MACHINE
LEARNING

Data Splitting and Bias-Variance Tradeoff

Diptangshu Banik

Twitter - @dipbanik

Overview

- Data needs to be split into train set and test set so that we can validate our results from the model we build.
- Usual split proportion varies from 3:2 to 4:1 depending upon how much data you have.
- Important that your test results and train results are similar so that your model performs well in real world.

Code

```
>>> from sklearn.model_selection import
train_test_split

>>> X_train, X_test, Y_train, Y_test =
train_test_split(X, y, test_size = 0.2,
random_state = 103)
```

Bias

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

- Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before.

- High Variance models perform very well on training data but has high error rates on test data.

Mathematically

$$Y=f(X) + e$$

Where e is the error term

Noise

- Noise is unwanted data items, features or records which don’t help in explaining the feature itself, or the relationship between feature & target.
- Noise often causes the algorithms to miss out patterns in the data.
- Examples –
  - Mistyped information
  - Meaningless features

UnderFitting

- **Underfitting** happens when a model is unable to capture the underlying pattern of the data. These models usually have high bias and low variance.
- Reasons –
  - we have very less amount of data to build an accurate model
  - when we try to build a linear model with a nonlinear data.

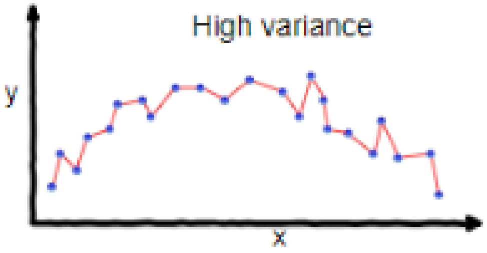
OverFitting

- **Overfitting** happens when our model captures the noise along with the underlying pattern in data. These models have low bias and high variance.

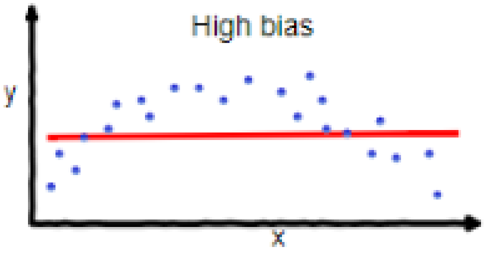
- Reasons –
  - we train our model a lot over noisy dataset
  - when we try to build a linear model with a nonlinear data.

These models are very complex like Decision trees which are prone to overfitting.

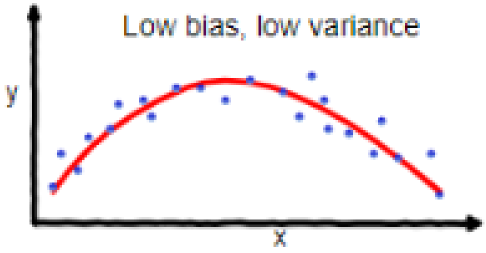
Graphically



overfitting



underfitting

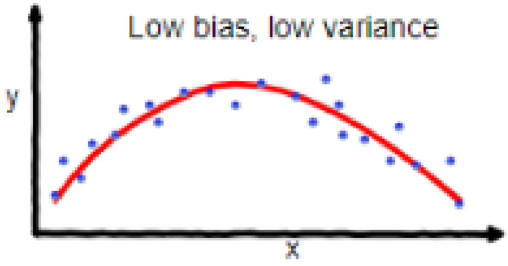


Good balance

Bias-Variance Tradeoff

- To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

- An optimal balance of bias and variance would never overfit or underfit the model.



Good balance