SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601 – Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 3: Bias and Variance

Overfitting and Underfitting

Overfitting and Underfitting

Overfitting

- A statistical model is said to be overfitted when we train it with a lot of data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
- The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.
- Then the model does not categorize the data correctly, because of too many details and noise.
- **Overfitting – High variance and low bias.**

Overfitting and Underfitting

Underfitting

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
- Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.
- It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.
- **Underfitting – High bias and low variance**

Overfitting and Underfitting

A

- Hobby = chating
- Not interested in class
- Doesn't pay much attention to professor

C

- Hobby = learning new things
- Eager to learn concepts.
- Pays attention to class and learns the idea behind solving a problem.

B

- Hobby = to be best in class.
- Mugs up everything professor says.
- Too much attention to the class work.

A

C

A

B

C

A	B	C
Not interested in learning	Memorizing the lessons	Conceptual Learning
Class test ~50% Test ~47%	Class test ~98% Test ~69%	Class test ~92% Test ~89%
Under-fit/ biased learning	Over-fit/ Memorizing	Best-fit

Overfitting and Underfitting

- This situation where any given model is performing too well on the training data but the performance drops significantly over the test set is called an overfitting model.

- For example, non-parametric models like decision trees, KNN, and other tree-based algorithms are very prone to overfitting.

- On the other hand, if the model is performing poorly over the test and the train set, then we call that an underfitting model. An example of this situation would be building a linear regression model over non-linear data.

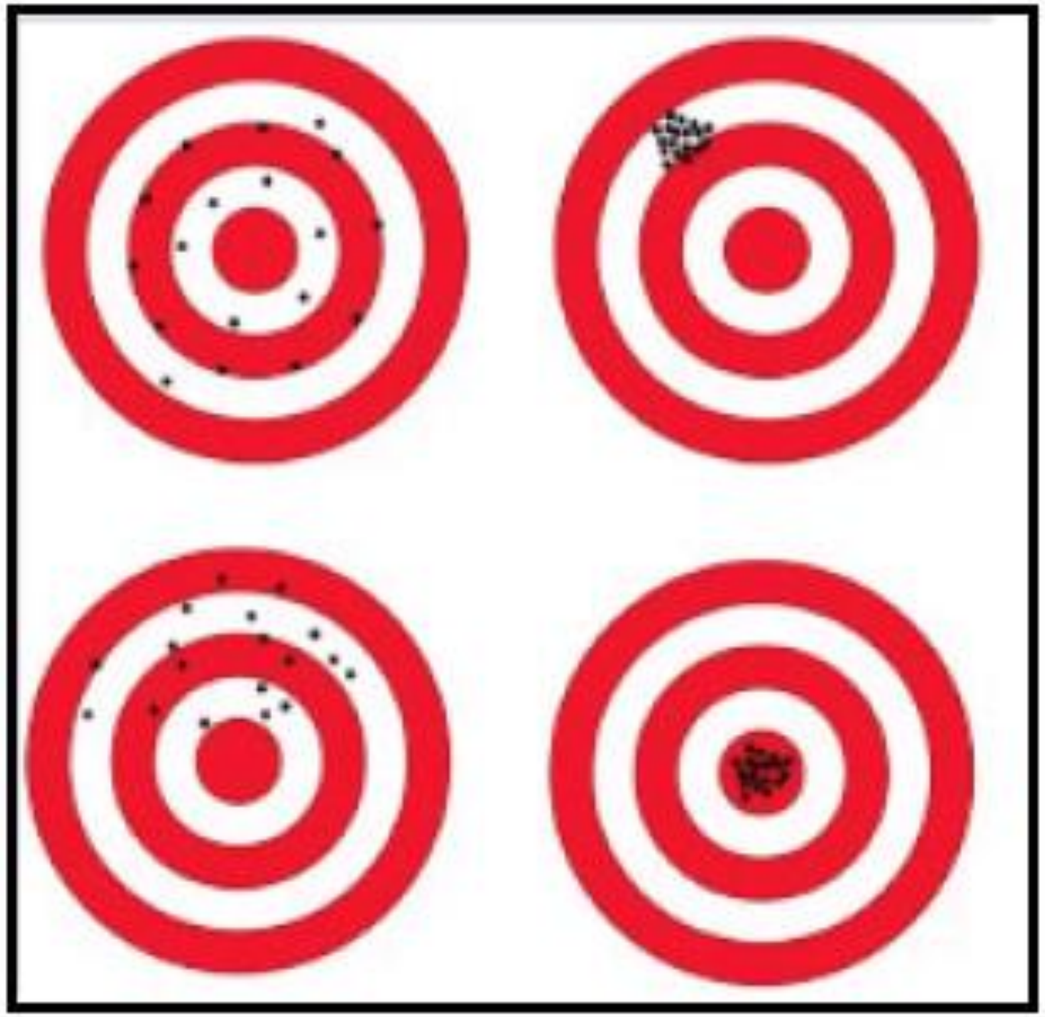
Bias / Variance

- Bias: It measures the difference between the model's prediction and the target value. If the model is oversimplified, the predicted value would be far from the ground truth resulting in more bias.

- Variance: Variance measures the inconsistency of different predictions over a varied dataset. Suppose the model's performance is tested on different datasets—the closer the prediction, the lesser the variance. Higher variance indicates overfitting, in which the model loses the ability to generalize.

- Variance is just a measure of how spread out, how scattered your predictions are.

Bias / Variance



Bias-Variance Trade-Off

Bias-Variance Trade-Off

- The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn the algorithm should achieve good prediction performance.

- There is no escaping the relationship between bias and variance in machine learning.
- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.

- At the end you're not out to just reduce bias or just reduce variance, you want to reduce error.
- A good practice is to check the training error and test error

- That's what really matters, and it turns out you can express error as a function of bias and variance:

$$Error = Bias^2 + Variance$$

K-fold cross-validation to avoid overfitting

- We split all of our data that we're building a machine learning model based off of into two segments
- A training dataset, and a test dataset.
- The idea is that we train our model only using the data in our training dataset, and then we evaluate its performance using the data that we reserved for our test dataset.
- K-fold cross-validation is one of the most common techniques used to detect overfitting.
- Here, we split the data points into k equally sized subsets in K-folds cross-validation, called "folds." One split subset acts as the testing set while the remaining groups are used to train the model.
- k-fold cross-validation splits the dataset into ‘k’ number of folds, then uses one of the ‘k’ folds as a validation set, and the other k-1 folds as a training set. This process is repeated k times, such that each of the k folds is used once as the test set. The scores obtained from this k times training and testing are then averaged to obtain the final score.

K-fold cross-validation to avoid overfitting

The idea, although it sounds complicated, is fairly simple:

1. Instead of dividing our data into two buckets, one for training and one for testing, we divide it into K buckets.
2. We reserve one of those buckets for testing purposes, for evaluating the results of our model.
3. We train our model against the remaining buckets that we have, K-1, and then we take our test dataset and use that to evaluate how well our model did amongst all of those different training datasets.
4. We average those resulting error metrics, that is, those r-squared values, together to get a final error metric from k-fold cross-validation.

THANK YOU