Bias Variance Analysis for Classification

For the Lab Exercise

Error

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

For the Bayes Classifier, once the distributions are known we know its error.

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 and the test set

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 $|D_i| = n$.

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

Note, $y = f(X; D_i)$ is the prediction for X while using the training set D_i

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

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.

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 $|D_s| = s$

We are using the notation t for the target and y for the prediction.

- 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.

In this Lab Exercise

- You need to generate 10 different training sets from the given source distributions, 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

- The learning method ie., the classifier
- The distributions (which includes apriori probabilities and class-conditional densities)
- You can use libraries (tool-box given) to generate data.

Note, For each test example, you need to find its main prediction which is y_m and y^* which is the Bayes prediction. For each test example, these predictions, changes (in general).