcal{D}_X^m}[\frac{1}{T}\sum_{i\in T} \mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i)] \\ &\geq \min_{S_X\in C^m} \frac{1}{T}\sum_{i\in T} \mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i). \end{aligned} \tag{1}$$

So, if we show that $\mathcal{A}$ is pretty bad even for the best S_X, it proves that $\mathcal{A}$ is pretty bad for some $\mathcal{D}_i$. Choose some $S_X \in C^m$. The intuition is that since $\mathcal{A}$ only observes the examples in S_X, it can't get the other examples right for all distributions $\mathcal{D}_i$.

$$\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i) \geq \frac{1}{2m} \sum_{x \in C \setminus S_X} \mathbb{I}[\mathcal{A}[S_X^i](x) \neq f_i(x)].$$

Let $j(i)$ such that

$$f_{j(i)}(x) = \begin{cases} f_i(x) & x \in S_X \\ -f_i(x) & x \notin S_X \end{cases}$$

Then

$$\begin{aligned} &\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i) + \mathrm{err}(\mathcal{A}[S_X^{j(i)}], \mathcal{D}_{j(i)}) \\ &\geq \frac{1}{2m} \sum_{x \in C \setminus S_X} (\mathbb{I}[\mathcal{A}[S_X^i](x) \neq f_i(x)] + \mathbb{I}[\mathcal{A}[S_X^{j(i)}](x) \neq f_{j(i)}(x)]). \end{aligned}$$

Crucial point: $\mathcal{A}[S_X^{j(i)}](x) = \mathcal{A}[S_X^i](x)$ because $S_X^i = S_X^{j(i)}$. So:

$$\begin{aligned} &\mathbb{I}[\mathcal{A}[S_X^i](x) \neq f_i(x)] + \mathbb{I}[\mathcal{A}[S_X^{j(i)}](x) \neq f_{j(i)}(x)] = \\ &\mathbb{I}[\mathcal{A}[S_X^i](x) \neq f_i(x)] + \mathbb{I}[\mathcal{A}[S_X^i](x) \neq -f_i(x)] = 1. \end{aligned}$$

Therefore

$$\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i) + \mathrm{err}(\mathcal{A}_{S_X^{j(i)}}, \mathcal{D}_{j(i)}) \geq \frac{|\{x \in C \setminus S_X\}|}{2m} \geq 1/2.$$

Therefore

$$\frac{1}{T} \sum_{i \in T} \mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i) = \frac{1}{T} \sum_{i \in T/2} (\mathrm{err}(\mathcal{A}[S_X^i], \mathcal{D}_i) + \mathrm{err}(\mathcal{A}_{S_X^{j(i)}}, \mathcal{D}_{j(i)})) \geq \frac{1}{T} \cdot T/2 \cdot 1/2 = 1/4.$$

Combining this with Eq. (1), we get that

$$\exists i \in [T], \mathbb{E}_{S \sim \mathcal{D}_i^m}[\mathrm{err}(\mathcal{A}[S], \mathcal{D}_i)] \geq 1/4.$$

□