1. What kind of learning is required to group a set of unlabelled dolls as Barbie, Frozen, and Others
2. Unsupervised learning c) Supervised learning
3. Reinforcement learning d) None of these
4. Cost function is also known as
5. Squared error function c) Mean square error
6. Least square estimator d) both a) and b)
7. If the output of a supervised model is a categorical output then it is a
8. Regression c) Classification
9. Prediction d) none of these
10. One of the methods of overcoming ‘overfitting’ phenomenon is called as –
11. Regularization c) Segmentation
12. Fragmentation d) None of these
13. Which of the following machine learning algorithm is based upon the idea of bagging?
14. Decision tree c) Classification
15. Random forest d) regression
16. Choose a disadvantage of decision trees among the following.
17. Decision trees are robust to outliers
18. Factor Analysis
19. Decision tree are porn to overfitt
20. All of the above
21. Which of the following is not a supervised learning
22. PCA c) Naïve basian
23. Decision tree d) regression
24. \_\_\_\_\_ is a part of machine learning that works with neural network
25. Artificial Intelligence c) both a and b
26. Deep Learning d) None of these
27. The unsupervised learning problems can be grouped as
28. Clustering c) regression
29. Association d) both a and b
30. A model evaluation technique which is collection of methods used to examine the underlying constructs influence of the response of variable is
31. Factor analysis c) ICA
32. PCA d) singular value composition

[Supervised vs Unsupervised Learning - Javatpoint](https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning)

Answer any three of the following 3 X 5 = 15

1. Define machine learning. How is supervised learning different from unsupervised learning techniques? 2+3

Answer: difference b/w supervised and unsupervised learning

|  |  |
| --- | --- |
| **Supervised Learning** | **Unsupervised Learning** |
| Supervised learning algorithms are trained using labelled data. | Unsupervised learning algorithms are trained using unlabelled data. |
| Supervised learning model takes direct feedback to check if it is predicting correct output or not. | Unsupervised learning model does not take any feedback. |
| Supervised learning model predicts the output. | Unsupervised learning model finds the hidden patterns in data. |
| In supervised learning, input data is provided to the model along with the output. | In unsupervised learning, only input data is provided to the model. |
| Supervised learning model produces an accurate result. | Unsupervised learning model may give less accurate result as compared to supervised learning. |
| Supervised learning can be categorized in **Classification** and **Regression** problems. | Unsupervised Learning can be classified in **Clustering** and **Associations** problems. |

2+3

1. Define standard deviation, variance, and Covariance. 5

Answer:

Variance

According to layman’s words, the variance is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as ‘σ2’.

Properties of Variance

* It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.
* Variance always has squared units. For example, the variance of a set of weights estimated in kilograms will be given in kg squared. Since the population variance is squared, we cannot compare it directly with the mean or the data themselves.

Standard Deviation

The spread of statistical data is measured by the standard deviation. Distribution measures the deviation of data from its mean or average position. The degree of dispersion is computed by the method of estimating the deviation of data points. It is denoted by the symbol, ‘σ’.

Properties of Standard Deviation

* It describes the square root of the mean of the squares of all values in a data set and is also called the root-mean-square deviation.
* The smallest value of the standard deviation is 0 since it cannot be negative.
* When the data values of a group are similar, then the standard deviation will be very low or close to zero. But when the data values vary with each other, then the standard variation is high or far from zero

Covariance:

In mathematics and statistics, covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

* **Positive covariance**: Indicates that two variables tend to move in the same direction.
* **Negative covariance**: Reveals that two variables tend to move in inverse directions.

1. The marks obtained by a student are dependent on his/her study time. Given the study time in minutes and marks out of 2000, find the relationship between study time and marks using the concept of linear regression. Also predict the marks for a student if he/she studied for 790 minutes. 5

|  |  |  |
| --- | --- | --- |
| S. No | Study Time (min) | Marks Obtained |
| 1 | 350 | 520 |
| 2 | 1070 | 1600 |
| 3 | 630 | 1000 |
| 4 | 890 | 850 |
| 5 | 940 | 1350 |
| 6 | 500 | 490 |
|  |  |  |

1. How is logistic regression different from linear regression? Why does logistic regression come under the category of classification problems?

Answer:

|  |  |
| --- | --- |
| Linear Regression | Logistic Regression |
| Linear regression is used to predict the continuous dependent variable using a given set of independent variables. | Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables. |
| Linear Regression is used for solving Regression problem. | Logistic regression is used for solving Classification problems. |
| In Linear regression, we predict the value of continuous variables. | In logistic Regression, we predict the values of categorical variables. |
| In linear regression, we find the best fit line, by which we can easily predict the output. | In Logistic Regression, we find the S-curve by which we can classify the samples. |
| Least square estimation method is used for estimation of accuracy. | Maximum likelihood estimation method is used for estimation of accuracy. |
| The output for Linear Regression must be a continuous value, such as price, age, etc. | The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc. |
| In Linear regression, it is required that relationship between dependent variable and independent variable must be linear. | In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable. |

1. Apply k-means algorithm in given data for k=2. Use C1(80) and C2(250) as initial cluster centres. Data: 234,123, 456, 23, 34, 56, 78, 90, 150, 116, 117, 118, 199.

Answer any three of the followings 3 X 15 =45

1. What is Bayes’ theorem? How does a naïve Bayes classifier help in classifying the output class? How is the Bayesian belief network different from naïve Bayes classifier? Predict a class label using naïve Bayes classification, given this sample training data. The tuple you have to classify is

X= (age=youth, income=medium, student=yes, credit=fair)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Age | Income | Student | Credit | Buy |
| 1 | <30 | High | No | Fair | No |
| 2 | <30 | High | No | Excellent | No |
| 3 | 31-40 | High | No | Fair | Yes |
| 4 | >40 | Medium | No | Fair | Yes |
| 5 | >40 | Low | Yes | Fair | Yes |
| 6 | >40 | Low | Yes | Excellent | No |
| 7 | 31-40 | Low | Yes | Excellent | Yes |
| 8 | <30 | Medium | No | Fair | No |
| 9 | <30 | Low | Yes | Fair | Yes |
| 10 | >40 | Medium | Yes | Fair | Yes |
| 11 | <30 | Medium | Yes | Excellent | Yes |
| 12 | 31-40 | Medium | No | Excellent | Yes |
| 13 | 31-40 | High | Yes | Fair | Yes |
| 14 | >40 | Medium | No | Excellent | No |

2+3+3+7

Answer:

Bayes’ theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of “causes”. For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability. In this article, let us discuss the statement and proof for Bayes theorem, its derivation, formula, and many solved examples.

**Naive Bayes Classifier:🡪**

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.  The class with the highest probability is considered as the most likely class. This is also known as **Maximum A Posteriori (MAP)**.

**The MAP for a hypothesis is:**

**MAP(H)**  
= max( P(H|E) )  
=  max( (P(E|H)\*P(H))/P(E))  
= max(P(E|H)\*P(H))

Difference B/w Bayesian belief network different from naïve Bayes classifier :--/

|  |  |
| --- | --- |
| Bayesian belief network | naïve Bayes classifier |
| A Bayesian network models relationships between features in a very general way. | A Naive Bayes classifier is a simple model that describes particular class of Bayesian network - where all of the features are class-conditionally independent. Because of this, there are certain problems that Naive Bayes cannot solve |
| Naive Bayes is just a restricted/constrained form of a general Bayesian network where you enforce the constraint that the class node should have no parents and that the nodes corresponding to the attribute variables should have no edges between them | Naive Bayes is a constrained form of a more general Bayesian network, this paper also talks about why Naive Bayes can and does outperform a general Bayesian network in classification tasks. |

**Advantage and Disadvantage of naïve bays classification**:

**Advantages of Naive Bayes**

The Naive Bayes is a popular algorithm due to its following advantages:

* This algorithm works very fast and can easily predict the class of a test dataset.
* You can use it to solve multi-class prediction problems as it’s quite useful with them.
* Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.
* If you have categorical input variables, the Naive Bayes algorithm performs exceptionally well in comparison to numerical variables.
* It can be used for Binary and Multi-class Classifications.
* It effectively works in Multi-class predictions.

Now let’s go through the disadvantages of Naive Bayes classifier MCQ**.**

**Disadvantages of Naive Bayes**

* If your test data set has a categorical variable of a category that wasn’t present in the training data set, the Naive Bayes model will assign it zero probability and won’t be able to make any predictions in this regard. This phenomenon is called ‘Zero Frequency,’ and you’ll have to use a smoothing technique to solve this problem.
* This algorithm is also notorious as a lousy estimator. So, you shouldn’t take the probability outputs of ‘predict\_proba’ too seriously.
* It assumes that all the features are independent. While it might sound great in theory, in real life, you’ll hardly find a set of independent features.

After understanding these disadvantages of Naive Bayes classifier MCQ, you can now better understand this algorithm’s applications.

**Applications of Naive Bayes Algorithm**

As you must’ve noticed, this algorithm offers plenty of advantages to its users. That’s why it has a lot of applications in various sectors too. Here are some applications of Naive Bayes algorithm:

* As this algorithm is fast and efficient, you can use it to make real-time predictions.
* This algorithm is popular for multi-class predictions. You can find the probability of multiple target classes easily by using this algorithm.
* Email services (like Gmail) use this algorithm to figure out whether an email is a spam or not. This algorithm is excellent for spam filtering.
* Its assumption of feature independence, and its effectiveness in solving multi-class problems, makes it perfect for performing Sentiment Analysis. Sentiment Analysis refers to the identification of positive or negative sentiments of a target group (customers, audience, etc.)
* Collaborative Filtering and the Naive Bayes algorithm work together to build recommendation systems. These systems use data mining and machine learning to predict if the user would like a particular resource or not.

1. What is decision tree? Define Information gain, gain ratio, and gini index. Apply the decision tree algorithm for the following dataset and find the optimal decision tree. Also predict the class label for the following Outlook = Sunny, Temp = Hot, Humidity = Normal, Windy = True, Play = ? 2+6+7

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S. No | **outlook** | **temp** | **humidity** | **windy** | **play** |
| 0 | sunny | hot | high | FALSE | no |
| 1 | sunny | hot | high | TRUE | no |
| 2 | overcast | hot | high | FALSE | yes |
| 3 | rainy | mild | high | FALSE | yes |
| 4 | rainy | cool | normal | FALSE | yes |
| 5 | rainy | cool | normal | TRUE | no |
| 6 | overcast | cool | normal | TRUE | yes |
| 7 | sunny | mild | high | FALSE | no |
| 8 | sunny | cool | normal | FALSE | yes |
| 9 | rainy | mild | normal | FALSE | yes |
| 10 | sunny | mild | normal | TRUE | yes |
| 11 | overcast | mild | high | TRUE | yes |
| 12 | overcast | hot | normal | FALSE | yes |
| 13 | rainy | mild | high | TRUE | no |

Answer:

Decision Tree is a Supervised ML that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

Information Gain, Gain Ratio and Gini Index are the three fundamental criteria to measure the quality of a split in Decision Tree.

**Entropy:🡪**

Entropy is the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity. It characterizes the impurity of an arbitrary class of examples.

*Entropy is the measurement of impurities or randomness in the data points.*

Here, if all elements belong to a single class, then it is termed as “Pure”, and if not then the distribution is named as “Impurity”.

**Information gain:🡪**

The concept of entropy plays an important role in measuring the information gain. However, “Information gain is based on the information theory”.

Information gain is used for determining the best features/attributes that render maximum information about a class. It follows the concept of entropy while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes.

Information gain computes the difference between entropy before and after split and specifies the impurity in class elements.

Information Gain = Entropy before splitting - Entropy after splitting

**Gain ratio🡪**

Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has. Formula of gini ratio is given by

Gain Ratio=Information Gain/Entropy

From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.

Be selected as splitting criterion, Quinlan proposed following procedure,

1. First, determine the information gain of all the attributes, and then compute the average information gain.
2. Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

**Gini index 🡪**

The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be “successful” or “failure” and hence conducts binary splitting only.

The degree of gini index varies from 0 to 1,

* Where 0 depicts that all the elements be allied to a certain class, or only one class exists there.
* The gini index of value as 1 signifies that all the elements are randomly distributed across various classes, and
* A value of 0.5 denotes the elements are uniformly distributed into some classes.

1. Define Confusion matrix with example and Calculate Precision, Recall, Specificity, and F-score. Why AUC (Area under the curve) is important? How to prevent overfitting in Machine Learning 4+5+3+3

Answer:

**How to prevent overfitting in Machine Learning:**

* **Cross-validation :** Cross-validation is a powerful preventative measure against overfitting.

The idea is clever: Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model.

* **Train with more data:** It won’t work every time, but training with more data can help algorithms detect the signal better. In the earlier example of modeling height vs. age in children, it’s clear how sampling more schools will help your model.
* **Regularization:** Regularization refers to a broad range of techniques for artificially forcing your model to be simpler.

The method will depend on the type of learner you’re using. For example, you could prune a decision tree, use dropout on a neural network, or add a penalty parameter to the cost function in regression.

* **Ensembling:** Ensembles are machine learning methods for combining predictions from multiple separate models. There are a few different methods for ensembling, but the two most common are: Bagging and boosting.
* **Early stopping**: Up until a certain number of iterations, new iterations improve the model. After that point, however, the model’s ability to generalize can weaken as it begins to overfit the training data.Early stopping refers stopping the training process before the learner passes that point.

1. What is the goal of SVM? What is the role of radial basis functions in separating nonlinear patterns? What are the advantages and disadvantages associated with SVM? Explain the working principals of bagging, bosting, and stacking. 3+ 4+ 3+5

Answer:

Goal: The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps −

First, SVM will generate hyperplanes iteratively that segregates the classes in best way.

Then, it will choose the hyperplane that separates the classes correctly.

**Advantage and Disadvantage:**

**Advantages of Support Vector Machine (SVM)**  
  
**1. Regularization capabilities:** SVM has L2 Regularization feature. So, it has good generalization capabilities which prevent it from over-fitting.  
  
**2. Handles non-linear data efficiently:** SVM can efficiently handle non-linear data using Kernel trick.  
  
**3. Solves both Classification and Regression problems:** SVM can be used to solve both classification and regression problems. SVM is used for classification problems while **SVR (Support Vector Regression)** is used for regression problems.  
  
**4. Stability:** A small change to the data does not greatly affect the hyperplane and hence the SVM. So the SVM model is stable.  
  
**Disadvantages of Support Vector Machine (SVM)**  
  
**1. Choosing an appropriate Kernel function is difficult:** Choosing an appropriate Kernel function (to handle the non-linear data) is not an easy task. It could be tricky and complex. In case of using a high dimension Kernel, you might generate too many support vectors which reduce the training speed drastically.   
  
**2. Extensive memory requirement:** Algorithmic complexity and memory requirements of SVM are very high. You need a lot of memory since you have to store all the support vectors in the memory and this number grows abruptly with the training dataset size.  
  
**3. Requires Feature Scaling:** One must do feature scaling of variables before applying SVM.  
  
**4. Long training time:** SVM takes a long training time on large datasets.  
  
**5. Difficult to interpret:** SVM model is difficult to understand and interpret by human beings unlike Decision Trees.

1. Write short notes on the following (any three) 3x 5
2. Reinforcement learning:-

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.

It is two type:

* **Positive Reinforcement**
* **Negative Reinforcement**

**Positive Reinforcement:**

The positive reinforcement learning means adding something to increase the tendency that expected behavior would occur again. It impacts positively on the behavior of the agent and increases the strength of the behavior.

**Negative Reinforcement:**

The negative reinforcement learning is opposite to the positive reinforcement as it increases the tendency that the specific behavior will occur again by avoiding the negative condition.

Applications of RL:

* Robotics:

RL is used in Robot navigation, Robo-soccer, walking, juggling, etc.

* Control:

RL can be used for adaptive control such as Factory processes, admission control in telecommunication, and Helicopter pilot is an example of reinforcement learning.

* Game Playing:

RL can be used in Game playing such as tic-tac-toe, chess, etc.

* Chemistry:

RL can be used for optimizing the chemical reactions.

* Business:

RL is now used for business strategy planning.

* Manufacturing:

In various automobile manufacturing companies, the robots use deep reinforcement learning to pick goods and put them in some containers.

* Finance Sector:

The RL is currently used in the finance sector for evaluating trading strategies.

1. Random forest:

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “the random subspace method”

*Key Benefits:*

* Reduce risks of overfitting
* Provide flexibility
* Easy to determine future importance

Applications of RF:

* Finance
* Healthcare
* E- commerce

1. KNN algorithm:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

Advantages of KNN Algorithm:

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.

1. K-Means clustering:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

**What is K-Means Algorithm?**

K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

The k-means [clustering](https://www.javatpoint.com/clustering-in-machine-learning) algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



**How does the K-Means Algorithm Work?**

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

1. **PCA:--**

**What is Principal Component Analysis?**

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

**What is a Principal Component**?

The Principal Components are a straight line that captures most of the variance of the data. They have a direction and magnitude. Principal components are orthogonal projections (perpendicular) of data onto lower-dimensional space.

Now that you have understood the basics of PCA, let’s look at the next topic on PCA in Machine Learning.

Applications of PCA in Machine Learning

* PCA is used to visualize multidimensional data.
* It is used to reduce the number of dimensions in healthcare data.
* PCA can help resize an image.
* It can be used in finance to analyze stock data and forecast returns.
* PCA helps to find patterns in the high-dimensional datasets.

**How does Principal Component Analysis Work?**

**1.** Normalized data

2. build the covariance matrix

3. find eigenvalue and eigenvector

4.  Sort the eigenvectors in highest to lowest order and select the number of principal components.

**Q: Features in Machine Learning:-**

Features are nothing but the **independent variables** in machine learning models. What is required to be learned in any specific machine learning problem is a set of these features (independent variables), coefficients of these features, and parameters for coming up with appropriate functions or models (also termed hyperparameters). The following represents a few examples of what can be termed as features of machine learning models:

* A model for predicting the risk of cardiac disease may have features such as the following:
  + Age
  + Gender
  + Weight
  + Whether the person smokes
  + Whether the person is suffering from diabetic disease, etc.

Features are of two types. They are following:

* **Continuous features**: Continuous features are numerical values that can take on any value within a certain range. This type of data is often used to represent things such as time, weight, income, temperature, etc. Continuous features are often used in machine learning applications, since they can provide a more detailed representation of data than discrete or categorical features. For example, imagine that you are trying to predict the weight of an animal based on its height.
* **Categorical or discrete features**: Categorical features are an important part of machine learning. Categorical data is data that can be divided into categories, such as “male” and “female” or “red” and “blue.” Categorical features can be used to help predict what category something belongs to, based on other features.

**Here are some characteristics of good features:**

* **Features must be found in most of the data samples**: Great features represent unique characteristics that can be applied across different types of data samples and are not limited to just one data sample. For example, can the “red” colour of the apple act as a feature? Not really. Because apples can be found in different colours. It might have happened that the sample of apples that was taken for evaluation contained apples of just a “red” color. If not found, we may end up creating models having **high bias**.
* **Features must be unique and may not be found prevalent with other (different) forms**: Great features are the ones that are unique to apple and should not be applicable to other fruits. The toughness characteristic of apples such as “hard to teeth” may not be a good feature. This is because guava can also be explained using this feature.
* **Features in reality**: There can be features that can be accidental in nature and is not a feature at all when considering the population. For example, in a particular sample of data, a particular kind of feature can be found to be prevalent. However, when multiple data samples are taken, the feature goes missing.

**Q: Matrix factorization:-**

This mathematical model helps the system split an entity into multiple smaller entries, through an ordered rectangular array of numbers or functions, to discover the features or information underlying the interactions between users and items.

Matrix factorization is a**simple embedding model**. Given the feedback matrix A ∈ R m × n, where m is the number of users (or queries) and n is the number of items, the model learns: A user embedding matrix U ∈ R m × d, where row i is the embedding for user i.

**Why or when use matrix factorization**

* **Reduced computation time while training.**This is one of the main reasons why you would want to use matrix factorization in highly dimensional datasets. It is faster and easier for your training algorithm (e.g. random forest, gradient boosting) to learn form compact information-dense features.
* **Building a recommender systems**.  
  By itself, the matrix factorization techniques such as singular value decomposition (SVD) can be used to build recommender systems.
* **Dimensionality reduction.**While dealing with high dimensional datasets, different transformation techniques can lead to different feature engineered datasets with huge dimensionality (e.g. for the same text data you can compute bag of words, n-grams, tf-idf, etc. see figure below). We can first reduce the dimensionality of each technique and then concatenate the datasets into a smaller dataset to train on.
* **As feature engineering.**We can compute the latent vectors on the levels of original features or on the transformed features (e.g. log) and have different training models. We could use each model in an ensemble.

**Q: Generative Model:-**

A **Generative Model** is a way of learning any kind of data distribution. It is used in **unsupervised machine learning** as a means to describe phenomena in data, enabling the computers to understand the real world.

In unsupervised machine learning, generative modelling algorithms process the training data and make reductions in the data. These models generally are run on neural networks and can come to naturally recognize the distinctive features of the data. The neural networks take these reduced fundamental understandings of real world data and then use them to model data that is similar or indistinguishable from real world data.

The **main aim** of all types of generative models is to **learn the true data distribution of the training set so that the new data points are generated with some variations**.

The two types of generative models are:

**Variational Autoencoder** (VAE)

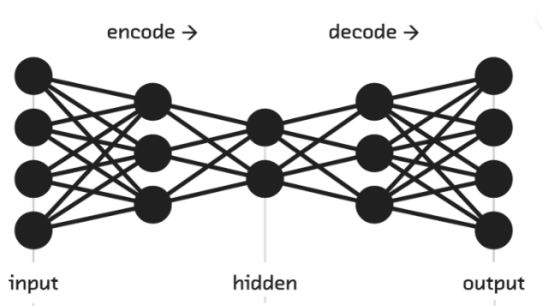
**Generative Adversarial Networks** (GAN)

**Variational Autoencoder (VAE)**

The definition of a VAE is that it “**provides probabilistic descriptions of observations in latent spaces**.” this means VAEs store latent attributes as probability distributions.

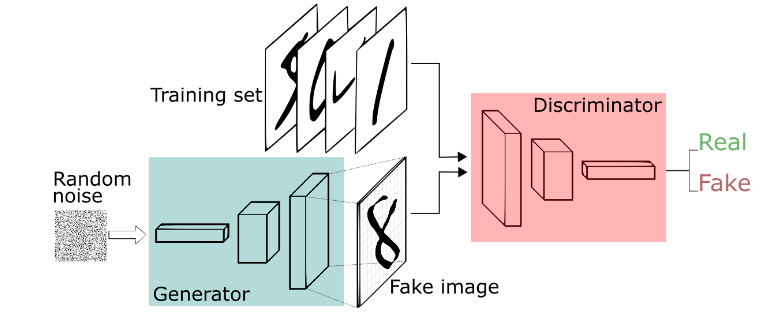
Let's first understand about the general architecure of VAE through the below diagram:

The **encoder** is responsible for the **mapping of input x to posteriors distributions** qθ(z∣x). similarly **p(x∣z) is then parametrized by the decoder**, combinely they form a generative network which takes latent variables z and parameters as inputs and projects them to data distributions pϕ(x∣z).



**Generative Adversarial Networks(GAN)**

Generative adversarial networks (GANs) are **deep neural net architectures comprised of two nets, pitting one against the other** (thus the “adversarial”).GAN is having a huge scope or potential because they can **learn to mimic any distribution of data**. That is, GANs can be taught to learn anything in any domain: images, music, speech, prose. They are **robot artists** in a sense, and their output is impressive.



**Q: Neural Networks:**

**Neural networks** are artificial systems that were inspired by biological neural networks. These systems learn to perform tasks by being exposed to various datasets and examples without any task-specific rules. The idea is that the system generates identifying characteristics from the data they have been passed without being programmed with a pre-programmed understanding of these datasets. Neural networks are based on computational models for threshold logic. Threshold logic is a combination of algorithms and mathematics. Neural networks are based either on the study of the brain or on the application of neural networks to artificial intelligence. The work has led to improvements in finite automata theory.

Propagation computes the input and outputs the output and sums the predecessor neurons function with the weight. The learning of neural network basically refers to the adjustment in the free parameters i.e. weights and bias. The learning rule modifies the weights and thresholds of the variables in the network. There are basically three sequence of events of learning process. These includes:

1. The neural network is simulated by an new environment.
2. Then the free parameters of the neural network is changed as a result of this simulation.
3. The neural network then responds in a new way to the environment because of the changes in its free parameters.

**The Advantages of Neural Networks**

* **Data:**One of the things that increased the popularity of the neural network is it gathers the massive amount of the data over the last years and decades. Neural networks give a better result when they gather all the data and information whereas traditional machines learning algorithms will reach a level, where more data doesn’t improve the performance.
* **Algorithms:** Neural networks are being popular due to the advancement made in the algorithms itself. These recent breakthroughs in the development of algorithms are mostly due to making them run much faster than before, which makes it possible to use more and more data.
* **Ability to work with incomplete information:** The data may generate performance after ANN training even with incomplete information. The performance loss here relies on the significance of the missing data.
* **Fault tolerance:** Corruption of one or more cells of ANN does not prevent it from generating output. This feature makes the networks fault-tolerant.
* **Dynamic:** Neural networks are good to model with nonlinear data with a large number of inputs; for example, images. It is reliable in an approach of tasks involving many features. It works by splitting the problem of classification into a layered network of simpler elements.
* **Parallel processing ability:** Artificial neural networks have numerical strength that can perform more than one job at the same time.
* **Computational power:** Computer power that is now accessible, enabling us to process more information. According to Ray Kurzweil, a leading figure in Artificial Intelligence, computational power is multiplied by a constant factor for each unit of time (e.g. doubling each year) rather than being incrementally added. This means there is an exponential increase in computational power.

**The Disadvantages of Neural Network**

**Black box:** One of the most distinguishing disadvantages of the neural network is their ‘’black box” nature. It means that we don’t know how and why the neural network came up with a certain output. For instance, when you put an apple’s picture into a neural network and predict it’s a cat, it’s very difficult to comprehend what led this forecast to come up with. When you have human interpretable characteristics, understanding the cause of your error is much simpler.

* **Amount of data:** Neural networks require much more data than any other traditional machine learning algorithms, as in at least thousands if not millions of labeled samples. This is a serious problem and many machine learning problems can be solved using fewer data in any other algorithms. This leads to the problem of over-fitting and generalization. The mode relies more on the training data and may be tuned to the data. Although there are some cases where the neural network has a deal with little data, most of the time they don’t.
* **Computationally expensive:** Neural networks are computationally expensive than any other traditional algorithms. Most of the traditional machine learning algorithms take much less time to train, ranging from a few minutes to a few hours or days. The amount of computational power that a Neural Network needs depends heavily on the size of your data, but also the depth and complexity of your network.
* **Determination of proper network structure:** There is no specific rule for determining the structure of a neural network. The appropriate network structure is achieved through experience and trial and error.
* **The duration of the network is unknown:** Reducing the network to a certain value of the sampling error implies completing the training. This value does not offer us the best outcomes.

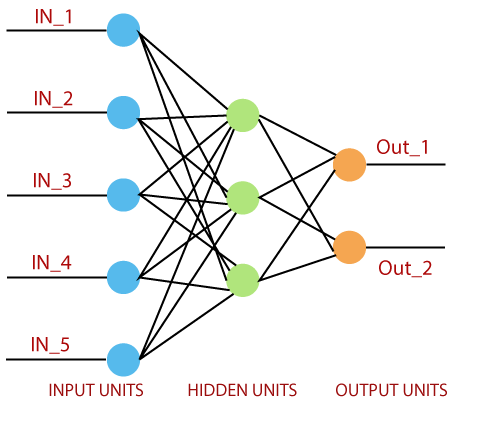
**Application:**

* Air traffic control
* Wagering on horse races, stock markets, sporting events, etc.
* Depth of river water could be predicted based on upstream reports, and the time and location of each report.

**Q: Draw a single layer perceptron of a neural network:-**

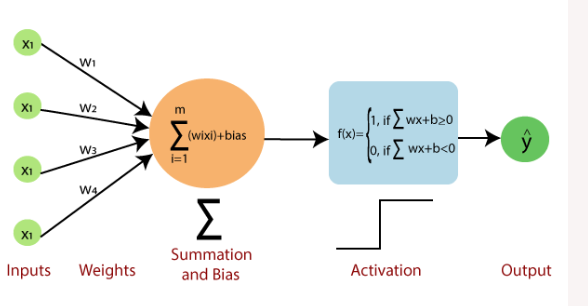
In this article we will go through a single-layer perceptron this is the first and basic model of the artificial neural networks. It is also called the feed-forward neural network. The working of the single-layer perceptron (SLP) is based on the threshold transfer between the nodes. This is the simplest form of ANN and it is generally used in the linearly based cases for the machine learning problems.

A perceptron is a neural network unit that does a precise computation to detect features in the input data. Perceptron is mainly used to classify the data into two parts. Therefore, it is also known as **Linear Binary Classifier**.



Perceptron uses the step function that returns +1 if the weighted sum of its input 0 and -1.

The activation function is used to map the input between the required value like (0, 1) or (-1, 1).



The perceptron consists of 4 parts.

* **Input value or One input layer:** The input layer of the perceptron is made of artificial input neurons and takes the initial data into the system for further processing.
* **Weights and Bias:**  
  **Weight:** It represents the dimension or strength of the connection between units. If the weight to node 1 to node 2 has a higher quantity, then neuron 1 has a more considerable influence on the neuron.  
  **Bias:** It is the same as the intercept added in a linear equation. It is an additional parameter which task is to modify the output along with the weighted sum of the input to the other neuron.
* **Net sum:** It calculates the total sum.
* **Activation Function:** A neuron can be activated or not, is determined by an activation function. The activation function calculates a weighted sum and further adding bias with it to give the result.

Now, We have to do the following necessary steps of training logistic regression-

* The weights are initialized with the random values at the origination of each training.
* For each element of the training set, the error is calculated with the difference between the desired output and the actual output. The calculated error is used to adjust the weight.
* The process is repeated until the fault made on the entire training set is less than the specified limit until the maximum number of iterations has been reached.

--------------------------------------------------------------------------------------------------------------------------------

**Q: Draw a multi-layer perception of neural network:?**

**Limitations of Single-Layer Perceptron:**

Well, there are two major problems:

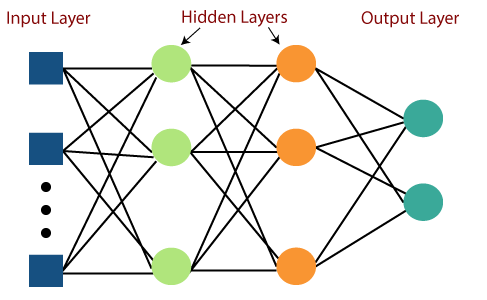
Single-Layer Percpetrons cannot classify non-linearly separable data points.

Complex problems, that involve a lot of parameters cannot be solved by Single-Layer Perceptrons.

**Multi-layer Perceptron:**

Multi-Layer perceptron defines the most complex architecture of artificial neural networks. It is substantially formed from multiple layers of the perceptron. TensorFlow is a very popular deep learning framework released by, and this notebook will guide to build a neural network with this library. If we want to understand what is a Multi-layer perceptron, we have to develop a multi-layer perceptron from scratch using Numpy.

The pictorial representation of multi-layer perceptron learning is as shown below-



MLP networks are used for supervised learning format. A typical learning algorithm for MLP networks is also called **back propagation's algorithm**.

A multilayer perceptron (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network. MLP is a deep learning method.

**Q: Difference b/w Machine learning and deep learning:**

| S. No. | Machine Learning | Deep Learning |
| --- | --- | --- |
| 1. | Machine Learning is a superset of Deep Learning | Deep Learning is a subset of Machine Learning |
| 2. | The data represented in Machine Learning is quite different as compared to Deep Learning as it uses structured data | The data representation is used in Deep Learning is quite different as it uses neural networks(ANN). |
| 3. | Machine Learning is an evolution of AI | Deep Learning is an evolution of Machine Learning. Basically, it is how deep is the machine learning. |
| 4. | Machine learning consists of thousands of data points. | Big Data: Millions of data points. |
| 5. | Outputs: Numerical Value, like classification of the score. | Anything from numerical values to free-form elements, such as free text and sound. |
| 6. | Its model takes less time in training due to its small size. | A huge amount of time is taken because of very big data points. |
| 7. | The results of an ML model are easy to explain. | The results of deep learning are difficult to explain. |