Data Science and Artificial Intelligence

Machine Learning

Decision Tree

Lecture No. 4

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic
- Topic
- Topic
- Topic
- Topic

decision tree

Topics to be Covered

- Topic
- Topic
- Topic
- Topic
- Topic

decision tree Space and time Complexity

Question

Bias and Variance


+ve mindset
MINDSET IS WHAT SEPARATES THE BEST FROM THE REST.
@daretocultivate|daretocultivate.com

- CART and ID3 Algorithm
- almost Similar algo
- only differ in the method of calculating Impurity.

Basics of Machine Learning

CART(Classification And Regression Tree) for Decision Tree

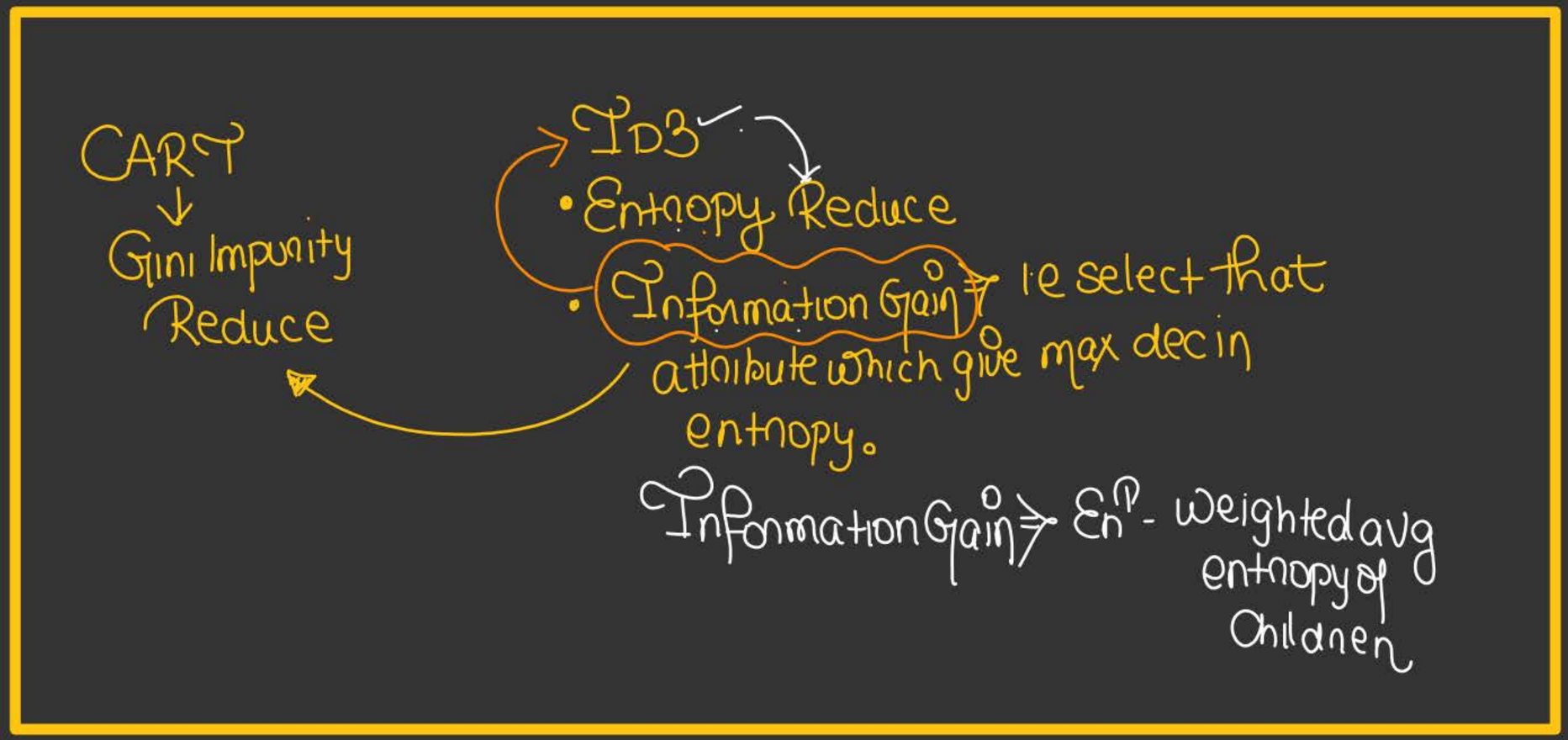
- **Splitting criteria:** CART uses a greedy approach to split the data at each node. It evaluates all possible splits and selects the one that best reduces the impurity of the resulting subsets.

- For classification tasks, CART uses Gini impurity as the splitting criterion. The lower the Gini impurity, the more pure the subset is.

- For regression tasks, CART uses residual reduction as the splitting criterion. The lower the residual reduction, the better the fit of the model to the data.

Iterative Dichotomiser 3 (ID3) Algorithms

- The ID3 algorithm uses a measure of impurity, such as entropy, to calculate the **information gain** of each attribute. **Entropy** is a measure of disorder in a dataset.
- A dataset with high entropy is a dataset where the data points are evenly distributed across the different categories. A dataset with low entropy is a dataset where the data points are concentrated in one or a few categories.


CART
Gini Impurity
Reduce
ID3
• Entropy Reduce
• Information Gain ⇒ ie select that attribute which give max dec in entropy.
Information Gain ⇒ En - weighted avg entropy of Children

Bias and Variance

Practice

Q7. Which one of the following is suitable? 1. When the hypothesis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.

- A. true, false
- B. false, true
- C. true, true
- D. false, false

You are tasked with building a decision tree to classify whether or not a flight will be delayed. One of the features is "Weather Conditions," which can take on three values: "Clear," "Rainy," and "Snowy." Another feature is "Number of Passengers," which is a numerical value representing the number of passengers on the flight. You decide to split the data on the "Weather Conditions" feature first. After the split, you calculate the Gini Impurity for each branch. Here are the results:

For "Clear" weather:

Gini Impurity = 0.25

For "Rainy" weather:

Gini Impurity = 0.50

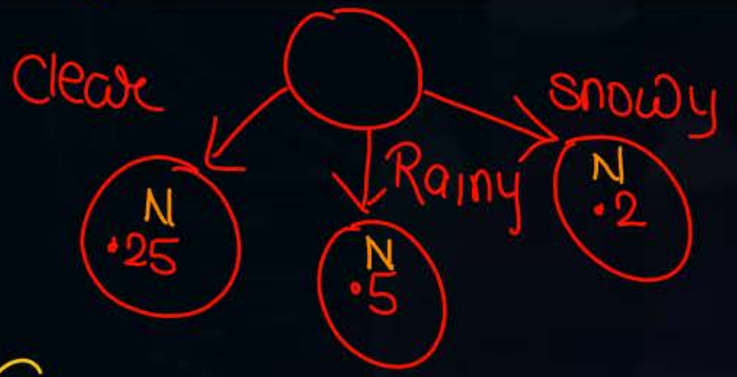
For "Snowy" weather:

Gini Impurity = 0.20

Now, you need to calculate the overall Gini Impurity for the entire split based on "Weather Conditions."

(A) 0.25 (B) ~~0.35~~ .32

(C) 0.40 (D) 0.45



GI ⇒ we have to assume that data is Balanced.

So GI = $\frac{\cdot 2 \times N + \cdot 5 \times N + \cdot 25N}{3N}$

$\Rightarrow \frac{\cdot 95}{3} \Rightarrow (\cdot 31\ldots)$

$\Rightarrow \cdot 32$

Q In LOO-CV, what is the size of the validation set in each iteration for a dataset with N samples?

(A) n

(B) n-1

(C) 1

(D) n/2

You are building a decision tree for predicting the price of houses based on several features. One of the features is the square footage of the house. You've collected data and divided the houses into three categories based on square footage: Small, Medium, and Large. You want to split the data based on square footage into these categories using the following thresholds:

Small: Up to 1,000 square feet

Medium: 1,000 to 2,000 square feet

Large: More than 2,000 square feet

You have 150 houses in your dataset. After splitting based on square footage, you end up with the following counts:

Small: 50 houses

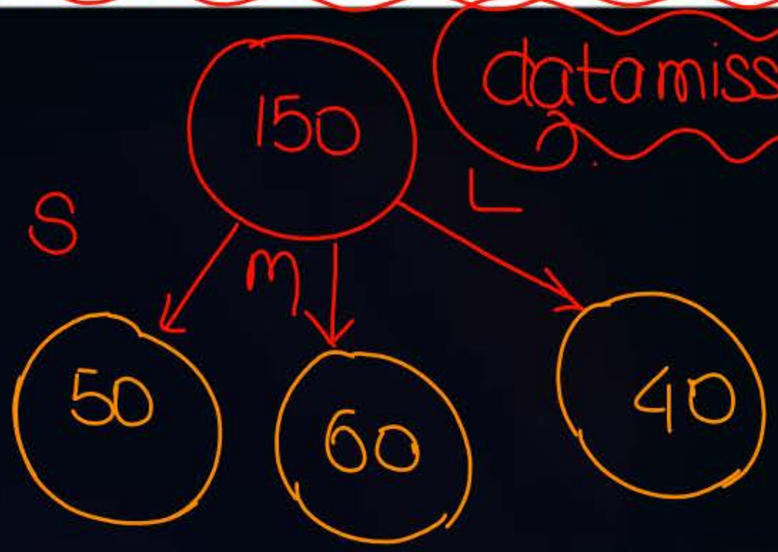
Medium: 60 houses

Large: 40 houses

Calculate the Gini Impurity for this split.

(A) 0.40 (B) 0.48

(C) 0.50 (D) 0.52



datamissing

SKIP

For Root Node.

$$GI = 1 - \Sigma p^2$$

$GI =$

Basics of Machine Learning

CART(Classification And Regression Tree) for Decision Tree

- **Splitting criteria:** CART uses a greedy approach to split the data at each node. It evaluates all possible splits and selects the one that best reduces the impurity of the resulting subsets.
- For classification tasks, CART uses Gini impurity as the splitting criterion. The lower the Gini impurity, the more pure the subset is.
- For regression tasks, CART uses residual reduction as the splitting criterion. The lower the residual reduction, the better the fit of the model to the data.

Iterative Dichotomiser 3 (ID3) Algorithms

- The ID3 algorithm uses a measure of impurity, such as entropy, to calculate the **information gain** of each attribute. **Entropy** is a measure of disorder in a dataset.
- A dataset with high entropy is a dataset where the data points are evenly distributed across the different categories. A dataset with low entropy is a dataset where the data points are concentrated in one or a few categories.

Decision Tree

Practise

In a decision tree, what is the purpose of the leaf nodes?

(A) To represent the class label or value to be predicted

(B) To store the conditions for splitting the data

(C) To indicate the importance of a feature

(D) To represent the depth of the tree

Decision Tree

Practise

What is the primary advantage of using decision trees in machine learning?

A. They are computationally inexpensive

B. They are easy to interpret and visualize

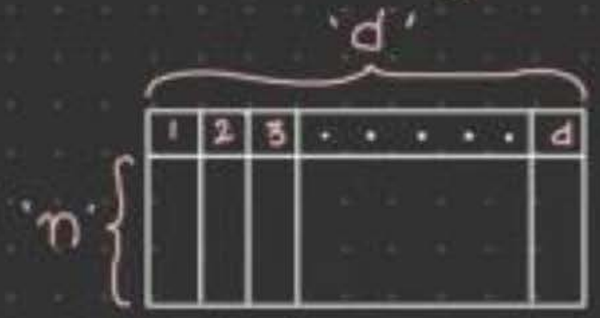
C. They can handle missing data

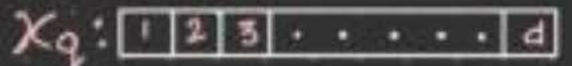
D. They have high predictive accuracy

Time and space complexity of Decision Tree

Time and Space Complexities ?

Training Decision Trees	Inferencing





Time: $O(n \log n * d) + O(n * d)$

$= O(n \log n * d)$

Space: $O(nodes)$

Time: $O(depth)$

Space: $O(nodes)$

Decision Tree

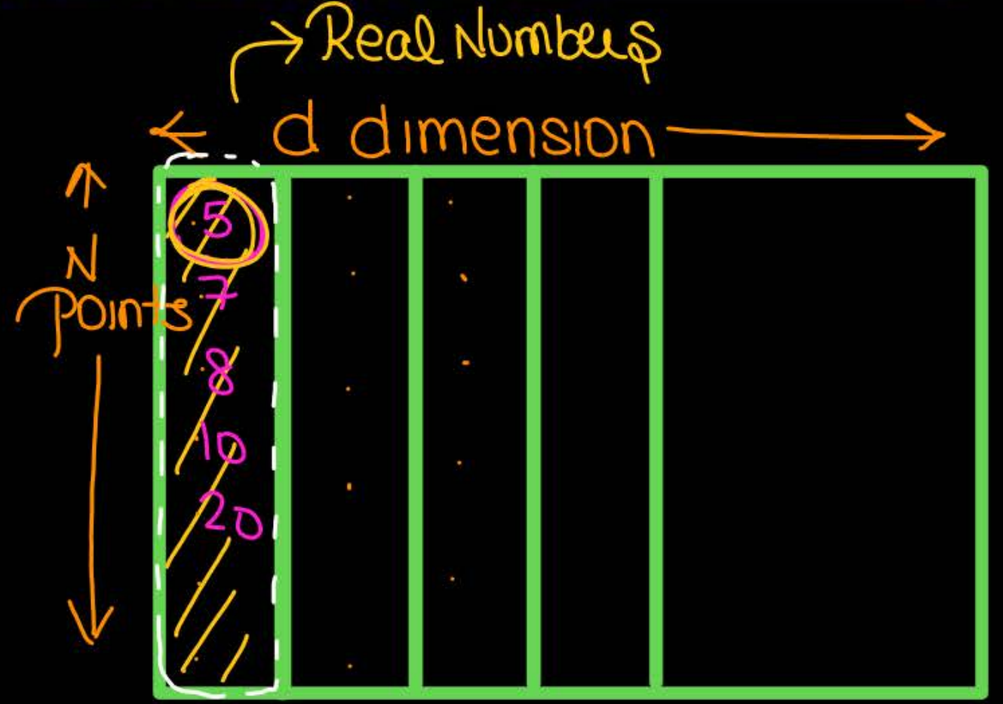
Time and space complexity of Decision Tree

* we have d dimension in data
* n number of points

Root Node

Npoints
ddimension

* we have to check all dimensions
* lets take any dimension



dimension has N values.

So to sort N values $(N \log N)$ Computation

- Now we have to check split of the decision tree at all values.
- Now Repeat the same proces for all dimension

For d dimension $= \{(N \log N) \times d\}$ ← This is order of Computation for a single node

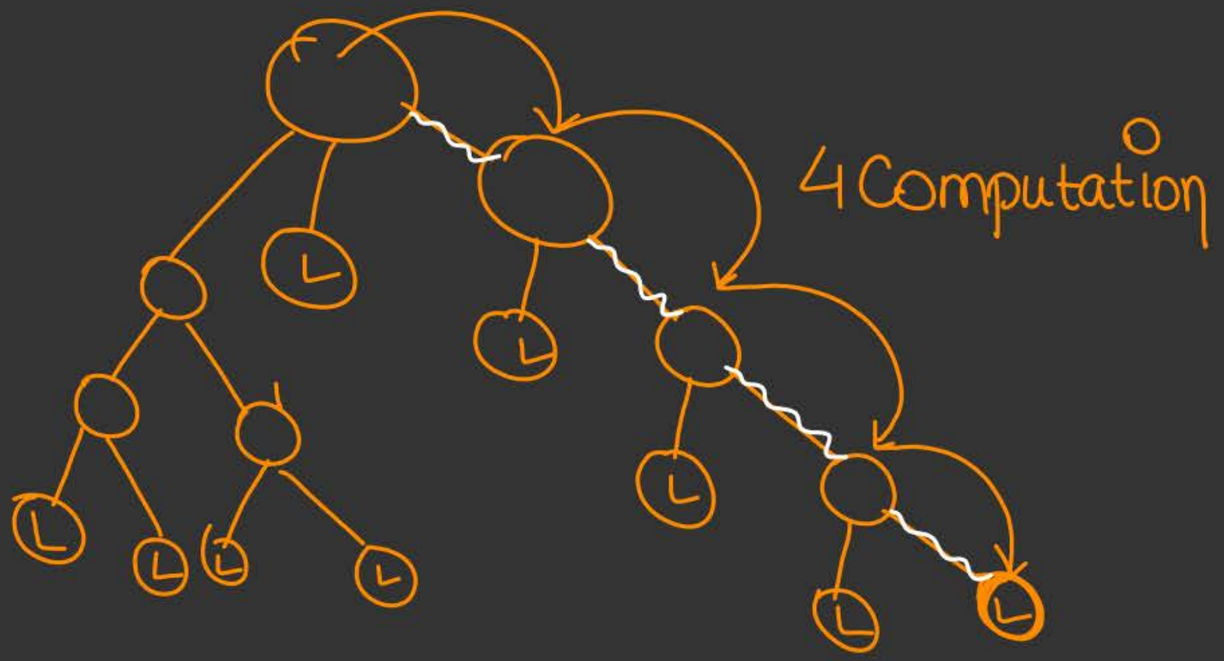
So the time Complexity of DT = $(Nd \log N)$

Space Complexity $\Rightarrow$ Order of Number of nodes.

Because in a DT the Conditions of nodes is to be stored

Testing time Complexity $\Rightarrow$ Order of depth of the tree

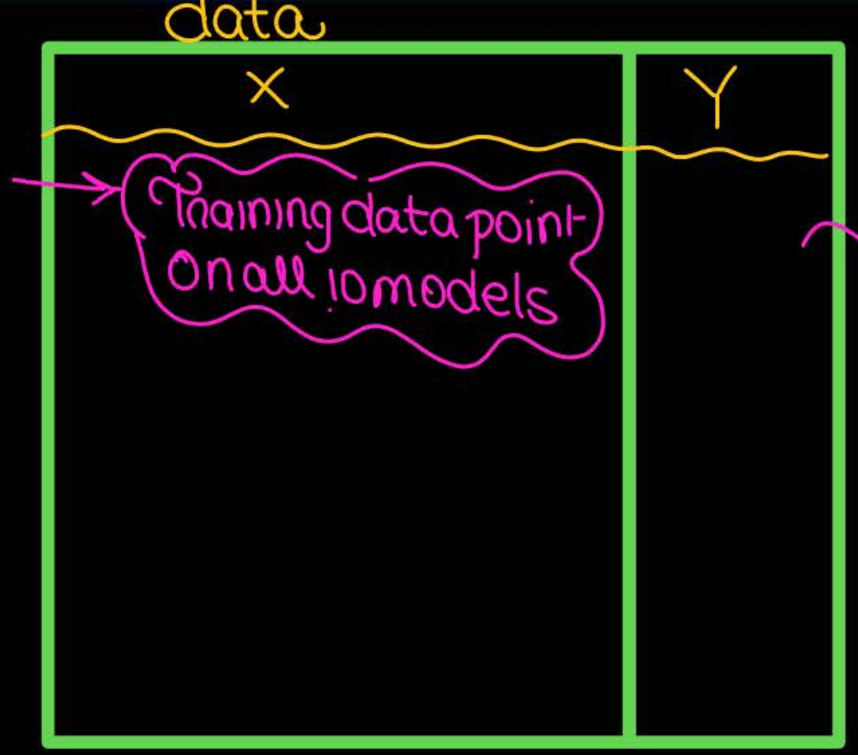
n


4 Computation

Bias and Variance

What is Bias ?

- The training error is called Bias.



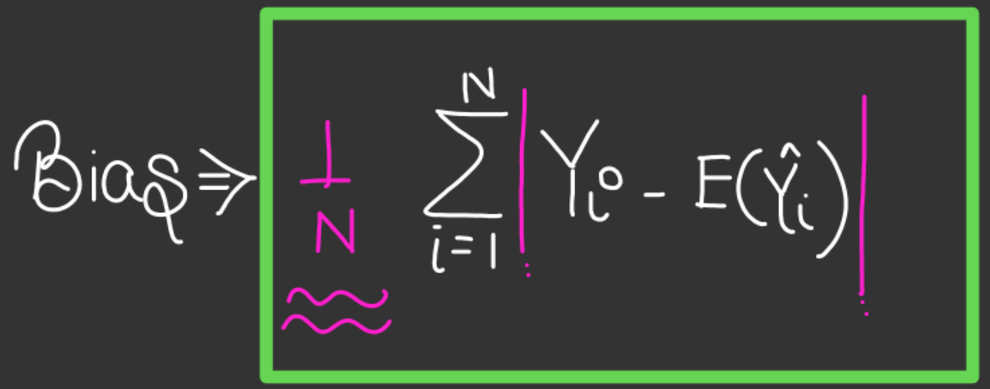
$\hat{Y}$ ⇒ Predicted value of Y from model

So we create many models from data

- $E(\hat{Y})$ ⇒ mean value of $\hat{Y}$

all models give $\hat{Y}$, then $E(\hat{Y})$ = mean of $\hat{Y}$

- Bias = $Y - E(\hat{Y})$


Bias ≈
1/N
N
i=1
|Yi - E(Ŷi)|

Underfitting ⇒ model is not able to understand data/datapattern

How to Remove this problem.

→ Inc No of training data

→ model Complexity Inc

Bias and Variance

What is Underfitting ?

- If the target function is kept too simple, it may not be able to capture the essential nuances and represent the underlying data well
- Many times underfitting happens due to the unavailability of sufficient training data.
- Underfitting can be avoided by
  - 1. using more training data
  - 2. increasing the model complexity
  - 3. Increasing the number of features into analysis
  - 4. Reduce Regularisation — Reduce λ.
  -

Bias and Variance

What is Underfitting ?

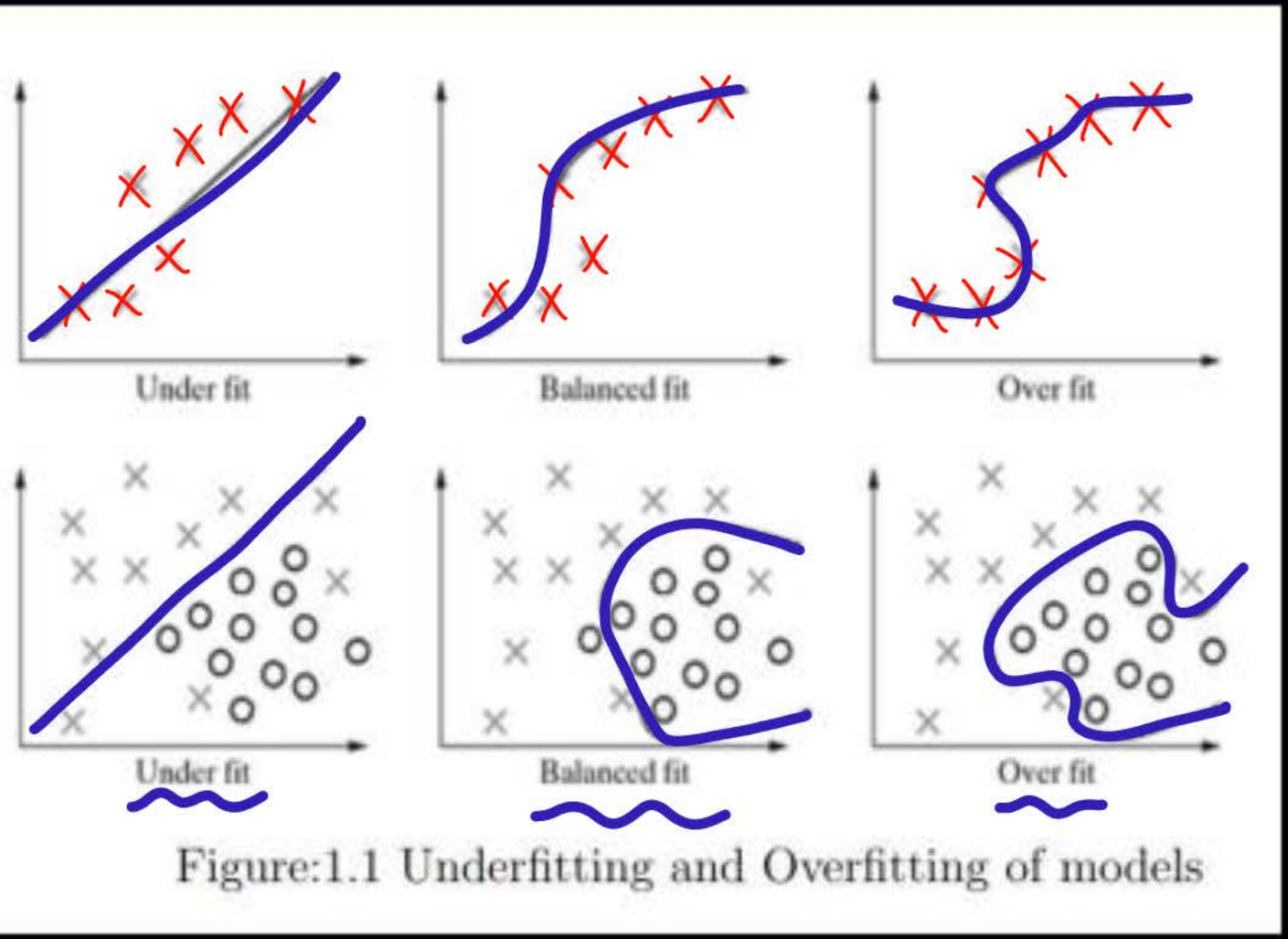


Figure:1.1 Underfitting and Overfitting of models

Bias and Variance

What is Over Fitting ?

- Overfitting refers to a situation where the model has been designed in such a way that it emulates the training data too closely. In such a case, any specific deviation in the training data, like noise or outliers, gets embedded in the model. It adversely impacts the performance of the model on the test data.
- The target function, in these cases, tries to make sure all training data points are correctly partitioned by the decision boundary. However, more often than not, this exact nature is not replicated in the unknown test data set. Hence, the target function results in wrong classification in the test data set.

Bias and Variance

What is Bias ?

- Bias: Bias is the error due to overly simplistic assumptions in the learning algorithm.
- High bias can lead to underfitting, where the model is too simple to capture the underlying patterns in the data.
- Zero bias means overfitting.

Bias and Variance

What is Bias ?

- Formula for bias is ...

$$\Rightarrow \text{Bias} \Rightarrow \frac{1}{N}\sum_{i=1}^{N}|Y_i - E(\hat{Y})|$$

Bias and Variance

How to solve the problem of bias

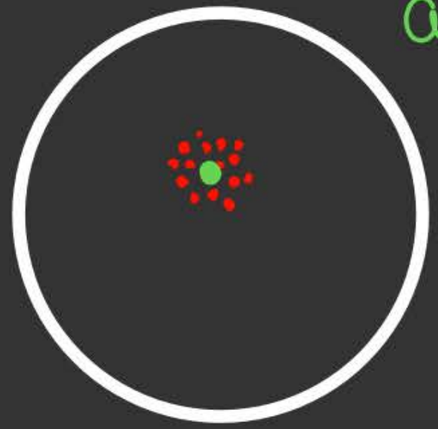
If Bias is large ⇒

1. Increase the model complexity
2. Use fewer assumptions → ex (in linear regression we assume data have linear pattern) etc
3. Add more features
4. Increase the size of training data
5. Reduce Regularisation

- Variance $\Rightarrow$ $E\left[\left(E(\hat{Y}) - \hat{Y}\right)^2\right]$

$\hat{Y}$ = Predicted value from model

$E(\hat{Y})$ mean of all predicted values from all model

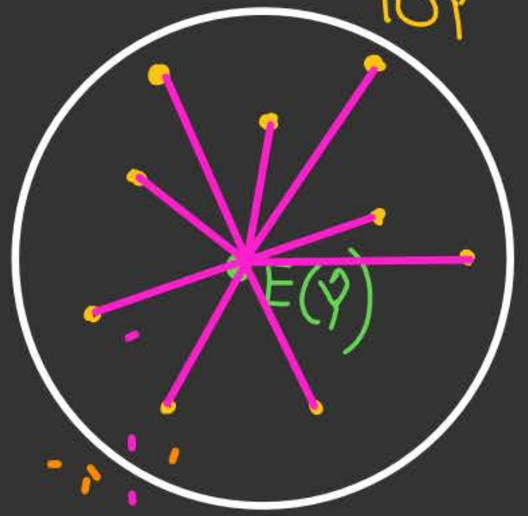


Single point

↓

10 models

↓ 10 $\hat{Y}$



Variance $\Rightarrow E\left[\left(E(\hat{Y})-\hat{Y}\right)^2\right]$

* If any model has high Variance then it will have high testing error
* If Variance is low then low testing error

Why.

- Because when the $\hat{Y}$ of the models are Seperated from each other then it means the models are very different from each other. ⇒ It means the algo used to model is highly data sensitive ⇒ Surely fail for any new point

  → It is imp to note that all the models are created from Portion of same data set.

- If variance is low then it means algo is less data Sensitive ⇒ low testing error.

Bias and Variance

What is Variance

- Variance: Variance is the error due to too much complexity in the learning algorithm. High variance can lead to overfitting, where the model is overly sensitive to noise in the training data and fails to generalize well to new, unseen data.
- Here the ML model is highly sensitive to the training data.
- More specifically, variance is the variability of the model that how much it is sensitive to another subset of the training dataset. i.e. how much it can adjust on the new subset of the training dataset.

Bias and Variance

What is Variance

- Formula for variance is ...

done

Reduce Variance =>

1) Pruning in DT
2) Regularisation
3) Inc the training data
4) Feature Selection, Remove unwanted features.

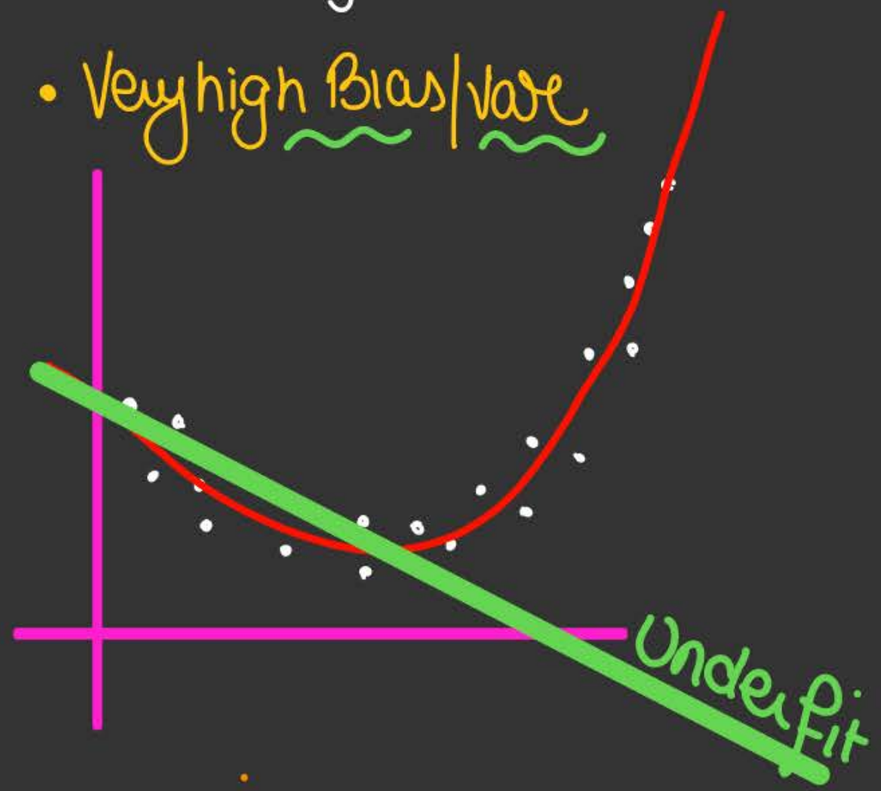
Bias and Variance

What is Variance

- How to reduce variance
1. Increase training data
2. Regularisation
3. Simplify the model
4. Reduce the features or remove unnecessary features
5. Cross validation on training data
6. Early stopping in decision trees.

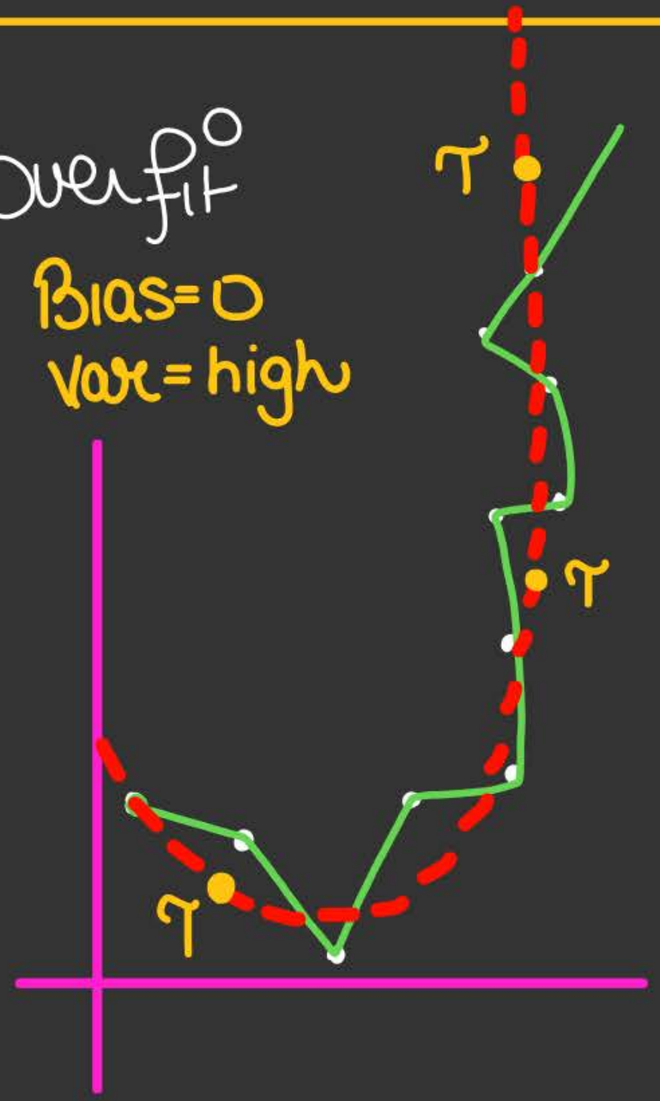
Underfit

- Very high Bias/Var



Overfit

Bias = 0
Var = high



Balanced fit

Variance low
Bias low.

Bias and Variance

Mathematics for Bias and Variance

Let Y be the true value of a parameter, and let $\hat{Y}$ be an estimator of Y based on a sample of data. Then, the bias of the estimator $\hat{Y}$ is given by:

$$\text{Bias}(\hat{Y}) = E(\hat{Y}) - Y$$

$$\text{Variance} = E[(\hat{Y} - E[\hat{Y}])^2]$$

how machine learning model is diff. from true funct fu

Total Error = Mean Square Error

$$= E_D\left\{(Y + \varepsilon - \hat{Y})^2\right\} = E_D\left((Y-\hat{Y}) + \varepsilon\right)^2$$

noise = normal distribution with mean = 0, variance = σ^2

$$= E_D\left[(Y-\hat{Y})^2 + \varepsilon^2 + 2\cdot\varepsilon\cdot(Y-\hat{Y})\right]$$

$E[a+b] = E[a] + E[b]$

$$= E_D\left[(Y-\hat{Y})^2\right] + E_D\left[\varepsilon^2\right] + 2\,E\left[\varepsilon\cdot(Y-\hat{Y})\right]$$

ε ⊥ $(Y-\hat{Y})$

$\mathrm{Var}(X) = E[X^2] - (E(X))^2$

$\mathrm{Var}(\text{noise}) = E(\varepsilon^2) - (E[\varepsilon])^2$

$$= E_D\left[(Y-\hat{Y})^2\right] + \underbrace{\mathrm{Var}(\varepsilon)}_{\sigma^2} - 2\,\underbrace{E_D(\varepsilon)}_{0}\,E_D\left[(Y-\hat{Y})\right]$$

$$= \underbrace{E_D\left[(Y-\hat{Y})^2\right]}_{\text{Reducible Error}} + \underbrace{\sigma^2}_{\text{Irreducible Error}}$$

$E[XY] = E[X]E[Y]$ if $X \perp Y$ and ...

$$E_D[(y-\hat{y})^2] = E_D[(y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2]$$

$E(a+b) = E(a) + E(b)$

$$= E_D[(y-E[\hat{y}])^2] + E[(E[\hat{y}]-\hat{y})^2]$$

$E(E(x)) = E(x)$

$$+ 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})]$$

$$= (y - E[\hat{y}])^2 + E[(E[\hat{y}]-\hat{y})^2] + 2E[y E[\hat{y}] - (E[\hat{y}])^2 - y\hat{y} + E[\hat{y}]\hat{y}]$$

Bias^2 — Variance

$E[ax] = aE[x]$

$E[E[\hat{y}]\cdot\hat{y}] = E[\hat{y}]\cdot E[\hat{y}] = (E[\hat{y}])^2$

$$+ 2[y E[\hat{y}] - (E[\hat{y}])^2 - y E[\hat{y}] + (E[\hat{y}])^2]$$

$= 0$

Total Error = Bias^2 + Variance + Irreducible Error

Total Error = Bias^2 + Variance + Irreducible Error

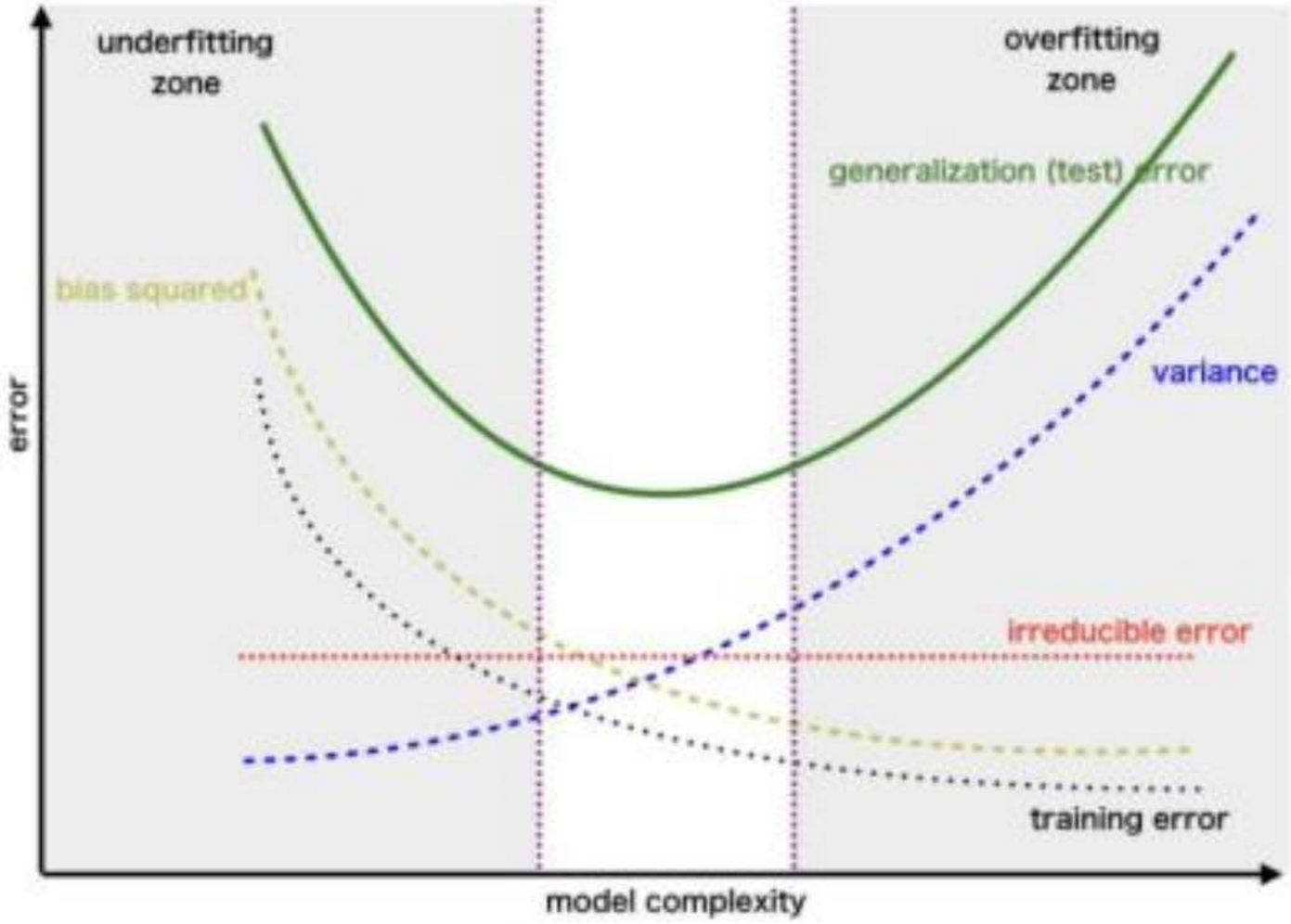
↓ minimize

↓ minimize

X no control

↓ minimize it

Bias and Variance



Region for the Least Value of Total Error

Bias and Variance

Practice

3) Which of the following statements are True? Check all that apply: ***1 point***

- [ ] If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.
- [ ] A model with more parameters is more prone to overfitting and typically has a higher variance.
- [ ] When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.
- [ ] Increasing degree of the polynomial in curve fitting will increase the bias in the model

Bias and Variance

Practice

5) Suppose you have implemented a regularized linear regression model. You observe that on the held out testing set, the model makes unacceptably large errors with its predictions. However, you observe that the model performs well (has a low error) on the training set. Which of the following steps can be incorporated to lower the error on testing dataset. Select all that apply. ***1 point***

- [ ] Try using a smaller set of the features
- [ ] Try decreasing the regularization parameter λ
- [ ] Get more training examples
- [ ] Use fewer training examples

Bias and Variance

Practice

6) Suppose you have implemented a regularized linear regression model. You observe that on the held out testing set, the model makes unacceptably large errors with its predictions. Furthermore, you observe that the model performs **poorly** on the training set. Which of the following steps can be incorporated to lower the error on the testing dataset. Select all that apply ***1 point***

- [ ] Try to obtain an additional set of features
- [ ] Try increasing the regularization parameter λ
- [ ] Get more training examples
- [ ] Try adding polynomial features

Bias and Variance

Practice

7) Suppose you are training a regularized linear regression model. Check which of the following ***1 point*** statements are true? Select all that apply.

- [ ] The regularization parameter λ value is chosen so as to give the lowest training set error
- [ ] The regularization parameter λ value is chosen so as to give the lowest cross validation error
- [ ] The regularization parameter λ value is chosen so as to give the lowest test set error
- [ ] The performance of a learning algorithm on the training set will typically be better than its performance on the test set

Bias and Variance

Practice

Q1. Impact of high variance on the training set ?

- A. overfitting
- B. underfitting
- C. both underfitting & overfitting
- D. depends upon the dataset

Bias and Variance

Practice

Q2. How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?

- A. ridge has larger bias, larger variance
- B. ridge has smaller bias, larger variance
- C. ridge has larger bias, smaller variance
- D. ridge has smaller bias, smaller variance

Bias and Variance

Practice

Q4. You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data. Which is false.

- A. this is an instance of overfitting
- B. this is an instance of underfitting
- C. the training was not well regularized
- D. the training and testing examples are sampled

Bias and Variance

Practice

Q5. Suppose your model is demonstrating high variance across the different training sets. Which of the following is NOT valid way to try and reduce the variance?

- A. increase the amount of training data in each training set
- B. improve the optimization algorithm being used for error minimization.
- C. decrease the model complexity
- D. reduce the noise in the training data

Bias and Variance

Practice

Q6. Which of the following are components of generalization Error?

- A. bias
- B. variance
- C. both of them
- D. none of them

Bias and Variance

Practice

Q7. Which one of the following is suitable? 1. When the hypothesis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.

- A. true, false
- B. false, true
- C. true, true
- D. false, false

Bias and Variance

Practice

Q8. MLE estimates are often undesirable because

- A. they are biased
- B. they have high variance
- C. they are not consistent estimators
- D. none of the above

Bias and Variance

Practice

Q9. Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

- A. you will add more features
- B. you will remove some features
- C. all of the above
- D. none of the above

Bias and Variance

Practice

Q10. We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, What do you expect will happen with the mean training error?

- A. increase
- B. decrease
- C. remain constant
- D. can't say

Bias and Variance

Practice

Q11. We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?

- A. bias increases and variance increases
- B. bias decreases and variance increases
- C. bias decreases and variance decreases
- D. bias increases and variance decreases

Bias and Variance

Practice

Q12. Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to the ideal model.

(i) Models which overfit are more likely to have high bias

(ii) Models which overfit are more likely to have low bias

(iii) Models which overfit are more likely to have high variance

(iv) Models which overfit are more likely to have low variance

Bias and Variance

Practice

Q13. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A. bias will be high, variance will be high
- B. bias will be low, variance will be high
- C. bias will be high, variance will be low
- D. bias will be low, variance will be low

https://mcqmate.com/topic/3/machine-learning-set-4

THANK - YOU