Bias (accuracy) Variance (consistency) Trade-off

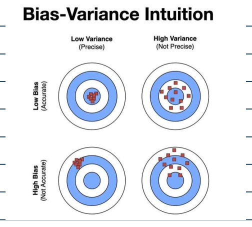
Friday 26 April 2024 12:09 PM

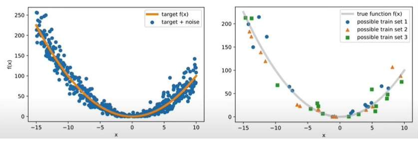
→ Want a model to generalize well to unseen data

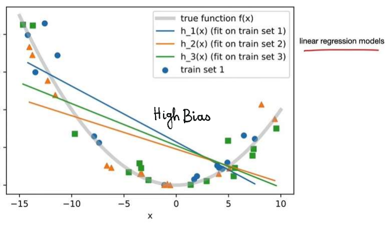
→ Want high generalization accuracy or low generalization error

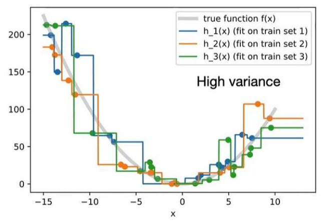
→ Underfitting: both the training & test error are high

→ Overfitting: gap b/w training & test error (Where test error is larger) [Complex model]


Bias-Variance Intuition

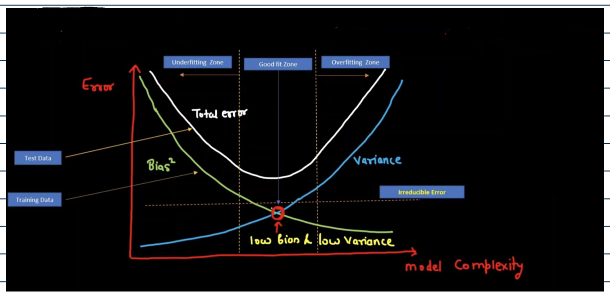


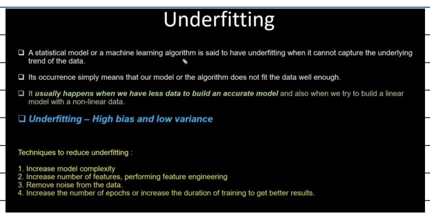




→ Bias - the inability of a ML model to truly capture the relationship in the training data.

→ Variance - the changes in the model when using different portions of the training data set, i.e. variability in the model prediction.



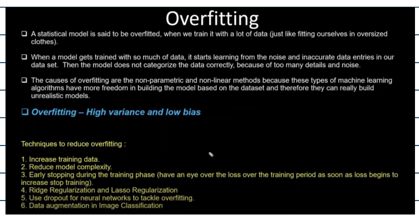


Underfitting

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

Techniques to reduce underfitting :

1. Increase model complexity
2. Increase number of features, performing feature engineering
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.

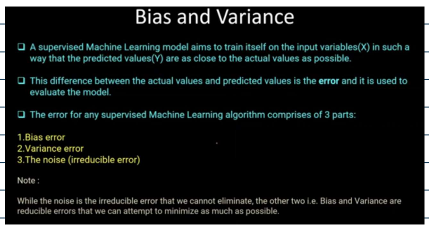


Overfitting

- ❑ A statistical model is said to be overfitted, when we train it with a lot of data (just like fitting ourselves in oversized clothes).
- ❑ When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too many details and noise.
- ❑ The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- ❑ ***Overfitting – High variance and low bias***

Techniques to reduce overfitting :

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.
6. Data augmentation in Image Classification



Bias and Variance

- ❑ A supervised Machine Learning model aims to train itself on the input variables(X) in such a way that the predicted values(Y) are as close to the actual values as possible.
- ❑ This difference between the actual values and predicted values is the **error** and it is used to evaluate the model.
- ❑ The error for any supervised Machine Learning algorithm comprises of 3 parts:

1.Bias error
2.Variance error
3.The noise (irreducible error)

Note :

While the noise is the irreducible error that we cannot eliminate, the other two i.e. Bias and Variance are reducible errors that we can attempt to minimize as much as possible.

→ point estimator $\hat{\theta}$ of some parameter θ

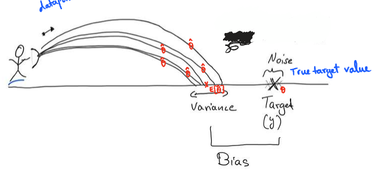
$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$ → Averaging point estimators [The difference b/w the expected prediction of our model & the correct value we are trying to predict]

$\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$ → Spread of point estimators [How much the predictions for a given point vary b/w different relizations of the model]

$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$ $\qquad$ $\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$

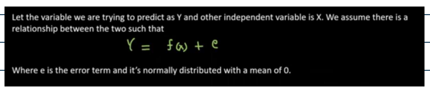
Predictions for a given datapoint

Intuition



→ Bias-Variance decomposition is a way of analyzing a learning algorithm's expected generalization error w.r.t a particular problem by expressing it as a sum of 3 very different quantities: bias, variance, & irreducible error.

mlextend



Let the variable we are trying to predict as Y and other independent variable is X. We assume there is a relationship between the two such that

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

Where e is the error term and it's normally distributed with a mean of 0.

→ $\text{MSE} = (y - \hat{y})^2$

$= (\theta + \varepsilon - \hat{\theta})^2$

$= (\theta - \hat{\theta} + \varepsilon)^2$

$= (\theta - \hat{\theta})^2 + 2(\theta - \hat{\theta})\varepsilon + \varepsilon^2$

$$E[MSE] = E\left[(\theta-\hat{\theta})^2 + \varepsilon^2 + 2\varepsilon(\theta-\hat{\theta})\right]$$

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

$$= E[(\theta-\hat{\theta})^2] + E[\varepsilon^2] + E[2]\cancel{E[\varepsilon]}E[(\theta-\hat{\theta})]$$

$$= E[(\theta-\hat{\theta})^2] + E[\varepsilon^2] \longrightarrow Var(\varepsilon) = \sigma_\varepsilon^2 = E[(\varepsilon - \underbrace{E[\varepsilon]}_{0})^2]$$

$$= E[(\theta-\hat{\theta})^2] + \sigma_\varepsilon^2 \qquad = E[\varepsilon^2]$$

$$= E[(\theta - E[\hat{\theta}] + E[\hat{\theta}] - \hat{\theta})^2] + \sigma_\varepsilon^2$$

$$= E\left[(\theta-E[\hat{\theta}])^2 + 2(\theta-E[\hat{\theta}])(E[\hat{\theta}]-\hat{\theta}) + (E[\hat{\theta}]-\hat{\theta})^2\right] + \sigma_\varepsilon^2$$

$$= E[(\theta-E[\hat{\theta}])^2] + \underbrace{E[2(\theta-E[\hat{\theta}])(E[\hat{\theta}]-\hat{\theta})]} + E[(E[\hat{\theta}]-\hat{\theta})^2] + \sigma_\varepsilon^2$$

$$\rightarrow E[2]E[(\theta-E[\hat{\theta}])]E[E[\hat{\theta}]-\hat{\theta}]$$

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

$$\underbrace{2(\theta-E[\hat{\theta}])(\underbrace{E[\hat{\theta}]-E[\hat{\theta}]}_{0})}_{0}$$

$$= E[(\theta-E[\hat{\theta}])^2] + E[(E[\hat{\theta}]-\hat{\theta})^2] + \sigma_\varepsilon^2$$

$$= (E[\hat{\theta}]-\theta)^2 + E[(E[\hat{\theta}]-\hat{\theta})^2] + \sigma_\varepsilon^2$$

$$= \boxed{\underbrace{Bias^2 + variance}_{\rightarrow \text{reduicable}} + \sigma_\varepsilon^2}$$

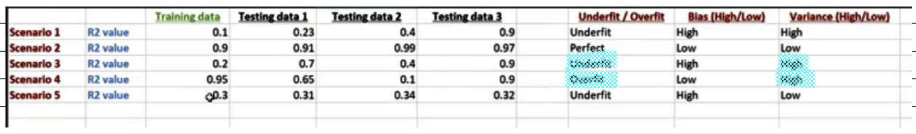
Bias

[overly simplistic model]

[Occurs due to wrong assumption in the model]

[No | less relevant features] [Base assumption is wrong]

look only at training data results



		Training data	Testing data 1	Testing data 2	Testing data 3		Underfit / Overfit	Bias (High/Low)	Variance (High/Low)
Scenario 1	R2 value	0.1	0.23	0.4	0.9		Underfit	High	High
Scenario 2	R2 value	0.9	0.91	0.99	0.97		Perfect	Low	Low
Scenario 3	R2 value	0.2	0.7	0.4	0.9		Underfit	High	High
Scenario 4	R2 value	0.95	0.65	0.1	0.9		Overfit	Low	High
Scenario 5	R2 value	0.3	0.31	0.34	0.32		Underfit	High	Low

Variance

[When there are more than the required relevant features]

[noise in the data]

[overly influenced by the training data]

[Model becomes sensitive to small fluctuations in the data]

[doesn't generalize well]

look at training & test data results spread.

☆ Bias → Underfitting ✓

Variance → overfitting X

→ A special case of variance is overfitting

High Variance doesn't always mean overfit