Understanding the Relationship between Training Error, Generalization Error, and the Information Bottleneck Method

November 30, 2017

The goal of this document is to find an analytical connection between the classic error and the Information Bottleneck Method. Here is a cleaned version of my scrap work so far. First some definitions:

- X = input features
- Y = response variable
- $\tilde{X}$ = compressed (Tishby doesn't treat this R.V. as the predictions, we must clarify the distinction)
- $f_\theta : X \to \hat{Y}$ = map function of input features to predictions ($\hat{Y}$)

1. Training error $\implies$

   The training error is defined as the following:

   $$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(f_\theta(x^{(i)}) \neq y^{(i)})$$

   Let us define the empirical joint distribution as the following:

   $$\hat{P}(x, y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(x^{(i)} = x)\mathbb{1}(y^{(i)} = y)$$

   We can define the parametrized joint distribution as the following:

   $$P_\theta(x, y) = P(x, y)$$

   As shown in homework 3 we know the following:

   $$argmin_\theta D(\hat{P}||P_\theta) = argmax_\theta \sum_{i=1}^{m} log(P_\theta(x^{(i)}, y^{(i)}))$$

   (still need to investigate)

2. Generalization error $\implies$

The generalization error ($I[f_\theta]$) is defined as the following where $V(x, y)$ is some loss function.

$$I[f_\theta] = \sum_{x,y} V(f_\theta(x), y)p(x, y)$$

Thus, we can define the generalization accuracy as the following:

$$1 - \sum_{x,y} V(f_\theta(x), y)p(x, y)$$

Let us use the **0-1 loss function** $V(x, y) = \mathbb{1}(f_\theta(x) \neq y)$. Thus, the general accuracy is:

$$\begin{aligned}
1 - I[f_\theta] &= 1 - \sum_{x,y} \mathbb{1}(f_\theta(x) \neq y)p(x, y) \\
&= 1 - \sum_{x,y} (1 - \mathbb{1}(f_\theta(x) = y))p(x, y) \\
&= 1 - \sum_{x,y} p(x, y) + \sum_{x,y} \mathbb{1}(f_\theta(x) = y)p(x, y) \\
&= 0 + \sum_{x,y} \mathbb{1}(f_\theta(x) = y)p(x, y) \\
&= \sum_{x,y} p(x, f_\theta(x)) \\
&= |Y| \sum_{x} p(x, f_\theta(x))
\end{aligned}$$

Thus, the θ that maximizes the general accuracy for the 0-1 loss is defined as follows:

$$\begin{aligned}
\operatorname*{argmax}_\theta 1 - I[f_\theta] &= \operatorname*{argmax}_\theta |Y| \sum_x p(x, f_\theta(x)) \\
&= \boxed{\operatorname*{argmax}_\theta \sum_x log(p(x, f_\theta(x)))}
\end{aligned}$$

Now let us use the **cross-entropy loss function** $V(x, y) = H(f_\theta(x), y) = H(f_\theta(x)) + D(f_\theta(x)||y)$. We can express the generalization accuracy as the following:

(still need to investigate)

3. Information Bottleneck Method $\implies$

Notice the following equality of the Information Bottleneck Method:

$$\underset{\theta}{\operatorname{argmax}}\, I(\tilde{X}, Y) - I(\tilde{X}, X) = \underset{\theta}{\operatorname{argmax}}\, I(\tilde{X}, Y) - I(X, Y) - I(\tilde{X}, X)$$

We can simplify this sum of mutual informations as follows:

$$\begin{aligned} I(\tilde{X}, Y) - I(X, Y) - I(\tilde{X}, X) &= H(\tilde{X}) + H(Y) - H(\tilde{X}, Y) \\ &\quad - H(X) - H(Y) + H(X, Y) \\ &\quad - H(\tilde{X}) - H(X) + H(\tilde{X}, X) \\ \\ &= H(X, Y) + H(\tilde{X}, X) - 2H(X) - H(\tilde{X}, Y) \qquad * \\ &= H(Y|X) + H(\tilde{X}|X) - H(\tilde{X}, Y|X) \\ &= I(Y, \tilde{X}|X) \end{aligned}$$

* Need to check this step.

Thus, we can express the θ that optimizes the information bottleneck as the following:

$$\boxed{\underset{\theta}{\operatorname{argmax}}\, I(Y, \tilde{X}|X)}$$