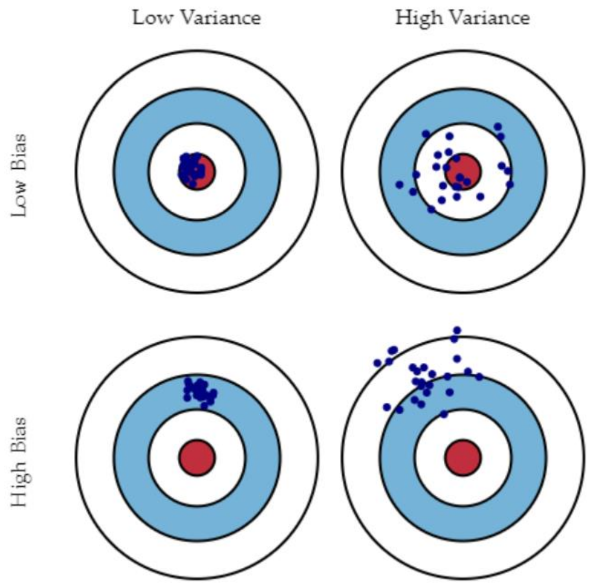
Week 5 Quiz

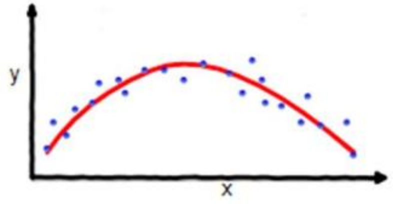
CASA0006

Which of the following statements about bias and variance (in the machine learning context) are correct? (134)

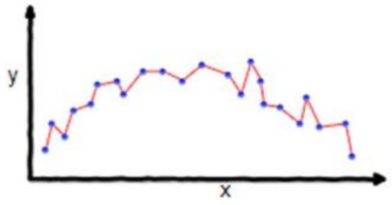
1. To measure bias or variance, multiple runs of models are necessary, which is similar to estimating the mean or variance of a variable.
2. Variance can be measured by the difference between the expected (or average) prediction of the model and the correct value.
3. A model with high variance is sensitive to small fluctuations in the training data.
4. The goal in many machine learning problems is to find a balance between these two sources of error and build a model that generalises well to new, unseen data.



$Y = ax^2+bx+c$



$Y = \sum_{i=0}^{N} a_i x_i$



Which of the following are true about model complexity? (123)

1. Random forest models are generally more complex than linear regression model.
2. The more splits a tree model has, the more complex it is.
3. The more layers/neurons an ANN model has, the more complex it is.
4. The model complexity of a neural network is independent of its number of hidden layers.

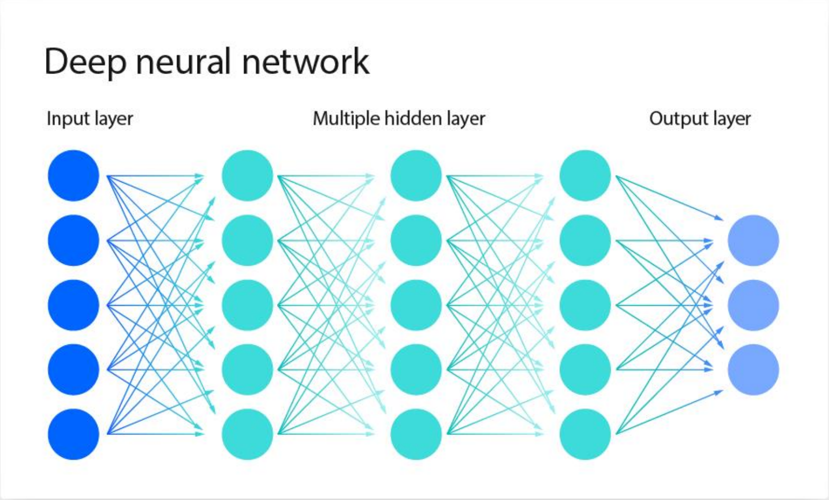
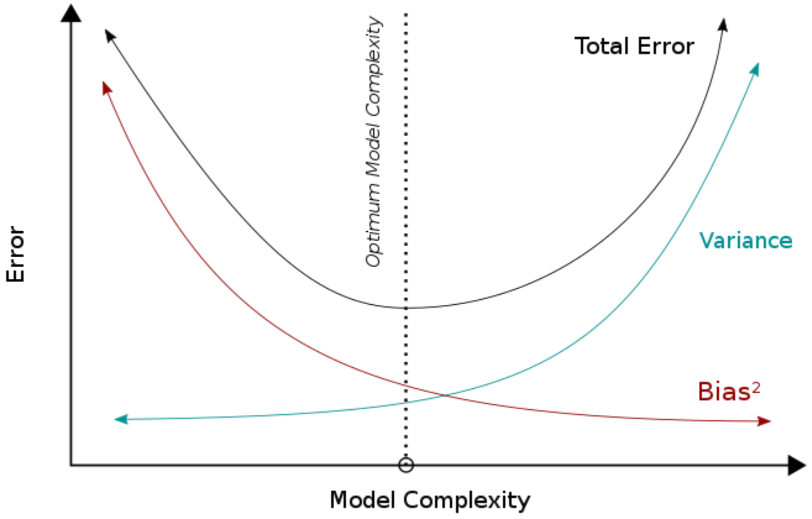
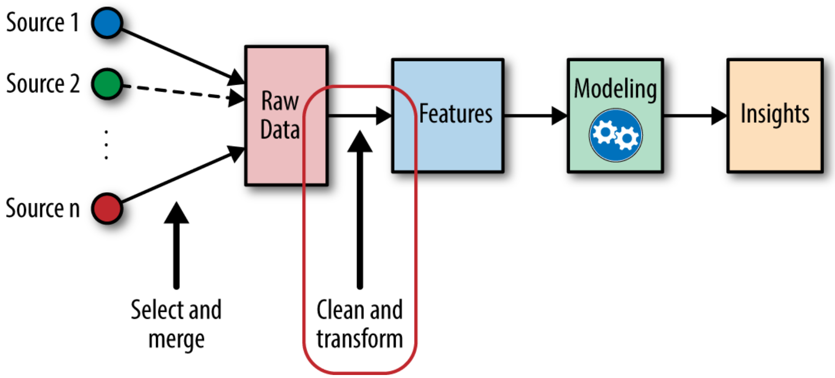


Image Source: https://www.ibm.com/topics/neural-networks

Which of the following is helpful in reducing bias? (134)

1. Increase model size/complexity
2. Use simpler models
3. Feature engineering, e.g., constructing new x variables from existing ones.
4. Using XGBoost to replace decision tree

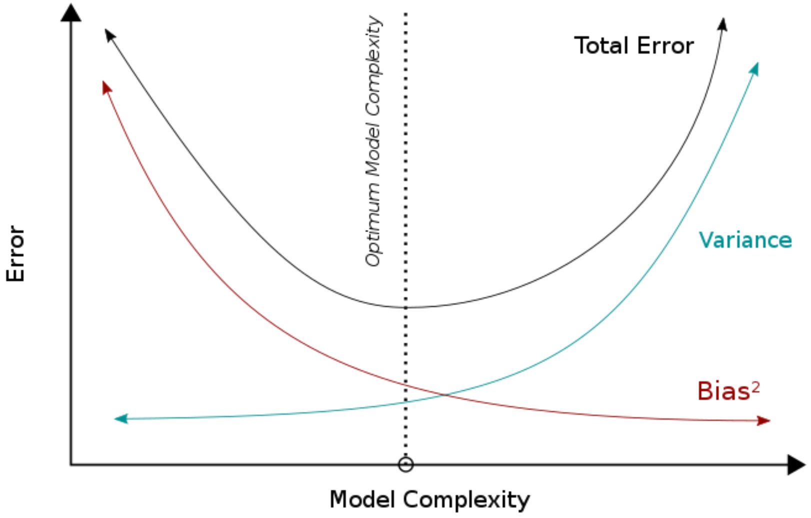


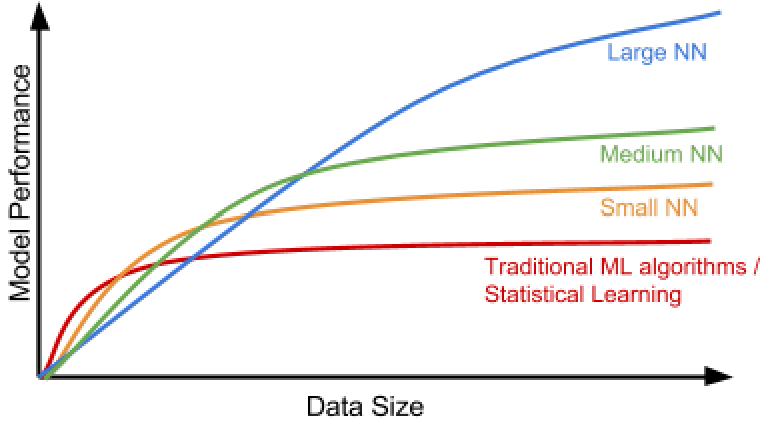


Feature engineering

Which of the following is helpful in reducing variance? (1234)

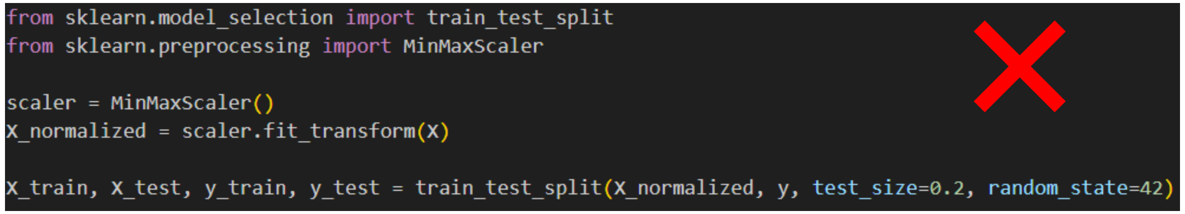
1. Reduce the maximum tree depth of a decision tree
2. Gather more data (Increased representation/diversity, reduce overfitting)
3. Reduce the layers of a large neural network
4. Replace a large neural network model with a smaller one





Which of the following process might have data leakage problem? (24)

1. Hold back a testing dataset for final model performance check.
2. Normalise the data before doing train-test split.
3. Do train-test split first and then normalise the data using only training set.
4. When you are predicting house prices in London, you get the datasets from two companies, Rightmove and Zoopla. Without checking the overlap between these two datasets, you train a neural network model using the Rightmove data and evaluate the performance using the Zoopla data.

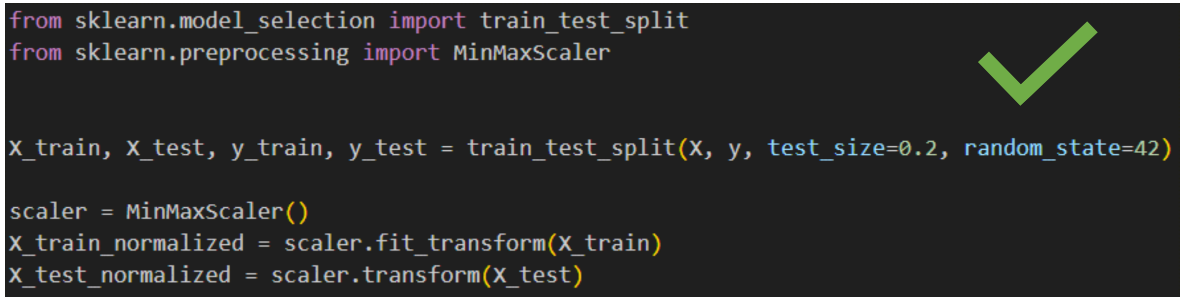


```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_normalized, y, test_size=0.2, random_state=42)
```

✗



```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = MinMaxScaler()
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)
```

✓

Never use any test data during training!