Bias Variance Analysis

Ref: Domingos, Pedro. "A unified bias-variance decomposition." *Proceedings of 17th international conference on machine learning*. Morgan Kaufmann, 2000.

For Regression

Error

Error of a learning method can be decomposed into bias and variance.

Bias: Deviation from the true value (inherent weakness of the learning method)

Variance: Variance of the prediction over different training sets (variation in the prediction because the training set is varied).

Error

Error depends on the learning method (f) and on the training set (D_i).

For the given test example X, the prediction is $f(X)$.

Actually since this depends on the training set D_i, we can write $f(X;\ D_i)$

Training sets

Let $D = \{D_1, \ldots, D_i, \ldots, D_{10}\}$, i.e., we are having 10 different training sets drawn from the same distribution.

Size of each D_i is the same. Let us say $|D_i| = n$.

Each training set will be like

$$D_i = \{(X_1, t_1), \ldots, (X_n, t_n)\}$$

Note, t is the target and $y = f(X; D_i)$ is the prediction

Mean prediction for x is called y_m

- $y_m = Mean\,\{f(X; D_1), f(X; D_2), \ldots, f(X;\ D_{10})\}$
- This is nothing but average prediction over the training sets.

Square Loss

$$L(y_i, y_j) = (y_i - y_j)^2$$

In regression we usually use the square loss.

Bias

- On average, deviation from the true prediction.
- For x, bias in the prediction is, $B(x) = L(y, y_m) = (y - y_m)^2$

$$= \left(y - \frac{f(x; D_1) + \cdots + f(x; D_{10})}{10}\right)^2$$

Variance

- Variance in the prediction
- For x, variance in the prediction is, $V(x) = \frac{L(f(x; D_1), y_m) + \cdots + L(f(x; D_{10}), y_m)}{10}$

Note, for a single example (x,y) these are bias and variance (across the training sets)

Bias and Variance

- Bias and variance has to be found by averaging over the entire feature-space.
- Bias $= E_X[B(X)]$
- Variance $= E_X[V(X)]$

- In practice, we take average over the Test Set.

Test Set

- Let D_S be the test set.
- Let $D_s = \{(X_1, t_1), \ldots, (X_s, t_s)\}$
- Let $|D_s| = s$

- Bias $= \frac{1}{s}\sum_{k=1}^{s} B(X_k)$
- Variance $= \frac{1}{s}\sum_{k=1}^{s} V(X_k)$

- Note, this is the average Bias and Variance over the Test Set.
- The summation is over all Test Examples.

Distribution from which the training set is drawn

- We assume that the target which captures the correct relationship between X and t, is $t = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d$
- Fix the values $\theta_0, \theta_1, \ldots, \theta_d$
- Generate X from the given distribution (Say Gaussian with given parameters)
- Calculate t, the true prediction for X.
- Generate noise ϵ from the Gaussian distribution $N(0,1)$.
- Update $t := t + \epsilon$
- Then the training example generated is (X, t)
- Generate n examples to create the training set D_i. One has to generate D_i for $i = 1, \ldots, 10$. Each of the D_i is generated independently.

Lab Exercise

- You need to generate 10 different training sets from the given source distribution, each of size n.
- You need to generate the test set of size s from the same distribution, but independent from the training set.
- Let us fix s to 100.
- But n can be varied from 100 to 1000. For simplicity let n take values 100, 200, …,1000.
- Find Bias and Variance for each of the n value.
- You can plot n vs Bias and n vs Variance.

Below are given at the start of the Lab exercise

- The learning method ie., the classifier
- The distributions
- You can use libraries (tool-box given) to generate data.

Note, For each test example, you need to find its mean prediction which is y_m which can change for each test example.

Formal Derivation

Formally (we introduce t …)

- Let D be the set of training sets (each of same size)
- Let t be the target (true value) for the given x.
- Note, t is a random variable. It depends on x. But is independent of any training set.
- y is the prediction which depends on the training set, hence we write $y = f(x, D_i)$ where D_i is the training set. y depends on D_i but is independent of t.

For the given x, let t be the target, and D be the set of training sets

- The expected loss incurred for the prediction of x is

$$E_{D,t}[L(t,y)] = E_{D,t}[(t-y)^2]$$
$$= E_{D,t}(t - E_t t + E_t t - y)^2$$
$$= E_t(t - E_t t)^2 + E_{D,t}(E_t t - y)^2$$
$$= E_t(t - E_t t)^2 + E_{D,t}(E_t t - E_D y + E_D y - y)^2$$
$$= E_t\,(t - E_t t)^2 + (E_t t - E_D y)^2 + E_D(E_D y - y)^2$$
$$= N(x) + B(x) + V(x)$$

Here $N(x)$ is the noise, $B(x)$ is the Bias and $V(x)$ is the variance for x.

For Classification

0-1 Loss

$$L(y_i, y_j) = \begin{cases} 0, \text{ if } y_i = y_j \\ 1, \text{otherwise} \end{cases}$$

In classification we usually use 0-1 Loss

Main prediction for X is called y_m

- $y_m = Mode\,\{f(X; D_1), f(X; D_2), \ldots, f(X, D_{10})\}$
- This is nothing but majority (most frequent) prediction of f over the training sets.
  - In case of a Tie we break it randomly

The Bayes prediction for X

- y^* is the Bayes prediction for X
- Since we know the distributions, this can be found from the Bayes Classifier

Bias

- Deviation from the Bayes classifier.
- For X, bias in the prediction is, $B(X) = L(y_m, y^*)$

Variance

- Variance in the prediction
- For X, variance in the prediction is, $V(X) = \frac{L(f(X; D_1), y_m) + \cdots + L(f(X; D_{10}), y_m)}{10}$

Note, y_m , y^* are the main prediction and the Bayes prediction for the x, respectively.

Note

- We have taken 10 different training sets.
- In general, one may consider different number of training sets to estimate Bias and Variance.

For formal derivations, refer to

- **Ref:** Domingos, Pedro. "A unified bias-variance decomposition." *Proceedings of 17th international conference on machine learning*. Morgan Kaufmann, 2000.