MACHINE LEARNING

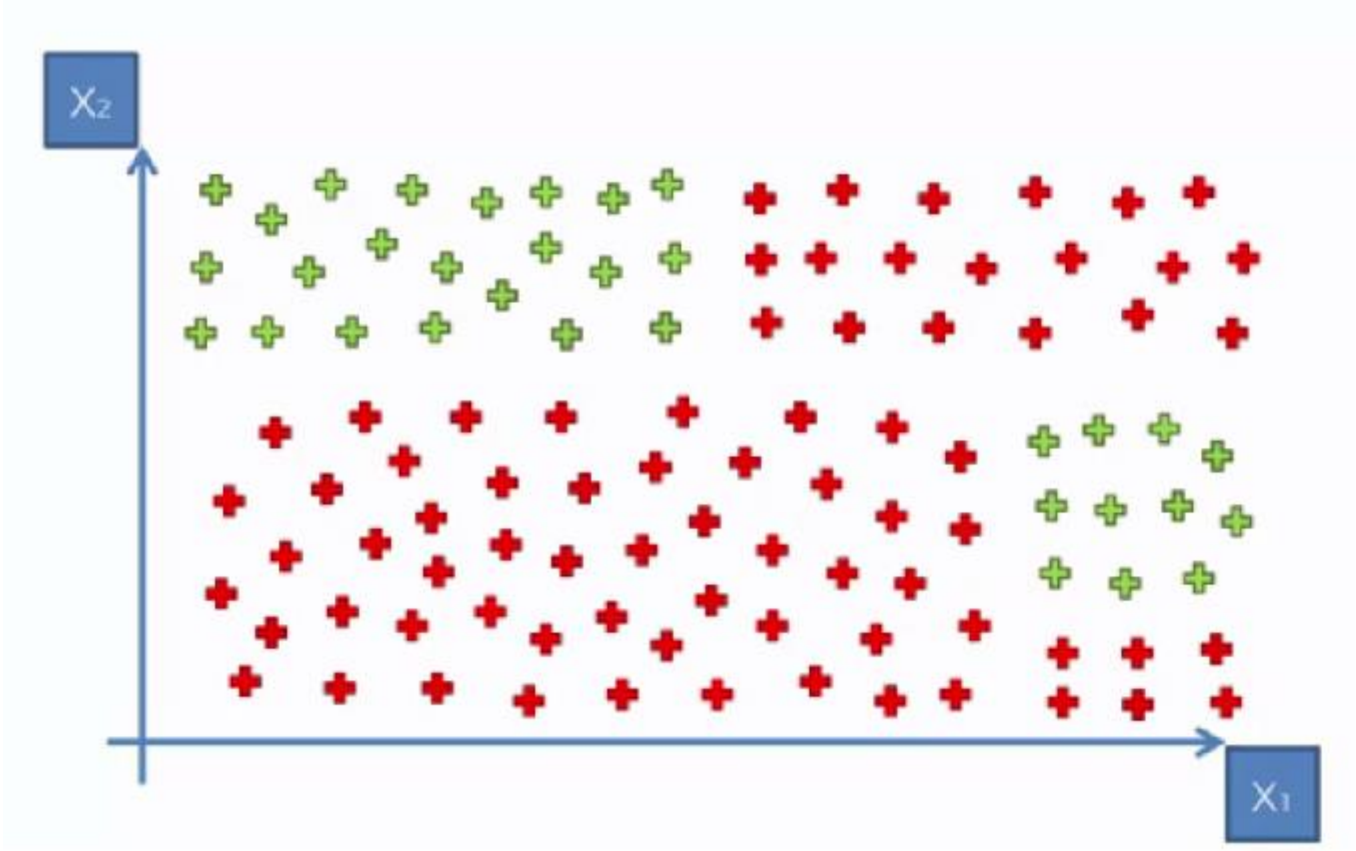
Decision Trees

SHIV NADAR
UNIVERSITY
CHENNAI

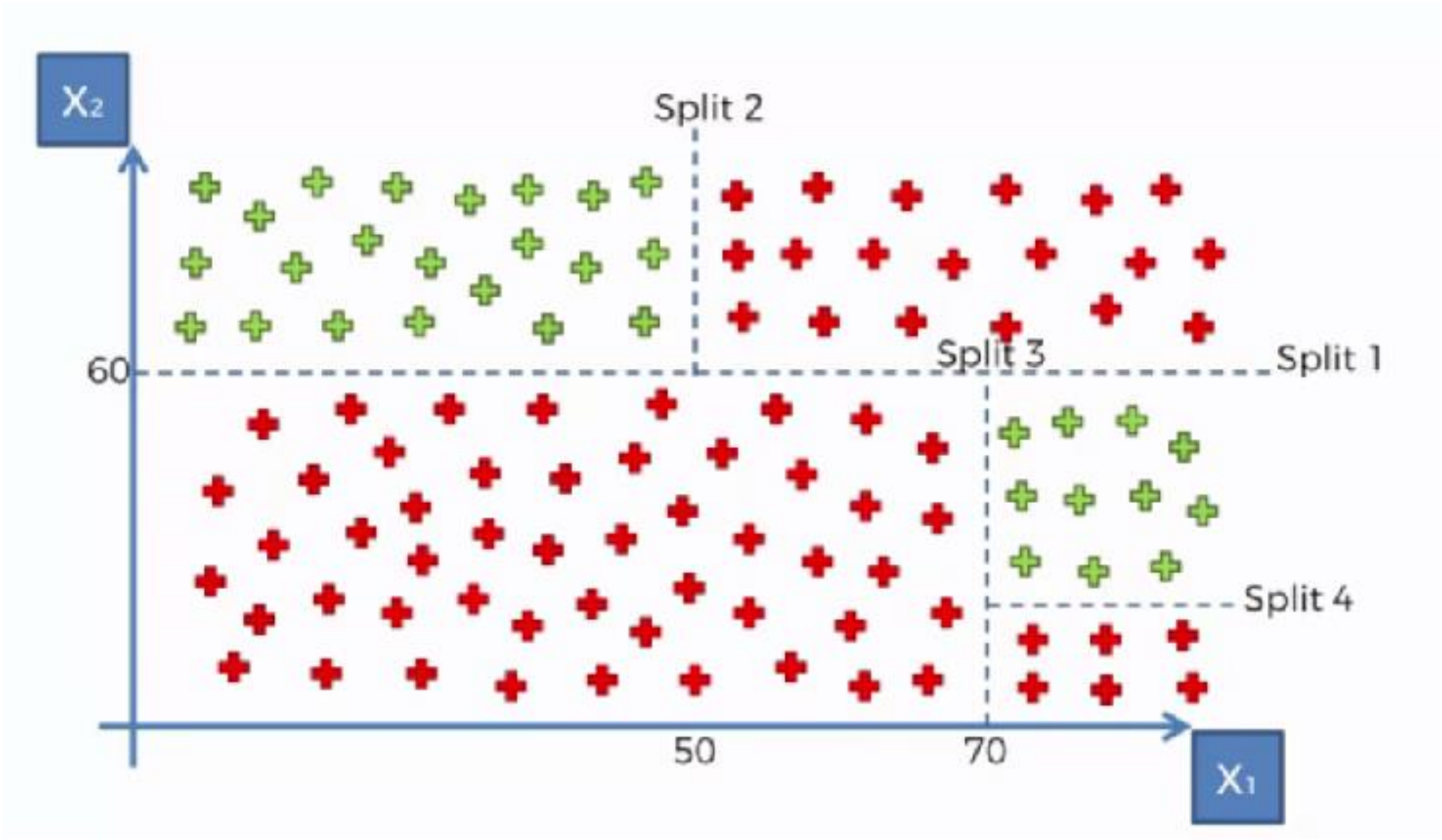
CART

- Classification Trees and Regression Trees
- Classification trees – predict categorical variables.
- Regression Trees – predict continuous values

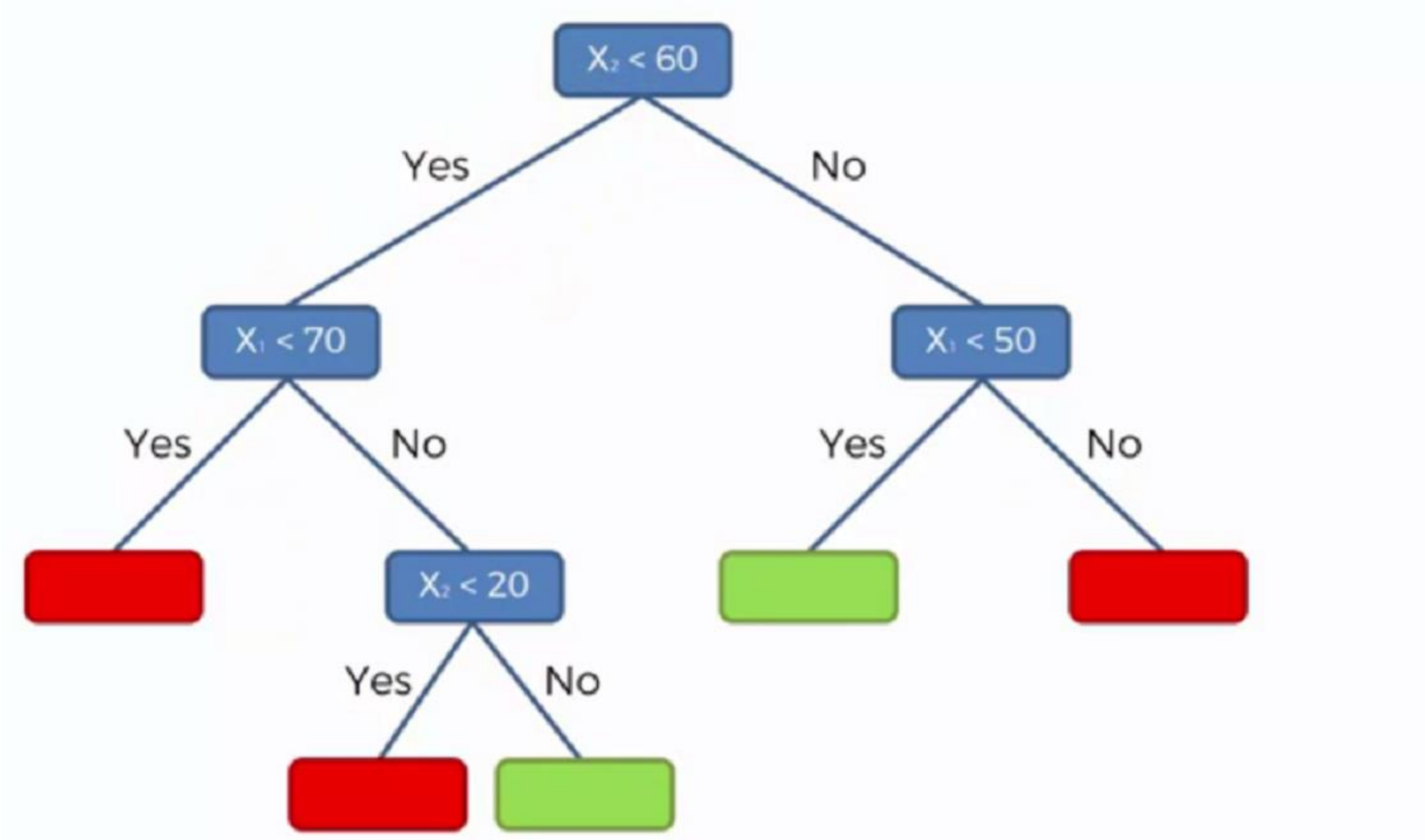
CLASSIFICATION TREES



CLASSIFICATION TREES CONTD..



VISUAL FORM OF DECISION TREE



DECISION TREES

- Decision Trees can be explained by two different ways:
  - Decision Nodes
  - Leaves
- The decision nodes are the nodes where the split occur.
- The leaves are the final outcomes.
- The two main attributes associated with decision trees are:
  - Entropy
  - Information Gain

SHIV NADAR UNIVERSITY CHENNAI

ENTROPY

- Entropy, also called as Shannon Entropy is denoted by H(s) for a finite set s, is the measure of uncertainty or randomness in data.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

- Intuitively, this tells about the predictability of the event. If the entropy is zero, there is absolutely no randomness or in other words, the event can be accurately predicted. For example, for a biased coin which has heads on both sides, the entropy for predicting heads is 0 because it is very certain that every time only heads is going to be the outcome.
- The lesser the entropy, less is the uncertainty. More the entropy, higher the uncertainty.

INFORMATION GAIN

- Information Gain is the effective change in entropy after deciding on a particular attribute.

$$IG(S,A) = H(S) - H(S,A)$$

$$IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

SHIV NADAR UNIVERSITY CHENNAI

SAMPLE DATA

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

CALCULATE THE ENTROPY

$$Entropy(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

$$Entropy(S) = -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right)$$

$$= 0.940$$

INFORMATION GAIN

$$IG(S, Wind) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

1. $H(S_{weak})$
2. $H(S_{strong})$
3. $P(S_{weak})$
4. $P(S_{strong})$
5. $H(S) = 0.94$

SHIV NADAR UNIVERSITY CHENNAI

INFORMATION GAIN

- Wind = weak : 8
- Wind = strong : 6

$$P(S_{weak}) = \frac{Number\ of\ Weak}{Total}$$

$$= \frac{8}{14}$$

$$P(S_{strong}) = \frac{Number\ of\ Strong}{Total}$$

$$= \frac{6}{14}$$

- Out of 8 weak : 6 play golf and 2 do not play

$$Entropy(S_{weak}) = -\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right)$$

$$= 0.811$$

- Similarly for the 6 strong wind: 3 play golf and 3 don't.

$$Entropy(S_{strong}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right)$$
$$= 1.000$$

- With all the values, lets calculate the final information gain for WIND

$$IG(S,Wind) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

$$IG(S,Wind) = H(S) - P(S_{weak}) * H(S_{weak}) - P(S_{strong}) * H(S_{strong})$$

$$= 0.940 - \left(\frac{8}{14}\right)(0.811) - \left(\frac{6}{14}\right)(1.00)$$

$$= 0.048$$

INFORMATION GAIN

- In the same way, calculate the entropy for Outlook, Temperature and Humidity

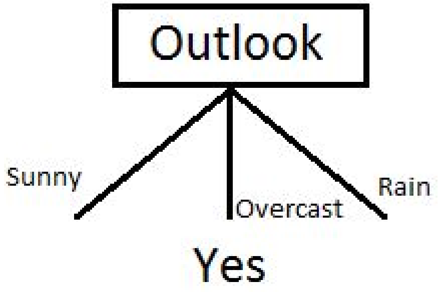
$$IG(S, Outlook) = 0.246$$

$$IG(S, Temperature) = 0.029$$

$$IG(S, Humidity) = 0.151$$

$$IG(S, Wind) = 0.048 \text{ (Previous example)}$$

- We could see from the values that Information Gain for Outlook seems to be highest. Hence, pick that attribute as the root node and start building the decision tree.



C5.0 ALGORITHM

- The decision tree algorithm works as follows:

1. Create root node for the tree
2. If all examples are positive, return leaf node ‘positive’
3. Else if all examples are negative, return leaf node ‘negative’
4. Calculate the entropy of current state H(S)
5. For each attribute, calculate the entropy with respect to the attribute ‘x’ denoted by H(S, x)
6. Select the attribute which has maximum value of IG(S, x)
7. Remove the attribute that offers highest IG from the set of attributes
8. Repeat until we run out of all attributes, or the decision tree has all leaf nodes.

GINI INDEX/ GINI IMPURITY INDEX

- Gini Index = 1- $\sum p^2$
- Eg: Iris data : has 3 classes: Setosa, Versicolor and Virginica
- This data also has 4 columns: Sepal Width, Sepal length, petal Width, petal length
- Condition for splitting: select the column with least variance

BIAS AND VARIANCE

- Inability of a machine learning model to capture the true relationship is called as **Bias.**

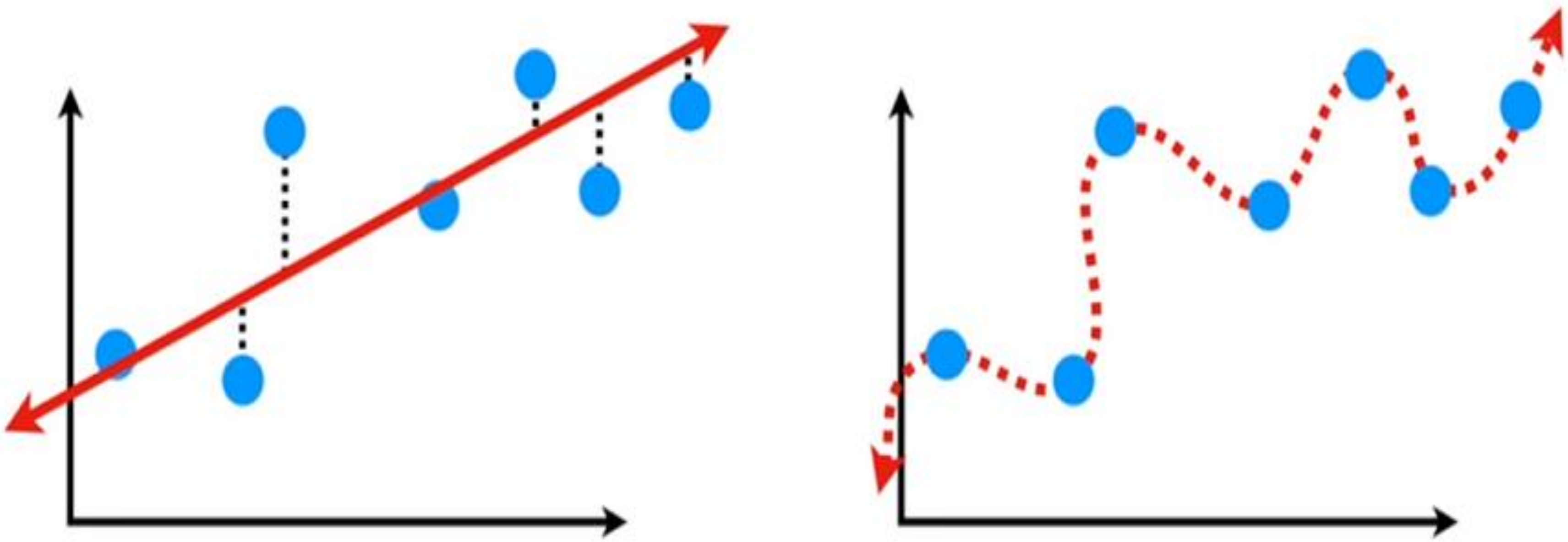
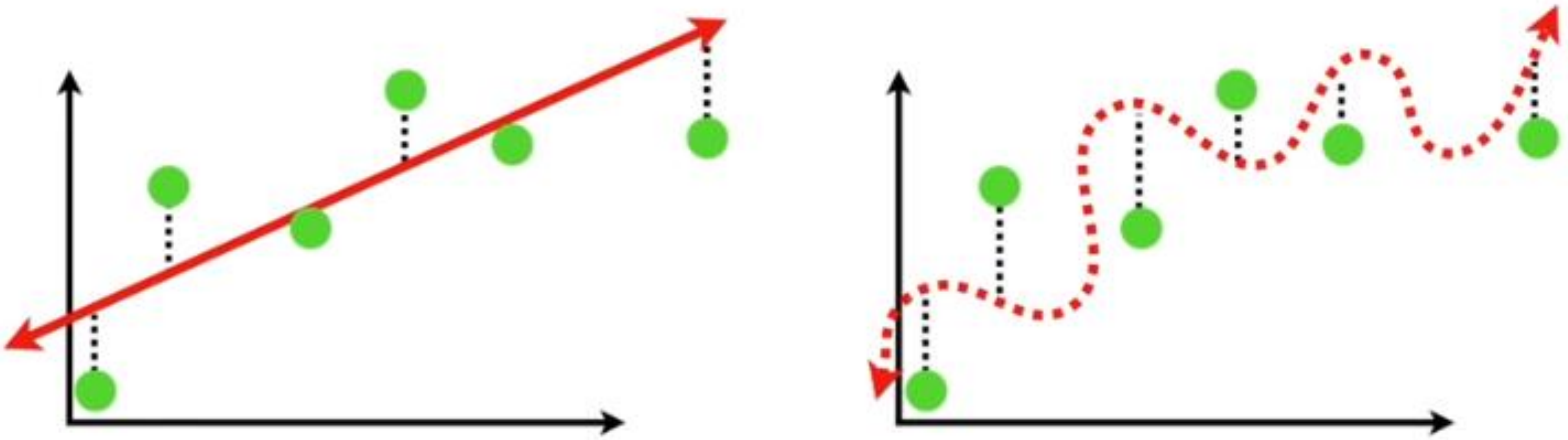


Image Courtesy: StatQuest- Bias and Variance

BIAS AND VARIANCE

- The difference between the test data and the model is called as **Variance.** In other words, variance is the deviation of the test results from the machine learning model.



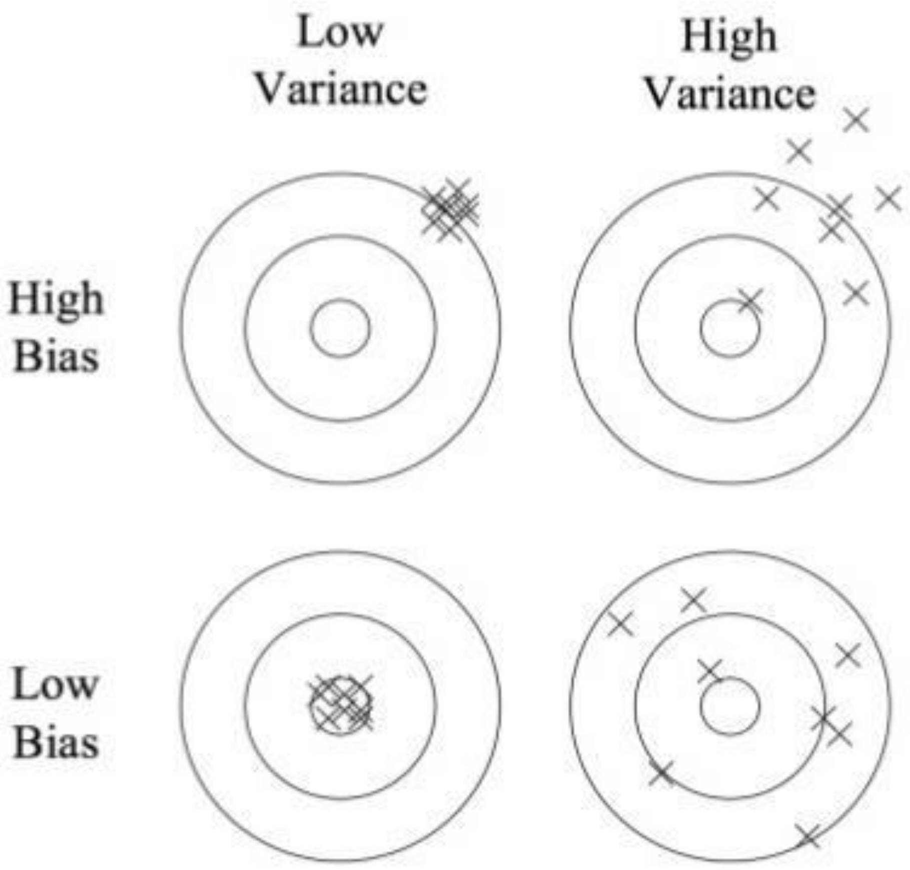
INFERENCES

- The straight line has high bias but less variance. This line can give good predictions but can never give great predictions (underfitting).
- The curved line has low bias but very high variance. This line can sometimes give a great prediction or sometimes gives worst predictions (overfitting).
- An ideal algorithm is the one that has low bias and low variance. But this is not achievable at all instances.

INFERENCES

- Model having a low bias and very high variance is called as overfitted model.
- Model having a low variance and high bias is called underfitted model.
- High Bias is commonly caused due to assumptions in model. Eg:
  - The target variable has a linear relationship between the indep. variables
- Bias of $f'(x) = E[f'(x)] - f(x)$
- Variance $f'(x) = E[(f'(x) - E(f'(x))^2]$
- Trade-off?

BIAS VARIANCE TRADEOFF



FINDING A SWEET SPOT BETWEEN THE SIMPLE MODEL AND COMPLEX MODEL

- 3 common methods are available:
  - Regularization – L1 & L2 (Lasso and Ridge respectively)
  - Boosting – eXtreme Gradient Boosting
  - Bagging – Random Forest