CSCI 5525 Machine Learning Fall 2019

Lecture 12: Learning Theory (Part 1)

March 2nd 2020

Lecturer: Steven Wu *Scribe: Steven Wu*

We have now seen several methods for supervised machine learning, and we have mostly talked about how to solve the empirical risk minimization problem–that is given a finite set of data, find a model in some class to minimize some loss function. We haven't talked about how these learned models will perform on future instances or the underlying distribution. We will now study several theoretical tools for analyzing the *generalization error* of a learning algorithm.

1 Bias-Variance Tradeoff

Let us revisit the general supervised learning setting. Suppose the learning algorithm given a data set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from some distribution P over $\mathcal{X} \times \mathcal{Y}$. For this part, we consider a regression setting with $y \in \mathcal{R}$ and square loss. A couple definitions are in order:

Expected Label for $x \in \mathcal{X}$.

$$\bar{y}(x) = E_P\left[Y \mid X = x\right] = \int\limits_y y\, P(y \mid x)\, dy$$

If you happen to know the distribution P perfectly, to minimize the square loss given an example x is to predict $\bar{y}(x)$. This is called the *Bayes optimal predictor*.

Now consider a learning algorithm $\mathcal{A}$ that takes the data set D drawn i.i.d. from P as input and outputs a predictor, denoted $f_D = \mathcal{A}(D)$. We can define:

Expected Test Error for f_D.

$$E_{(x,y)\sim P}\left[(f_D(x) - y)^2\right] = \int\limits_x \int\limits_y \left(f_D(x) - y\right)^2 P(x, y)\, dy\, dx$$

Now we should also take into account that the data set D is also drawn randomly from P, and so there is randomness in the predictor produced by $\mathcal{A}$ as well.

Expected Classifier for a given algorithm $\mathcal{A}$.

$$\bar{f} = E_{D\sim P^n}\left[f_D\right] = \int\limits_D f_D P(D)\, dD$$

We can also use the fact that f_D is a random variable to compute the expected test error only given $\mathcal{A}$, taking the expectation also over D.

Expected Test Error given $\mathcal{A}$.

$$E_{\substack{(x,y)\sim P \\ D\sim P^n}}\left[\left(f_D(x)-y\right)^2\right]$$

To be clear, D is our training points and the (x, y) pairs are the test points.

This quantity measures the expected accuracy of the underlying learning algorithm $\mathcal{A}$, which is the quantity we would like analyze. Let us decompose this expected test error.

Decomposition of Expected Test Error

$$\begin{aligned}
&E_{x,y,D}\left[\left[f_D(x)-y\right]^2\right] && (1)\\
&= E_{x,y,D}\left[\left[\left(f_D(x)-\bar{f}(x)\right)+\left(\bar{f}(x)-y\right)\right]^2\right] && \\
&= E_{x,D}\left[\left(f_D(x)-\bar{f}(x)\right)^2\right] + 2\ E_{x,y,D}\left[\left(f_D(x)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right] + E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] && (2)
\end{aligned}$$

You will show that the middle term of the above equation can be shown to be 0.

Now let us deal with the variance and another term:

$$E_{x,y,D}\left[\left(f_D(x)-y\right)^2\right] = \underbrace{E_{x,D}\left[\left(f_D(x)-\bar{f}(x)\right)^2\right]}_{\text{Variance}} + E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] \tag{3}$$

We can break down the second term in the above equation as follows:

$$\begin{aligned}
E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] &= E_{x,y}\left[\left[\left(\bar{f}(x)-\bar{y}(x)\right)+\left(\bar{y}(x)-y\right)\right]^2\right] && (4)\\
&= \underbrace{E_{x,y}\left[\left(\bar{y}(x)-y\right)^2\right]}_{\text{Noise}} + \underbrace{E_{x}\left[\left(\bar{f}(x)-\bar{y}(x)\right)^2\right]}_{\text{Bias}^2} + 2\ E_{x,y}\left[\left(\bar{f}(x)-\bar{y}(x)\right)\left(\bar{y}(x)-y\right)\right] && (5)
\end{aligned}$$

You will also show that the third term in the equation above can be shown to be 0.

This gives us the decomposition of expected test error as follows

$$\underbrace{E_{x,y,D}\left[\left(f_D(x)-y\right)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{x,D}\left[\left(f_D(x)-\bar{f}(x)\right)^2\right]}_{\text{Variance}} + \underbrace{E_{x,y}\left[\left(\bar{y}(x)-y\right)^2\right]}_{\text{Noise}} + \underbrace{E_{x}\left[\left(\bar{f}(x)-\bar{y}(x)\right)^2\right]}_{\text{Bias}^2}$$