Managing Error & Complexity:

In model prediction there are two main sources of errors:

1) Bias: due to model being unable to represent the complexity of of the data.

2) Variance: due to overly sensitive to the limited data it is being trained on.

Bias: it occurs when there is enough data but the data is not complex enough

1) Underfitting: as a result of data being bias (not complex enough) the model consistently and systematically misrepresents the data leading to low accuracy in prediction.

Hence bias occurs when we have a inadequate models.

* An example is when our OBJECTS are classified by color & shape but our model is only designed to partition and classify objects by color. Here, if we have easter eggs which are colorful, and a rainbow. The model can label rainbow as easter eggs since rainbows are colorful as well. This is an example of bias - under fitting, where we have ample amount of data but not complex enough introducing a consistent and systematic misrepresentation in the model leading to a underfitted prediction (low accuracy)

PTO

NEXT PAGE

Variance : we usually train data using a subset of a population. The predictions vary to an extent depending on the data in the subset.
Here, variance is the measure of how much predictions vary for any gird given test sample.
Some variance is normal, but too much variance indicates the model is unable to generalize the prediction to the larger population.

Overfitting : high sensitivity to the training set is also known as overfitting. It generally occurs when the model is too complex or we do not have enough data to support it.
We can reduce the variability (sensitivity) of the prediction and increase the precision by training on more data.
If more data is not available then we can control the variance by reducing its' complexity.

* We learning_curve from SKLEARN and plot the results to see how much bias or variance is in the model.

* There is a trade off in value of the simplicity of or complexity of a model given a fixed set of data.
* if the model is too simple, our model cannot learn about the data and misrepresents the data.
* If our model is too complex, we need more data to learn underlying relationship
* other wise if the data is underfitted or over fitted it is very common for the model to infer relationships that might not exist.
* The key is to find a sweet point that minizes bias and variance by find the right level of model complexity.
* Given more data any model can improve and different models may be optimal.

high bias
* pay little attention
over simplified
* high error on training set (low r^2, large SSE)
Sum of squared errors: SSE

high variance
* pays to much attention to data (does not generalize well)
* much higher error on the test set than on training set.
* many features, carefully optimized performace

* you want to use min features, with low SSE, this is a SWEET SPOT.

* you have to generalize a prediction solution from the subsets.

Curse of Dimentionality:

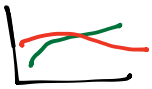
* When number of features grow or dimensions grows, The amount of data we need to generalize accurly grows exponentially!

* LEARNING CURVE: It is a graph which compares the performance of a model on training and testing data over a varying number of training instances.
* we look at relationships between training and performance, we generally see performance improve as number of training points increases.

* by seperating training and testing sets and performance seperately on plots, we get a better idea how we the model can generalize to unseen data.

* learning curve helps us to verify when the model has learned as much possible about the data.
* when this occurs, the performance on both training and test sets plateau and there is consistent gap between the two error rates.

* Bias: when the training and testing data converge and are quite high this means the model is biased. no matter how much data we feed into it it does not change. therefore systematic high errors.



bias ML or

* Variance: when the training and testing data are far apart. this usually means the model suffers from high variance. Unlike, biased model, models that suffer from variance need more data to improve. we can limit variance by simplifying the model to represent only the

* ideal learning curve: has a small gap between the testing and the training curves at similar values. Yielding a good performance.

* The smaller the gap between the training and testing sets, better our m

m

model generalizes.

* model complexity: The visual technique is not limited to learning. with most models we can change complexity by changing inputs or parameters.

* A model complexity looks at training and testing curves as the model complexity varies. The most common trend is that as model complexity increases, bias will fall off & variance will increase/rise.

* sklearn → provides a tool for validation curves which can monitor model complexity.

* Model Complexity VS Learning Curves: if we were to take the same machine learning model with same fixed set of data., but create several graphs at different levels of model complexity, all the learning curve graphs will fit together into a 3D model complexity graph.

we we took the final testing and training errors for each model complexity and visualize them along the complexity of the model, we will be able to see how the model performs as the model complexity increases

* Cross Validation:

* how do we split data into training & testing set.

* cross validation: you split the data into equal bins and use some for tra training & some for testing and then rerun the bins using different testing & training sets. You do this n number of times.

$\frac{K}{10}$-fold C.V