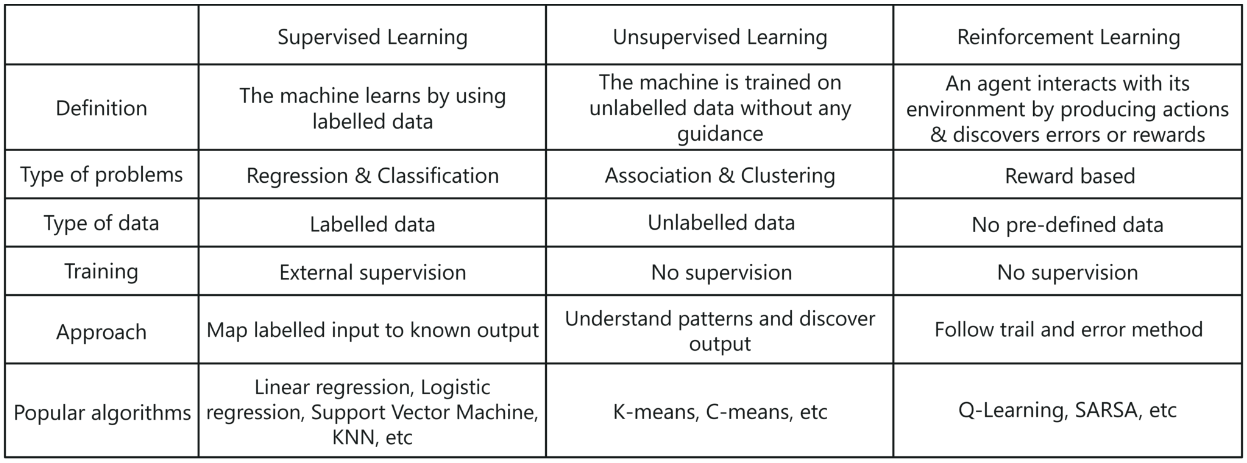
ML Interview Questions

41 Key Machine Learning Interview Questions with Answers

Machine learning interview questions are an integral part of the data science interview and the path to becoming a data scientist, machine learning engineer, or data engineer. Springboard created a free guide to data science interviews, so we know exactly how they

https://www.springboard.com/blog/machine-learning-interview-questions/



	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance	An agent interacts with its environment by producing actions & discovers errors or rewards
Type of problems	Regression & Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trail and error method
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc	Q-Learning, SARSA, etc

Classification	Regression
• Classification is the task of predicting a discrete class label • In a classification problem data is labelled into one of two or more classes • A classification problem with two classes is called binary, more than two classes is called a multi-class classification • Classifying an email as spam or non-spam is an example of a classification problem	• Regression is the task of predicting a continuous quantity • A regression problem requires the prediction of a quantity • A regression problem with multiple input variables is called a multivariate regression problem • Predicting the price of a stock over a period of time is a regression problem

Training data and test data:

- The **training data** is used to make sure the machine recognizes patterns in the data, the **cross-validation data** is used to ensure better accuracy and efficiency of the algorithm used to train the machine,
- And the **test data** is used to see how well the machine can predict new answers based on its training.

Difference Between Algorithm and Model in Machine Learning - Machine Learning Mastery

Last Updated on August 19, 2020 Machine learning involves the use of machine learning algorithms and models. For beginners, this is very confusing as often " machine learning algorithm" is used interchangeably with " machine learning model." Are they the same thing or something different?

https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

Terminologies of Machine Learning

- **Model:** A model is a **specific representation** learned from data by applying some machine learning algorithms. A model is also called a **hypothesis**.
- **Feature:** A feature is an individual measurable property of our data. A set of numeric features can be conveniently described by a **feature vector**. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, **etc. Note:** Choosing informative, discriminating and independent features is a crucial step for effective algorithms. We generally employ a **feature extractor** to extract the relevant features from the raw data.
- **Target (Label):** A target variable or label is the value to be predicted by our model. For the fruit example discussed in the features section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.
- **Training:** The idea is to give a set of inputs(features) and it’s expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.

- **Prediction:** Once our model is ready, it can be fed a set of inputs to which it will provide a

Understanding Hyperparameters and its Optimisation techniques

Model parameters are the properties of training data that will learn on its own during training by the classifier or other ML model. For example,

- Weights and Biases [

[**Weights** control the signal (or the strength of the connection) between two neurons. In other words, a weight decides how much influence the input will have on the output.

Biases, which are constant, are an additional input into the next layer that will always have the value of 1. Bias units are not influenced by the previous layer (they do not have any incoming connections) but they do have outgoing connections with their own weights. The bias unit guarantees that even when all the inputs are zeros there will still be an activation in the neuron. Also see **Deep Learning bias**]

- Split points in Decision Tree

Model Hyperparameters are the properties that govern the entire training process. The below are the variables usually configure before training a model.

- Learning Rate
- Number of Epochs
- Hidden Layers
- Hidden Units
- Activations Functions

Hyperparameters Optimisation Techniques

> The process of finding most optimal hyperparameters in machine learning is called hyperparameter optimisation.

Common algorithms include:

- Grid Search
- Random Search
- Bayesian Optimisation

Explain false negative, false positive, true negative and true positive with a simple example.

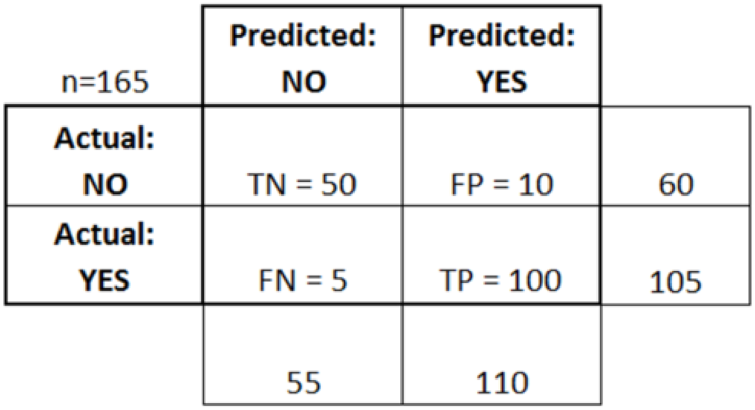
Let’s consider a scenario of a fire emergency:

- **True Positive:** If the alarm goes on in case of a fire. *Fire is positive and prediction made by the system is true.*
- **False Positive:** If the alarm goes on, and there is no fire. *System predicted fire to be positive which is a wrong prediction, hence the prediction is false.*

- **False Negative:** If the alarm does not ring but there was a fire.*System predicted fire to be negative which was false since there was fire.*
- **True Negative:** If the alarm does not ring and there was no fire.*The fire is negative and this prediction was true.*

Q8. What is a Confusion Matrix?

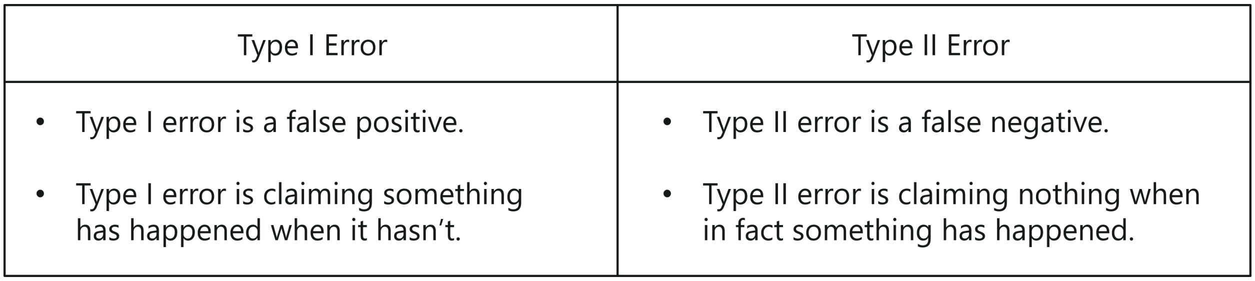
A confusion matrix or an error matrix is a table which is used for summarizing the performance of a classification algorithm.



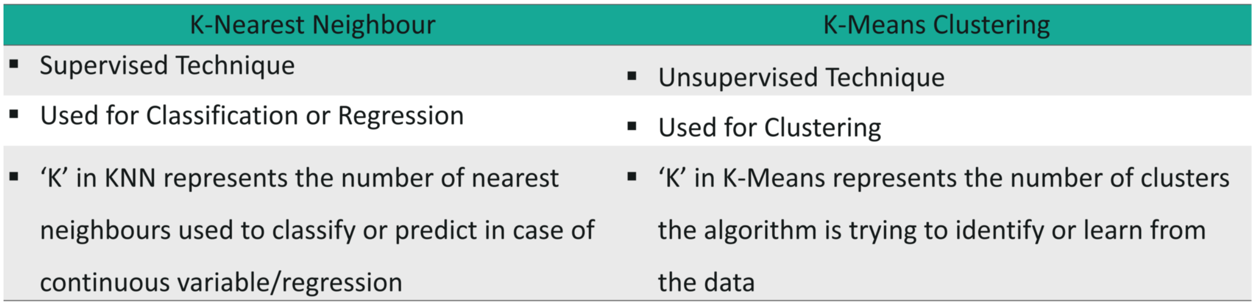
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Consider the above table where:

- TN = True Negative
- TP = True Positive
- FN = False Negative
- FP = False Positive



Type I Error	Type II Error
• Type I error is a false positive. • Type I error is claiming something has happened when it hasn't.	• Type II error is a false negative. • Type II error is claiming nothing when in fact something has happened.



K-Nearest Neighbour	K-Means Clustering
▪ Supervised Technique	▪ Unsupervised Technique
▪ Used for Classification or Regression	▪ Used for Clustering
▪ ‘K’ in KNN represents the number of nearest neighbours used to classify or predict in case of continuous variable/regression	▪ ‘K’ in K-Means represents the number of clusters the algorithm is trying to identify or learn from the data

Which is more important to you – model accuracy or model performance?

- Model accuracy is only a subset of model performance. The accuracy of the model and performance of the model are directly proportional and hence better the performance of the model, more accurate are the predictions.

How to avoid over-fitting?

- This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

1- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.

2- Use cross-validation techniques such as k-folds cross-validation.

3- Use regularization techniques such as LASSO that penalize certain model parameters if they’re likely to cause overfitting.

What is bias?

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

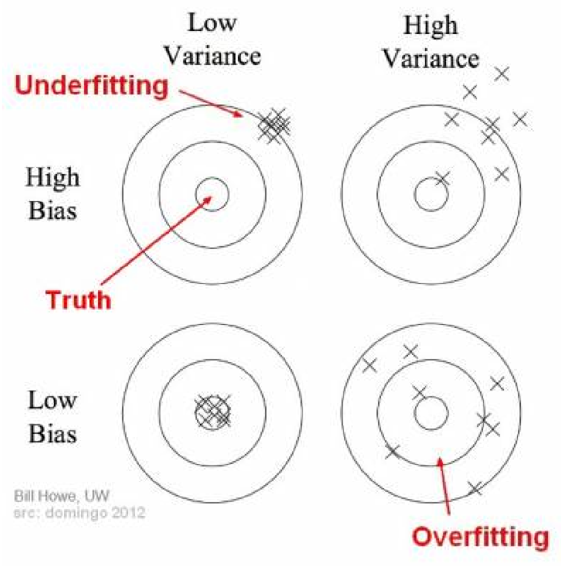
▼ A model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high errors in training and test data.

What is variance?

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

▼ A 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.

▼ As a result, such models perform very well on training data but have high error rates on test data.



Bill Howe, UW
src: domingo 2012

Explain the Bias-Variance Tradeoff.

Predictive models have a tradeoff between bias (how well the model fits the data) and variance (how much the model changes based on changes in the inputs).

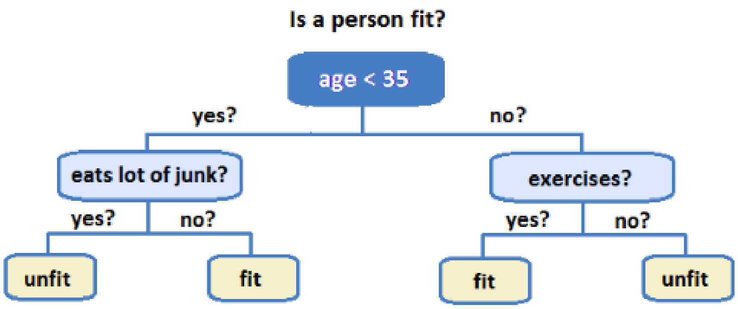
Simpler models are stable (low variance) but they don't get close to the truth (high bias).

More *complex models* are more prone to being overfit (high variance) but they are expressive enough to get close to the truth (low bias).

The best model for a given problem usually lies somewhere in the middle.

Explain Decision Tree algorithm in detail.

Decision tree is a supervised machine learning algorithm chiefly used for regression and classification. The dataset is continually split up into smaller subsets of similar value, to develop a decision tree incrementally. The result is a decision tree where each node represents a feature (attribute), each branch represents a decision (rule) and each leaf represents an outcome (categorical or continuous value).



A SIMPLE DECISION TREE