BIAS VARIANCE TRADEOFF, OVERFIT AND UNDERFIT

Q) What are the prediction error in Machine Error?

Ans - Prediction error can be broken into 3 parts

i) Bias error   ii) Variance Error   iii) Irreducible error

Q) What are irreducible error?

Ans - Irreducible error cannot be reduced regardless of what algorithim is used. It is the error introduced from the choosen framing of the problem and may be caused by the factors like unknown variables that influence the mapping of input variable to the output variable.

Q) What is Bias error?

Ans - Bias are the simplifying assumptions made by a model to make the target function easier to learn.

i) Low bias → Suggest less assumption about the form of target function

Example - Decision tree, KNN, support vector machine.

ii) High bias → Suggest more assumption about the form of target function

Example - Linear regression, Logistic regression.

Q) What is Variance error?

Ans - Variance is the amount that estimates of the target function will change if different training data was used.

i) Low variance → Suggest small changes to estimate of the target function with changes to the training dataset.

Example - Linear regression, Logistic regression.

ii) High variance → Suggest larger changes to estimate of the target function with changes to the training dataset.

Example - Decision tree, KNN, SVM.

Q) Then how to avoid high variance and high bias. (Bias Variance tradeoff)

Ans - Goal of any supervised ML is to acheive low bias & low variance. But in general there is no escaping the relationship between bias & variance

- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.

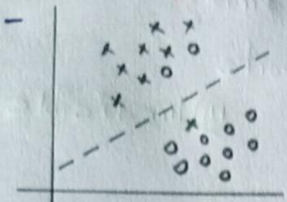
Generally Linear ML algorithim have ~~bias~~ high bias but low variance.

Non Linar ML algorithim have Low bias but high variance.

Tradeoff means balancing between Bias and Variance. Because if one increase other decrease and most suitable is low bias and low variance.

Q) Difference between Linear Data and Non-Linear Data.

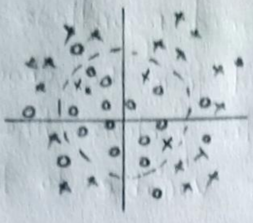
Linear Data -



In this we can draw a line to differentiate between data classes.

Algo is linear → i) Linear Regression.
ii) Logistic Regression.

Non Linear Data -



Here we cannot separate two classes by straight line.

Algo is Non linear - i) Classification & regression tree.
ii) Naive bayes iii) KNN
iv) SVM.

Q) So from programming, how to check bias and variance?

Ans - Basically, bias is a systematic error of the model, which couse model to do false predictions nearly all the time due to incorrect modelling, unrelated data usage etc. We can use different regression model and calculate differences between predicted values and target value in order to assess bias. (RMSE)

- Variance is the dispersion of predicted value over target values with different train-sets. Again, we can train our regression models with those train-sets but this time, we can assess how do predictions vary through each training sets.

- Overall, bias is about the model or data itself and variance is about model sensitivity to training observations.

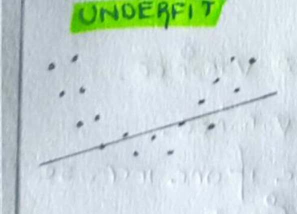
Q) What is underfit and overfit?

Ans - Underfitting: Poor performance on training data and poor genarlization to other data. It follows training data to closely but fails to learn underlying relationship between input & output.

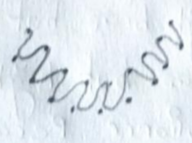
Overfitting: Good performance on training data, poor generalization on other data.

Underfitting is Low variance, high bias. Overfit means high variance, low bias.

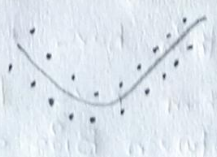
UNDERFIT



OVERFIT



BEST FIT



Solution toward best fit
i) Cross validation
ii) Early stopping
iii) Pruning
iv) Regularization.