

LDA (https://www.youtube.com/watch?v=DWJYZq_fQ2A)

<https://www.youtube.com/watch?v=fCmlceNqVog>

GENERATIVE PROCESS EXAMPLE

-Say we have a group of articles and we assume that all of those articles can be characterized by three topics: Animals, Cooking, and Politics.

-Each of those topics can be described by the following words:

- * Animals: dog, chicken, cat, nature, zoo
- * Cooking: oven, food, restaurant, plates, taste, delicious
- * Politics: Republican, Democrat, Congress, ineffective, divisive

-Say we want to generate a new document that is 80% about animals and 20% about cooking.

1. We choose the length of the article (say, 1000 words)
2. We choose a topic based on our specified mixture (so, out of our 1000 words, roughly 800 will come from the topic "animals")
3. We choose a word based on the word distribution for each topic (i.e.

WORKING BACKWARDS

-Suppose you have a corpus of documents

-You want LDA to learn the topic representation of K topics in each document and the word distribution of each topic.

-LDA backtracks from the document level to identify topics that are likely to have generated the corpus.

WORKING BACKWARDS (CONT.)

1. Randomly assign each word in each document to one of the K topics.
2. For each document d :
 - Assume that all topic assignments except for the current one are correct.
 - Calculate two proportions:
 1. Proportion of words in document d that are currently assigned to topic $t = p(\text{topic } t \mid \text{document } d)$
 2. Proportion of assignments to topic t over all documents that come from this word $w = p(\text{word } w \mid \text{topic } t)$
 - Multiply those two proportions and assign w a new topic based on that probability, $p(\text{topic } t \mid \text{document } d) * p(\text{word } w \mid \text{topic } t)$
3. Eventually we'll reach a steady state where assignments make sense

CONCLUSIONS

Documents are probability distributions over latent topics.

Topics are probability distributions over words.

LDA takes a number of documents. It assumes that the words in each document are related. It then tries to figure out the “recipe” for how each document could have been created. We just need to tell the model how many topics to construct and it uses that “recipe” to generate topic and word distributions over a corpus. Based on that output, we can identify similar documents within the corpus.

CONCLUSIONS

ADVANTAGES

LDA is an effective tool for topic modeling.

Easy to understand conceptually

Has been shown to produce good results over many domains.

New applications

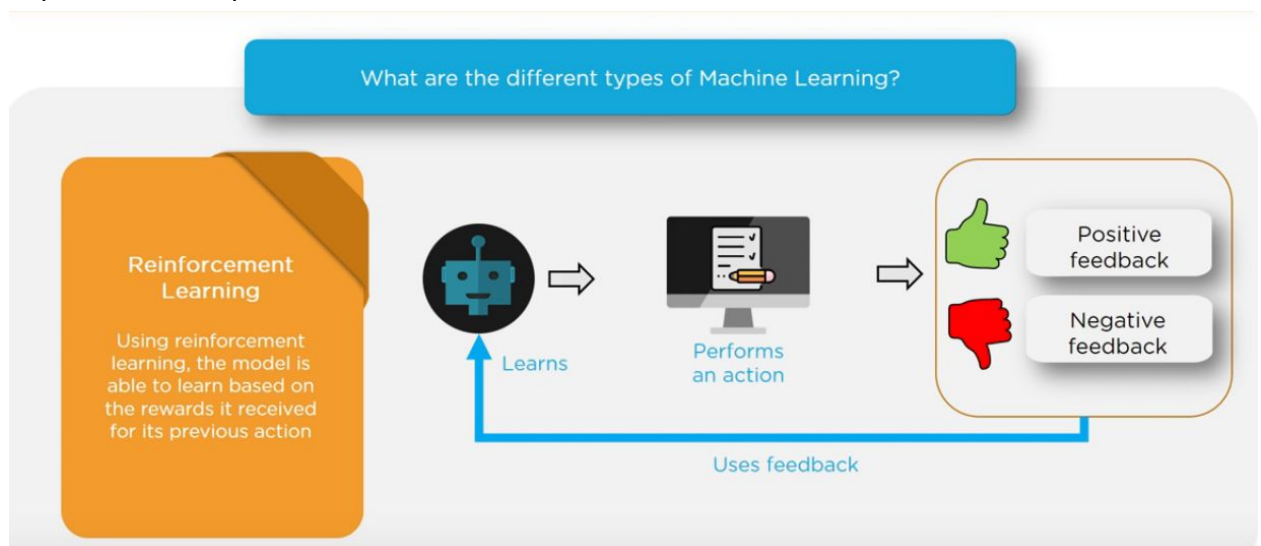
LIMITATIONS

Must know the number of topics K in advance

Dirichlet topic distribution cannot capture correlations among topics

Machine Learning interview questions:

1. Type of machine learning
Supervised, unsupervised, reinforcement



02 Machine Learning Interview Questions

What is overfitting? And how can you avoid it?

- ❑ Overfitting occurs when the model learns the training set too well. It takes up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data



Training Dataset



Testing Dataset

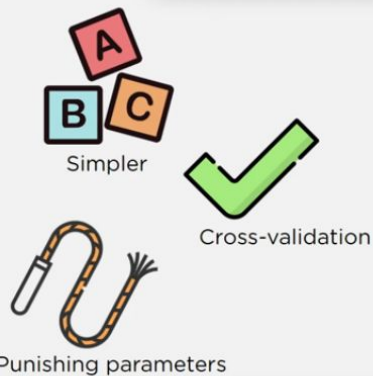


Misclassified Data

- ❑ Here the model is overfit to the training dataset and will give error when new testing dataset is introduced
- ❑ High loss and low accuracy is seen in the test dataset

02 Machine Learning Interview Questions

What is overfitting? And how can you avoid it?



There are three main methods to avoid overfitting:

- ❑ Regularization: This involves a cost term for the features involved with the objective function
- ❑ Make a simple model: With lesser variables and parameters, the variance can be reduced
- ❑ Cross-validation methods, like k-folds can also be used
- ❑ If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

04 Machine Learning Interview Questions

How do you handle missing or corrupted data in a dataset?

The ways to handle missing / corrupted data is to drop those rows / columns or replace them completely with some other value

There are two useful methods in Panda:

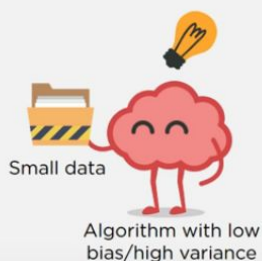
- a) `isnull()` and `dropna()` will help finding the columns / rows with missing data and drop them
- b) `fillna()` will replace the wrong values with a placeholder value(0)



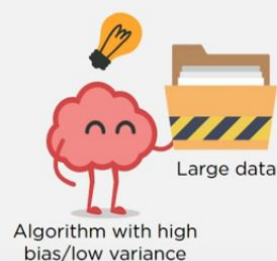
05 Machine Learning Interview Questions

How can you choose a classifier based on training set size?

When the training set is small, a model that has a high bias and low variance seems to work better because they are less likely to overfit. For e.g. Naïve Bayes works best



When the training set is large, models with low bias and high variance tend to perform better as they work fine with complex relationships. E.g. Decision Tree



06 Machine Learning Interview Questions

Explain confusion matrix with respect to Machine Learning algorithms.

- ❑ **Confusion matrix** (or error matrix) is a specific table that is used to measure the performance of an algorithm
- ❑ It is mostly used in **supervised learning** (in unsupervised learning it is called matching matrix)
- ❑ Confusion matrix has two dimensions:
 1. **Actual**
 2. **Predicted**
- ❑ It also has identical sets of features in both these dimensions

		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

07 Machine Learning Interview Questions

What is false positive and false negative and how are they significant?

		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

False Positive: 3

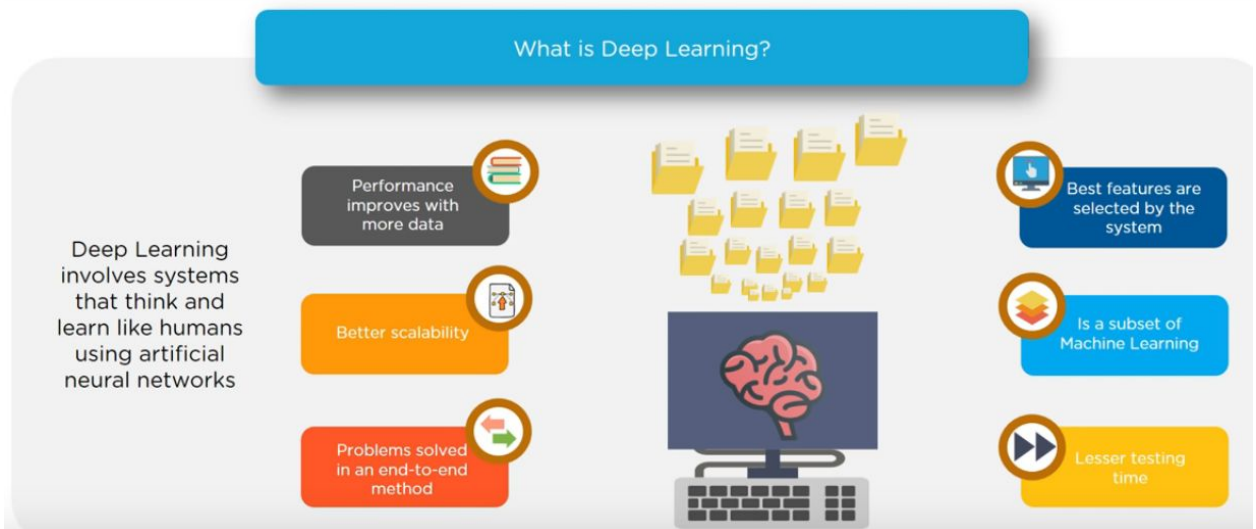
False Negative: 1

False Positive are those cases which wrongly get classified as **True** but are actually **False**

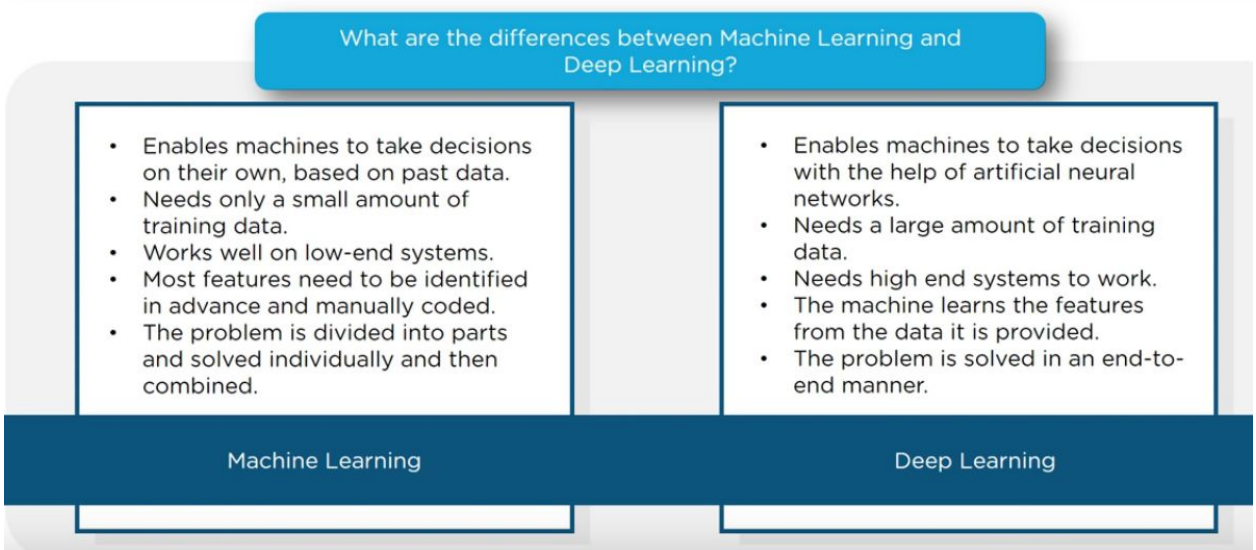
False Negative similarly are those cases which wrongly get classified as **False** but are **True**

True Positive: 12
False Positive: 3
True Negative: 9
False Negative: 1

09 Machine Learning Interview Questions



16 Machine Learning Interview Questions



11 Machine Learning Interview Questions

What are the applications of supervised Machine Learning in modern businesses?



Email Spam Detection



Sentiment Analysis



Healthcare Diagnosis



Fraud Detection

12 Machine Learning Interview Questions

What is semi supervised Machine Learning?

Supervised Learning uses training data that is completely labeled and Unsupervised Learning uses no training data

In case of Semi-Supervised Learning, the training data contains of a small amount of labeled and a large amount of unlabeled data



Training data



Learns



Test data



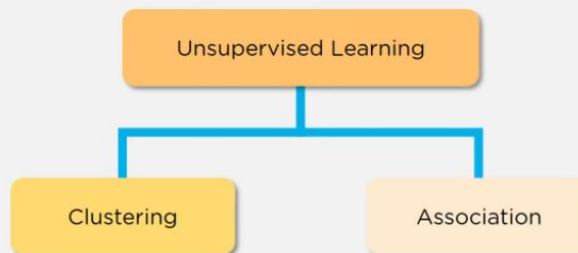
Analyzes



Makes predictions

13 Machine Learning Interview Questions

What are the unsupervised Machine Learning techniques?



13 Machine Learning Interview Questions

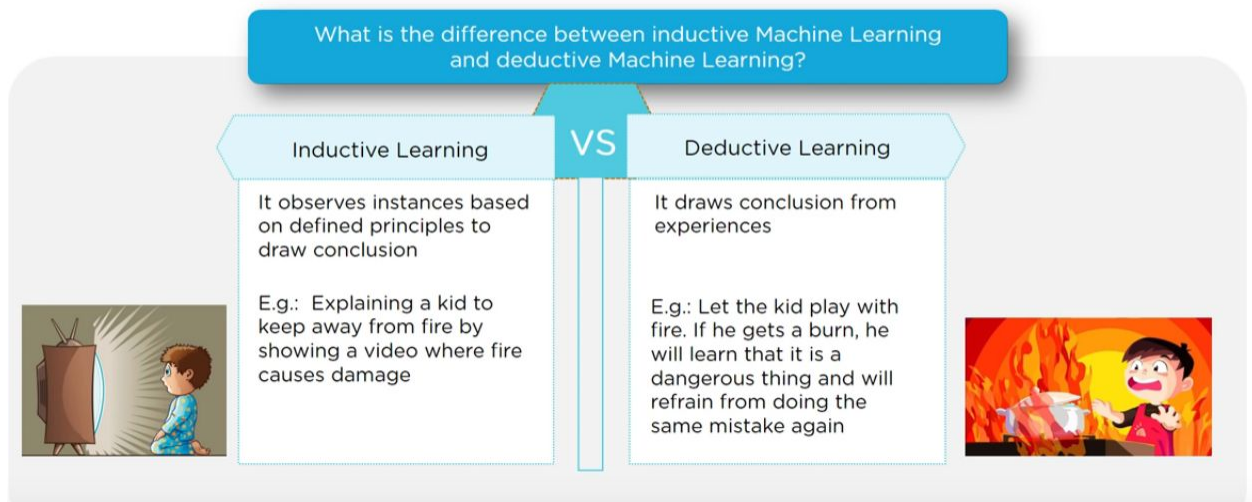
What are the unsupervised Machine Learning techniques?

Association

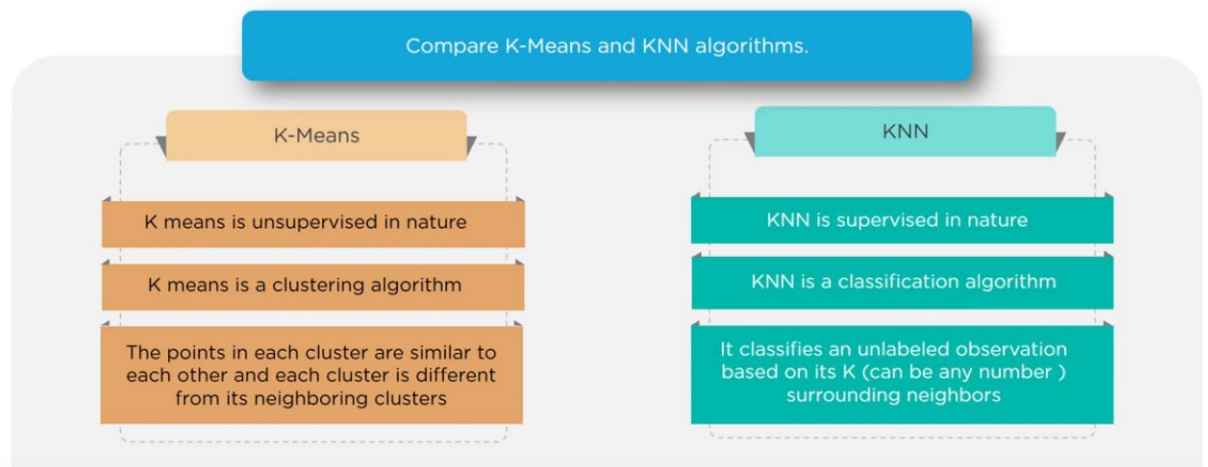
- ❑ In an Association problem, we identify patterns of associations between different variables or items
- ❑ In e-commerce websites, they're able to suggest other items for you to buy, based on the prior purchases that you've done, spending habits, items in your wish-list, other customers' purchase habits and so on.



15 Machine Learning Interview Questions



16 Machine Learning Interview Questions



17 Machine Learning Interview Questions

What is 'naive' in the Naive Bayes classifier?

- ❑ It's called "naive" because it makes assumptions that may or may not turn out to be correct
- ❑ The algorithm assumes the presence of one feature of a class is not related to the presence of any other feature (absolute independence of features), given the class variable
- ❑ For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, irrespective of other features and this assumption may or may not be right (E.g. Apple matches the description too).



19 Machine Learning Interview Questions

How will you know which machine learning algorithm to choose for your classification problem?

- If accuracy is a concern, then one can test different algorithms and cross validate them.
- If the training dataset is small, one should use models that have low variance and high bias
- If the training dataset is large, one should use models with high variance and low bias.



Cross validate different algorithms



Small data

Algorithm with low bias/high variance



Large data

Algorithm with high bias/low variance

20 Machine Learning Interview Questions

How is Amazon able to recommend other things to buy? How does it work?

- ❑ Once the user buys something from Amazon, it stores that purchase data for future references and finds products that are most likely to be also bought
- ❑ This is possible because of the **Association algorithm** which can identify patterns in a given data



21 Machine Learning Interview Questions

When will you use classification over regression?

Classification is used when your target variable is **Categorical** in nature. While Regression is used when your target variable is **Continuous** in nature. Both belong to the category of **Supervised Machine Learning Algorithms**.

Classification problems could be estimating the Gender of a person, the type of color, if the result is True or False, etc.

Regression problems could be estimating sale and price of a product, predicting sports score, amount of rainfall, etc.

22 Machine Learning Interview Questions

How will you design an email spam filter?

Building a spam filter involves the following processes:

- Email spam filter will be fed with thousands of emails
- Each of these emails will already have a label - 'spam' or 'not spam'
- The Supervised Machine Learning algorithm will then figure out which type of emails are being marked as spam based on spam words like lottery, free offer, no money, full refund, etc



Thousands of emails



Spam/Not Spam



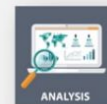
Learns

22 Machine Learning Interview Questions

How will you design an email spam filter?

Building a spam filter involves the following processes:

- The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like **Decision Trees** and **SVM** to figure out how likely it is that the email is spam
- If the likelihood, or probability, is high, it will label it as spam and the email won't hit your inbox
- Based on the accuracy of each model, we will use the algorithm with the highest accuracy after testing all the models



Statistical Analysis

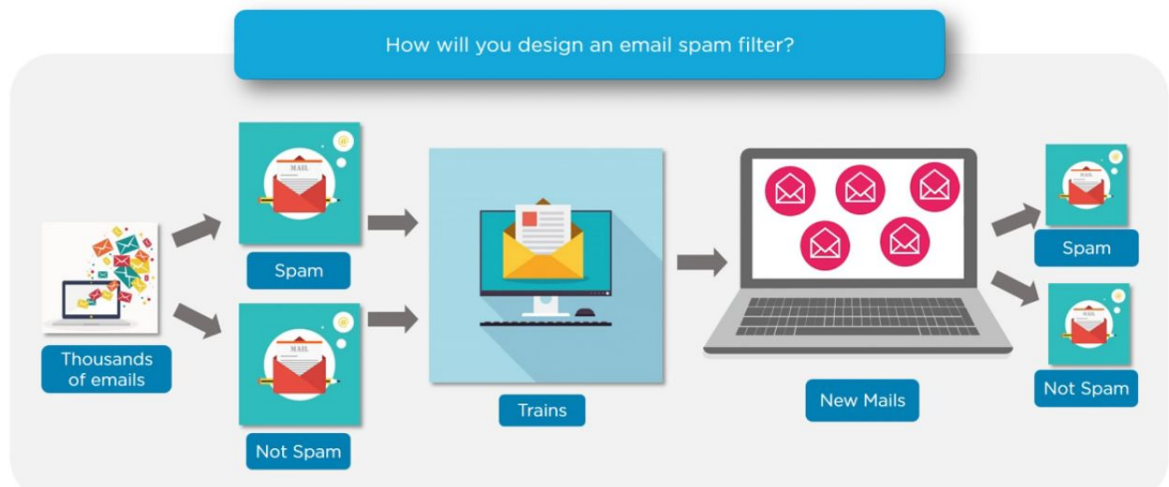


Spam



Not Spam

22 Machine Learning Interview Questions



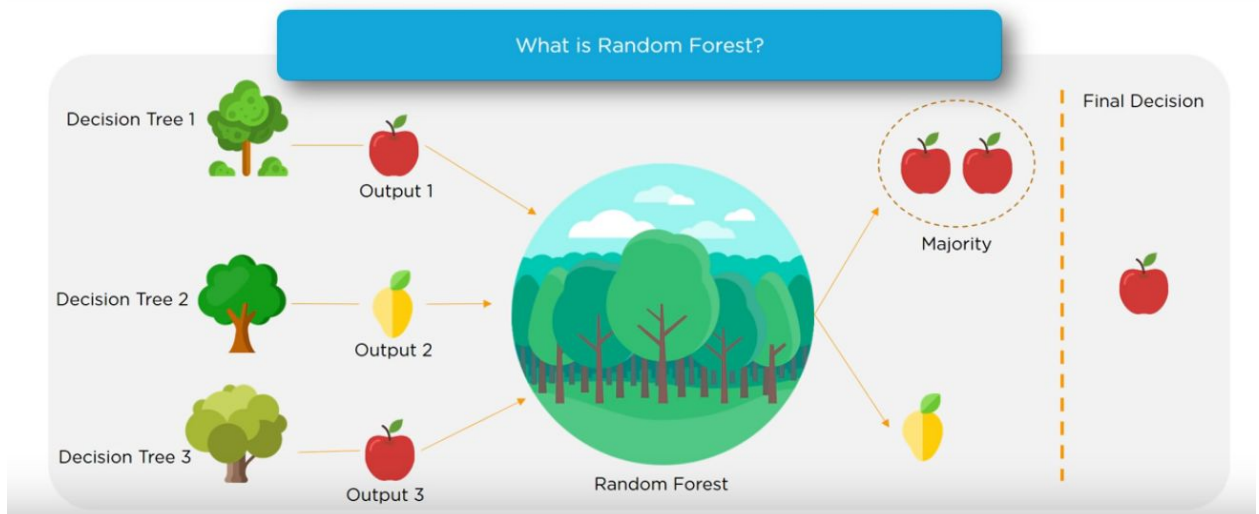
23 Machine Learning Interview Questions

What is Random Forest?

Random Forest is a Supervised Machine Learning Algorithm that is generally used for classification problems

Random forest operates by constructing multiple Decision Trees during training phase. The Decision of the majority of the trees is chosen by the random forest as the final decision

23 Machine Learning Interview Questions

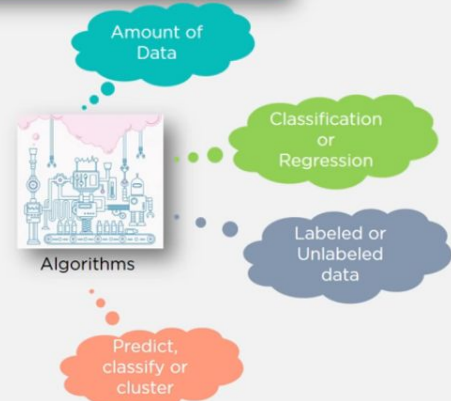


24 Machine Learning Interview Questions

Considering the long list of Machine Learning algorithm, given a data set, how do you decide which one to use?

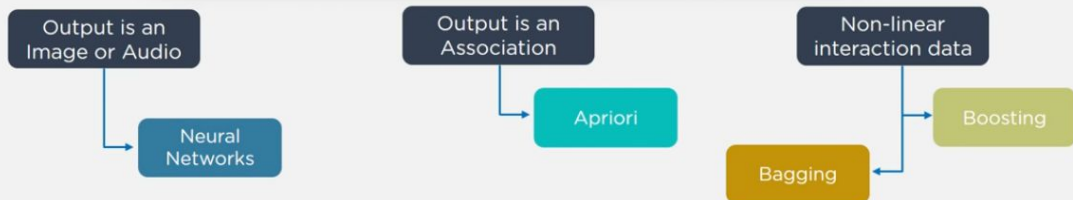
Choosing an algorithm depends on following questions:

- ❑ How much data do you have and is it continuous or categorical?
- ❑ Is the problem a classification, association, clustering or regression?
- ❑ Predefined variables (labeled), unlabeled or mix?
- ❑ What is the goal?



24 Machine Learning Interview Questions

Considering the long list of Machine Learning algorithm, given a data set, how do you decide which one to use?



There is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

25 Machine Learning Interview Questions

What is bias and variance in a Machine Learning model?

Bias in a Machine Learning model occurs when the predicted values are farther from the actual values

Low bias indicates a model where the prediction values are very close to the actual ones

Underfitting: High bias can cause an algorithm to miss the relevant relations between features and target outputs



High Bias
Low Variance

25 Machine Learning Interview Questions

What is bias and variance in a Machine Learning model?

Variance refers to the amount the target model will change when trained with different training data

For a good model, variance should be minimized

Overfitting: High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs



High Variance
Low Bias

26 Machine Learning Interview Questions

What's the trade-off between bias and variance?



- Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance
- You don't want either high bias or high variance in your model

26 Machine Learning Interview Questions

What's the trade-off between bias and variance?



High Bias
Low Variance

High Bias and Low Variance algorithms train models that are **consistent**, but inaccurate on average

High Variance and Low Bias algorithms train models that are **accurate** but inconsistent



High Variance
Low Bias

We need to find a balance of Bias and Variance so as to **minimize** the **total error**

27 Machine Learning Interview Questions

Define precision and recall.

Precision is the ratio of a number of events you can correctly recall to a number all events you recall (mix of correct and wrong recalls)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

In any 10 events, if you answer 10 times in which 8 events are correct and 2 events are wrong



27 Machine Learning Interview Questions

Define precision and recall.

Recall is the ratio of a number of events you can recall to the number of total events

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

If you can recall all 10 events, then, your recall ratio is 1.0 (100%)

100%

If you can recall 7 events, your recall ratio is 0.7 (70%)

70%



28 Machine Learning Interview Questions

What is pruning in decision trees and how is it done?



- ❑ Pruning is a technique in Machine Learning that reduces the size of **decision trees**
- ❑ It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting

28 Machine Learning Interview Questions

What is pruning in decision trees and how is it done?

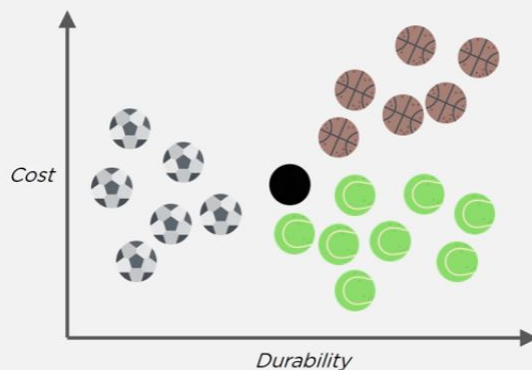


There is a popular pruning algorithm called **Reduced error pruning**

- ❑ Starting at the leaves, each node is replaced with its most popular class
- ❑ If the prediction accuracy is not affected then the change is kept
- ❑ Reduced error pruning has the advantage of **simplicity and speed**

30 Machine Learning Interview Questions

Explain K Nearest Neighbor algorithm.



K Nearest Neighbors algorithm works in a way that a new data point is assigned to a neighboring group it is most similar to.

In K Nearest Neighbors, K can be an integer greater than 1. So, for every new data point we want to classify, we compute to which neighboring group it is closest to.

Machine Learning pipeline

The following diagram shows an example pipeline:

