Bias Variance Tradeoff

As usual, we are given a dataset $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$ drawn iid from some distribution $P(X, Y)$.

We will assume a regression setting, i.e $y \in \mathbb{R}$.

→ Note: Proof in regression setting is easier that is why we will be using it.

Let us consider that for any given input x there might not exists a unique label y.

For example, x describes features of a house (eg bedrooms, carpet area, etc) and the label y its price. Two different house with identical description might get sold at a different price.

Hence, for any given feature vector x, there is distribution over possible labels.

expected label (given $x \in \mathbb{R}^d$):

$$\bar{y}(x) = E_{y|x}[Y] = \int_y y\, P(y|x)\, dy.$$

Alright, so we draw our training dataset D, consisting n inputs, i.i.d from the distribution P.

As second step we typically call some machine learning algorithm A on this data set to learn a hypothesis (aka classifier).

$$h_D = A(D)$$

For a given ~~model~~ h_D, learned on data set D with algorithm A, we can compute the generalization error (as measured in squared loss) as follows:

Expected Test Error (given h_D):

$$E_{(x,y)\sim P}\left[(h_D(x) - y)^2\right] = \iint (h_D(x) - y)^2 P(x,y)\, \partial y\, \partial x$$

Note - Other loss functions can also be used, we are using squared loss for the ease of proof.

Expected Classifier (given A)

$$\bar{h} = E_{D \sim P^n}[h_D] = \int_D h_D P(D) \partial D.$$

Here $P(D)$ is probability of drawing D from P^n.

Here, $\bar{h}$ is a weighted average over functions.

Expected test error (given A) -

$$E_{\substack{(x,y) \sim P \\ D \sim P^n}}\left[\left(h_D(x) - y\right)^2\right] = \int_D \int_x \int_y (h_D(x) - y)^2 \, P(D) \, \partial x \, \partial y \, \partial D.$$

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

This expression is of interest to us because it evaluates the quality of a machine learning algorithm A with respect to a data distribution $P(X, Y)$.

Now, let's decompose it further.

$$E_{x,y,D}\left[(h_D(x)-y)^2\right] = E_{x,y,D}\left[\left((h_D(x)-\bar{h}(x))+(\bar{h}(x)-y)\right)^2\right]$$

$$= E_{x,D}\left[(\bar{h}_D(x)-\bar{h}(x))^2\right] + 2E_{x,y,D}\left[(h_D(x)-\bar{h}(x)).(\bar{h}(x)-y)\right]$$

$$+ E_{x,y}\left[(\bar{h}(x)-y)^2\right]$$

The middle term of the above equation is 0.

$$E_{x,y,D}\left[(h_D(x)-\bar{h}(x))(\bar{h}(x)-y)\right]$$

$$= E_{x,y}\left[E_D[h_D(x)-\bar{h}(x)](\bar{h}(x)-y)\right]$$

$$= E_{x,y}\left[(E_D[h_D(x)]-\bar{h}(x))(\bar{h}(x)-y)\right]$$

$$= E_{x,y}\left[(\bar{h}(x)-\bar{h}(x))(\bar{h}(x)-y)\right]$$

$$= E_{x,y}[0] = 0.$$

Hence we are left with

$$E_{x,y,D}\left[(h_D(x)-y)^2\right] = \underbrace{E_{x,D}\left[(\bar{h}_D(x)-\bar{h}(x))^2\right]}_{\text{Variance.}} + E_{x,y}\left[(\bar{h}(x)-y)^2\right]$$

Scanned with CamScanner

Expanding the term,

$$E_{x,y,D}[(\bar{h}(x)-y)^2] = E_{x,y}[((\bar{h}(x)-\bar{y}(x)) + (\bar{y}(x)-y))^2]$$

$$= \underbrace{E_{x,y}[(\bar{y}(x)-y)^2]}_{\text{Noise}} + \underbrace{E_x[(\bar{h}(x)-\bar{y}(x))^2]}_{\text{Bias}^2}$$

$$+ 2\underbrace{E_{x,y}[(\bar{h}(x)-\bar{y}(x))(\bar{y}(x)-y)]}_{=0}.$$

Proving that third term will be zero.

$$E_{x,y}[(\bar{h}(x)-\bar{y}(x))(\bar{y}(x)-y)]$$

$$= E_x[E_{y|x}[\bar{y}(x)-y](\bar{h}(x)-\bar{y}(x))]$$

$$= E_x[E_{y|x}[\bar{y}(x)-y](\bar{h}(x)-\bar{y}(x))]$$

$$= E_x[(\bar{y}(x)-E_{y|x}[y])(\bar{h}(x)-\bar{y}(x))]$$

$$= E_x[(\bar{y}(x)-\bar{y}(x))(\bar{h}(x)-\bar{y}(x))$$

$$= E_x[0]$$

$$= 0.$$

Scanned with CamScanner

Finally,

$$E_{x,y,D}\left[(h_D(x) - y)^2\right]$$

$$= \underbrace{E_{x,D}\left[(h_D(x) - \bar{h}(x))^2\right]}_{\text{Variance}} + \underbrace{E_{x,y}\left[(\bar{y}(x) - y)^2\right]}_{\text{Noise}}$$

$$+ \underbrace{E_x\left[(\bar{h}(x) - \bar{y}(x))^2\right]}_{\text{Bias}^2}$$

Scanned with CamScanner