DAML & NLP.
NIKHILESH.M

Variance and Bias:

* The error induced because of estimating the population is known as bias.

$d \rightarrow$ estimate

$\theta \rightarrow$ parameter

$(d - \theta)$ is error

* d is a random variable so Expected value of d is $E(d)$.

$E[d] - \theta$ is defined as bias

The difference between $E[d] - \theta$ is error

* the expected value of statistics is equal to population parameters then this statistic is equal to Unbiased estimate population parameter

$\bar{x}$ is Unbiased estimate of μ

$$\mu = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Assignment:

1, Prove that $\bar{x}$ is Unbiased estimate of μ

$$\mu = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2, Prove that $\bar{x}$ is not Unbiased estimate of σ^2

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

3, In the process of proving 2nd proof. Prove that

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Evaluating an Estimator: Bias and Variance :

Let X be a sample from a population.

$$X = \{x^t, r^t\}$$

x^t → Input, r^t → output

g is an estimate of r

$g \sim r$. r is a population parameter

$E[(x-g)^2]$ is Mean square error (MSE)

$$MSE = E\left[(x - E(x)) + (E(x) - g)\right]^2$$

$$E\left[(x-E(x))^2 + (E(x)-g)^2 + 2(x-E(x))(E(x)-g)\right]$$

$$= E\left[(x-E(x))^2\right] + E\left[(E(x)-g)^2\right] + 0$$

Variance Bias.

ref notes

$2[E(x)-g][E(x)-E(x)]$

Constant

$\Rightarrow 2(E(x)+x)[E(x)-E(x)]$

$= 0$

Notes:

a, Variance is coming from data, It is scatered around mean.

b, the one which we can control is Bias

c, the one which we cannot control is variance

d, Statistic is estimation of the population parameter

* Variance: Variance measures how much, on average, x_i vary around the expected value (going from one dataset to another).

* Bias: Bias measures how much, the expected value varies from the correct value g

* Error: error is the sum of the variance and the square of the bias.

Decision tree:

A decision tree is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree composed of internal decision nodes and terminal leaves

It is an efficient nonparametric method, which can be used for classification and regression.

It is a binary tree.

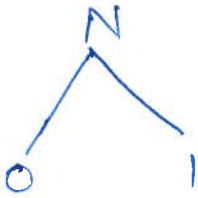
Classification problem:

* Complex decision based on several parameters
* Basian decision stage

*

lets Consider N as property called price, used to decide wether a car is family car or not

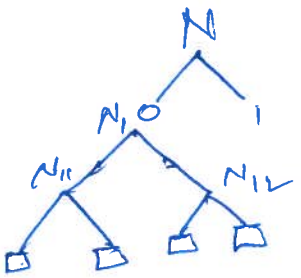
For n number of sample



If price < 500000 No represents 0 leaf in tree.

If price > 10,00000 Yes represents 1 leaf in tree

The above problem can be extended to multiple parameters example price & engine power.



→ node

→ branch

→ leaf nodes.

Each decision node N implements a test function $f_N(x)$ with discrete outcomes labelling the branches. Given an input, at each node, a test is applied and one

of the branches is taken depending on the outcomes.

This process starts at the root and is repeated recursively until leaf node is hit, at which point the value written in the leaf constitutes the output

* Decision tree can manage discrete data very well.
* Continous variable cannot be handled by decision tree, continous variable has to be discretize

Drawbacks of Decision tree:

- It gives best result when number of params are limited.
- It gives better result for very big amount of data which takes care of all possible values atleast once.
- theoritically each node should have atleast 30 points
- Decision tree is specific to sample

Advantages:

- No need to eliminated outliers
- error induced because of outliers is negligible