2/11/20 Udemy machine learning Tuesday

What is machine learning?

→ Machine learning is a form of artificial intelligence, where the machine are taught to do tasks without being explicitly programmed.

Characteristics of ML

→ we feed machines with some data and the machine learns the relationships amongst those data.

→ based on that relationships it can predict events for the data.

→ its improves performance with time, which means it has some intelligence like humans.

Applications

→ medical diagonosis, self driving cars, spam email filtering, stock market trading, automatic language translation, product recommendations, traffic predictions, image recognition, speech recognition etc.

Real world applications

→ detect cancer and brain tumors

→ Tesla: self driving cars

→ Gmail spam emails filtering

→ Google translators

→ Netflix, youtube, Amazon and other such companies Google maps face detection to

unlock a device, Siri, Cortana, Alexa all these technologies use machine learning in the background.

Methods in ML

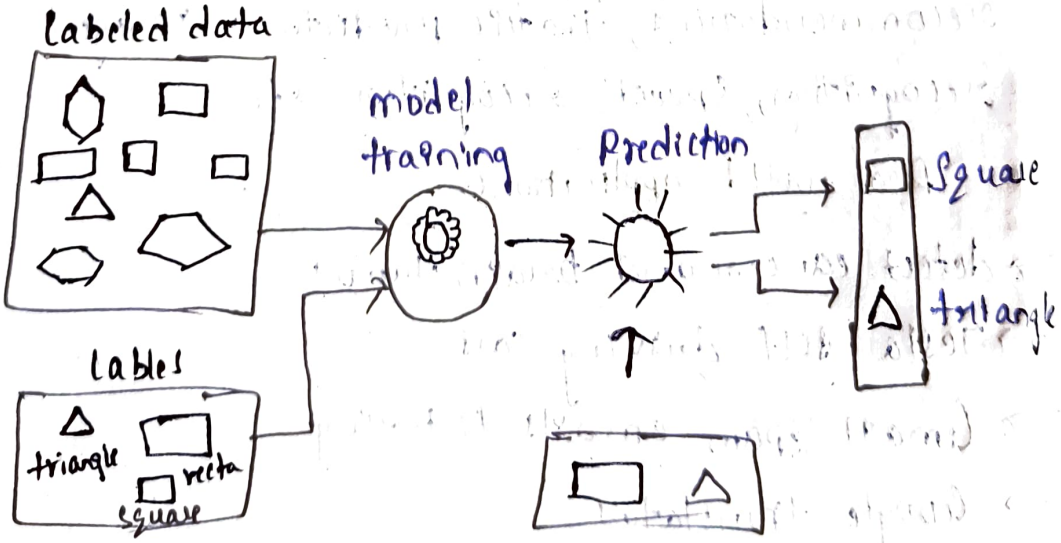
→ In ML tasks are generally classified into broad categories and these are based on how learning is received.

→ Two of the most widely adopted ML methods are supervised and unsupervised learning.

What is supervised learning

→ The term "supervise" refers to the act of watching and directing the completion of work, project & operation.

→ It uses labelled training data and a collection of training sample to estimate a function.



→ The computer is provided with example inputs that are labeled with their desired output.

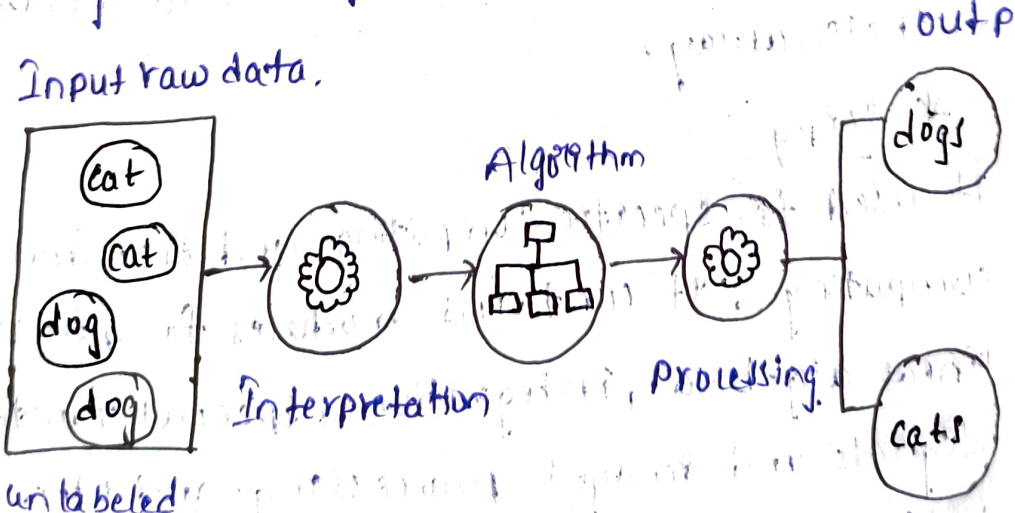
→ The purpose of this method is for the algorithm to be able to learn by comparing its actual output with the taught o/ps to find errors and modify the model accordingly.

→ Supervised learning therefore uses patterns to predict and put label values on the unlabeled data.

→ A common use case of supervised learning is to use historical data to predict statistically likely future events.

What is unsupervised learning

→ It is a kind of ML in which the training data is supplied to the system without any pre-assigned labels or values.



→ The unsupervised learning approach can discover hidden pattern within a datasets that are needed to classify raw data as

unlabeled.

→ ML methods that facilitate unsupervised learning are particularly valuable.

Python

→ Python is a widely used and efficient programming language that has lately become the language of choice for data scientists.

→ One of the reason behind the popularity of this language is that its well established and has robustly defined library such as numpy, cpt, matt, plot, lib pandas and CI kitlearn

* Numpy

→ It helps to deal with large multidimensional arrays and matrices, along with a large collection of high level mathematical function to operate on thele array.

* Scipy

→ It used for specific computing and technical computing. It contains modules for optimization linear algebra, integration, special functions, signals and images processing and other task common in science and engineering

* matplotlib

→ It is a plotting library and it is easy to use with numpy because its numeric is integrated with numpy.

* Pandas

→ pandas is designed for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

* scikit learn

→ It is a free s/w machine learning library for the python programming language. It features various classification, regression and clustering algorithm, and it is designed to interoperate with the python numerical and scientific library numpy and scipy.

Regression

→ It is a statistical method which predicts a relationship of a continous outcome based on the value of one or more input variables.

→ It is used in many disciplines such as statistics, economics, finance, investment industries data science and more

→ Some real world examples for regression analysis include predicting the price of a house given house

features, forecasting sales based on input parameters, predicting the weather and etc.

→ Regression is divided into 2 types

1. linear regression
2. nonlinear regression.

1. linear regression

The relationship between input variable and the output variable is linear

* Simple linear regression

only a single input variable and that single input variable is used to predict an output while,

* multiple linear regression

more than one input variable is involved

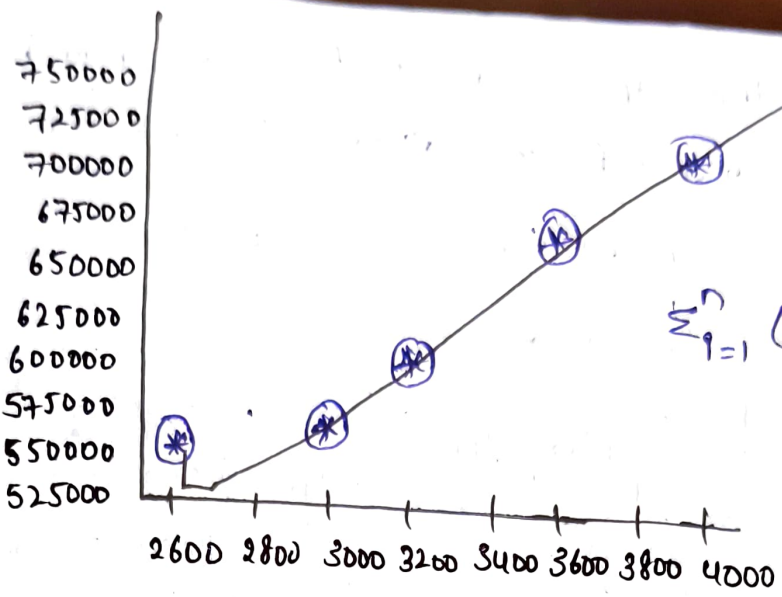
2. Nonlinear regression

output can be represented as a function of that is non-linear combination of the inputs.

*Working simple linear regression

area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000

using this data we will build a machine learning model that can predict the price of a 330 square



we can draw many line to fit that graph among those lines we take the best fit line

$y = mx + b$ where y = dependent variable

x = independent variable

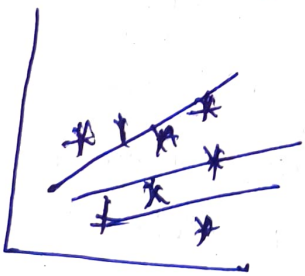
m = slope of line

b = y intercept

for the above example

price = m * area + b

* Working multiple linear regression



$y = A + B_1 x_1 + B_2 x_2 + B_3 x_3 + \dots$

where A is y intercept

B_1 = slope of line that represent the relation ship between the output and the first independent variable x_1,

→ B_1 indicates that if we increase the value of x_1 by one unit, then B_1 tells us that how

much the output value will be affected regarding B_2 B_3 and before they all work in a similar fashion

* model evaluation for regression

mean absolute error

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

n = no. of values
y_i = Actual value
$\hat{y}_i$ = predictive value

mean square error

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2$$

root mean square error

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

* classification

It is a predective modelling issue in machine learning where a class label is predicted for a given sample of input data.

→ It is a Supervised learning approach in which the computer program learns from the data given to it and makes new prediction or classification.

→ Here are some applications like filtering emails, speech and handwriting recognition, biometric fingerprint

→ Eg in fingerprint recognition system will recognize the fingerprint of the specific person as is

trained by different fingerprints of different persons
- so with the help of classification the system will identify the class of fingerprint that a specific person belongs to

→ The algorithm for classification are neural network regression and ~~case~~ k nearest neighbour

kNN Algorithm

→ k-nearest neighbour algorithm is classification method that utilize a set of labelled points to learn how to identify new ones

→ formula for calculation of Euclidean distance

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

→ choose a value for k

→ determine the distance of an unknown point from all other points

→ Now in training data set choose the k which are nearest to the unknown data point

→ Now predict the response of unknown datapoint

Example

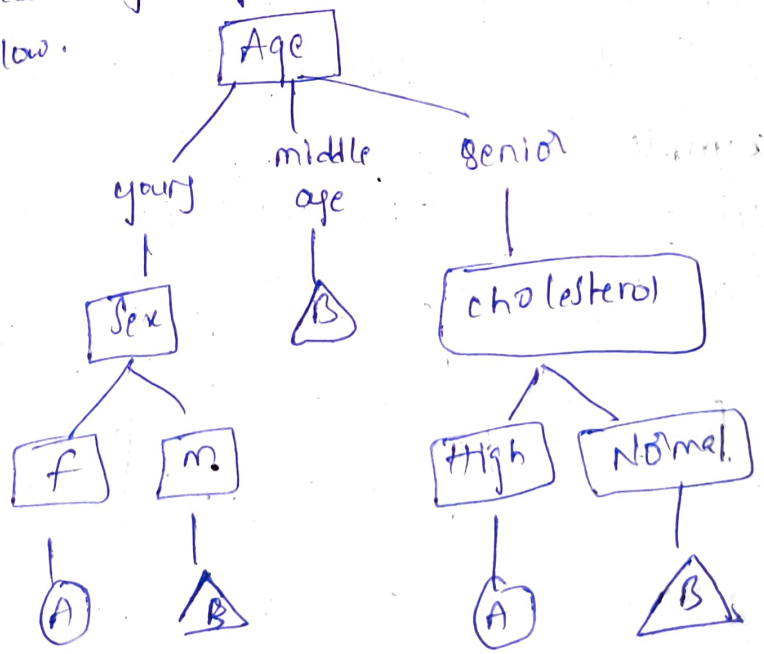
kNN classifier will be used in this example to classify the input image into cat or dog

* Steps to implement KNN

→ data pre processing like filtering
→ fitting KNN to the training set
→ predict the test result
→ Test accuracy of the result.

* Decision trees

→ It is a kind of supervised machine learning technique in which the data is being split continously based on parameters
→ It breaks down a dataset into smaller and smaller subsets
→ The final result is a tree with decision nodes and leaf nodes.
→ A decision node example age cholesterol has 2 or more branches.
→ The leaf node example middle edge represents a classification or decision.
→ The learning algorithm of decision tree is as follow.



→ 1st it choose a feature from the given dataset then it determines the importance of the feature in the segmentation of data.

→ data segmenting depend upon the best attribute. after that we repeat the above steps.

→ All those processes use some kind of criteria and measurements in order to make decision.

→ one of the such criteria is entropy.

what is entropy

→ It is the measure of the amount of uncertainty or randomness in the data.

→ Intuitively it tells us about the predictability of a certain event.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

$p(x)$ is a probability function of the event x to take place.

(Lowest value = 0 (no randomness)

Highest value = 1 (high randomness)

Eg consider a coin toss that has a probability of heads of 0.5 and a probability of tails of 0.5

→ Here the entropy is the highest possible sense.

→ There is no way of determining what the outcome might be

→ Alternatively consider a coin that has heads on both sides of the coin.

→ The entry of such an event can be predicted perfectly, since we know beforehand that it will always be heads.

→ In other words this event has no randomness Hence its entropy is zero.

→ In particular low values imply less uncertainty while high values imply high uncertainty.

→ So while choosing parameter for each node in a tree, the decision tree measures this values and choose the one with high certainty.

* Logistic Regression

→ logistic regression is used to predict the likelihood of a specific class or occurence

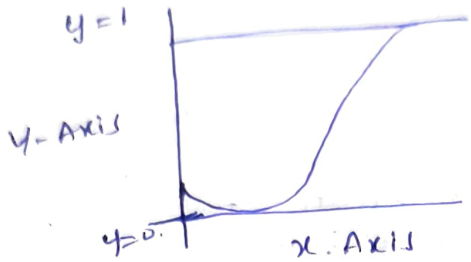
→ Here are some conditions to make use of logistic regression if the information is binary, if you require probabilistic outcome when a linear decision boundary is required Sure, these are the conditions

Applications

→ It is used to forecast survival in wounded patients in the medical profession to estimate

→ To estimate the likelihood of a person suffering a heart attack, to predict the likelihood of a procedure or a product failing and predict a hormones possibility of defaulting on a

lender



Steps to implement Logistic Regression

→ data preprocessing like following
→ Fitting KNN to the training set.
→ Predict the test result
→ Test accuracy of result.

Logistic Regression

→ we estimate the values of categorical variables using logistic regression

→ we discover the S-curve in logistic regression and use it to categorize the samples

→ It is not necessary to have a linear connection between the dependent and independent variables in logistic regression

→ There should be no collinearity between the independent variables in logistic regression

linear regression

→ we estimate the outcome of continous variables using linear regression

→ In linear regression we look for the best fit line, which allows us to predict the outcomes with ease.

→ It is necessary for the connection between the dependent and independent variables to be linear

→ There is a possibility of collinearity between the independent factors in linear regression.

*clustering

→ clustering is the task of grouping a set of objects in such a way that objects in the same group form a cluster.

→ It is an unsupervised learning technique as we can see in the diagram.

→ Arbitary colored points are being separated into classes by clustering.

→ Applications are pattern recognition, spatial data analysis, image processing, document classification, identification of similar entities, finding pattern of wheather behaviour

* use cases

→ we can use clustering for the following techniques like analyzing data from research, creating a summary, detecting noise, duplicate detection and soon.

k-means clustering

→ It is an unsupervised learning algorithm, which group the unlabeled data set into different clusters.

→ It is an iterative algorithm, that divides the unlabeled dataset into ~~k~~ different clusters in such a way that each dataset belongs to only one group that has similar properties

→ The leader is defines the no. of predefined clusters that need to be created in the process.

→ It allows us to cross the data into different group and in a convenient way.

→ This will allow it to have the category of group on the unlabeled dataset by itself and without the need of any ~~thing~~ training.

Steps for Algorithm

→ Partition clustering

→ without any cluster - internal structure, k - means divided data into non-overlapping groups.

→ within a cluster, examples are very similar.

* Elbow method

Q How do we determine the optimal value for k)

→ 1st we need to keep in mind that the performance of the k means clustering algorithm depends on the highly efficient clusters that it forms.

→ Hence choosing the optimal no. of clusters is a big critical task.

→ There are some different ways to find the optimal no. of clusters, but here we are discussing the most appropriate method to find the no. of clusters or value of k.

→ This method uses the concept of wcss value wcss stands for within cluster sum of square which defines the total variation within a cluster.

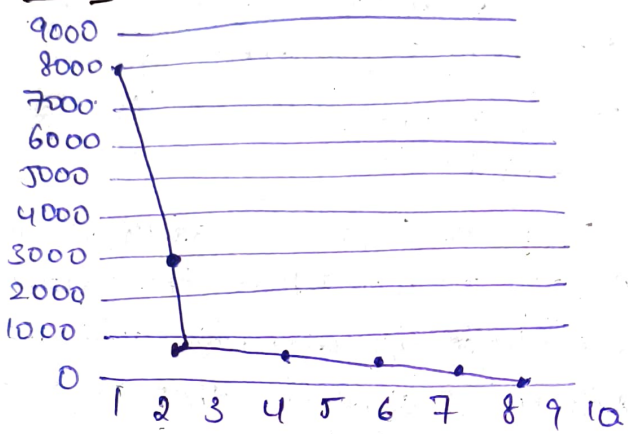
$$WCSS = \sum_{P_i \text{ in cluster } 1} \text{distance}(P_i C_1)^2 + \sum_{P_i \text{ in cluster } 2} \text{distance}(P_i C_2)^2 + \sum_{P_i \text{ in cluster } 3} \text{distance}(P_i C_3)^2$$

→ The elbow method is one of the most popular ways to find the optimal no. of clusters

→ In formula the sum of distance between each data point and its centroid squared.

→ To find the distance between datapoints we can use Euclidean distance or manhattan distance.

Steps :-

→ 

→ It executes the k-means clustering on given dataset for different k values (ranges from 1-10)

→ for each value of k, calculates the WCSS value

→ plots a curve between calculated WCSS values and the no. of clusters k.

→ The sharp point of bend or a point of the plot like an arm, then that point is considered as the best value of k.

Hierarchical clustering

→ It involves creating clusters that have predetermined ordering from top to bottom.

for example: all files unfold as on the hard disk organized by hierarchy.

There are 2 types of hierarchical clustering

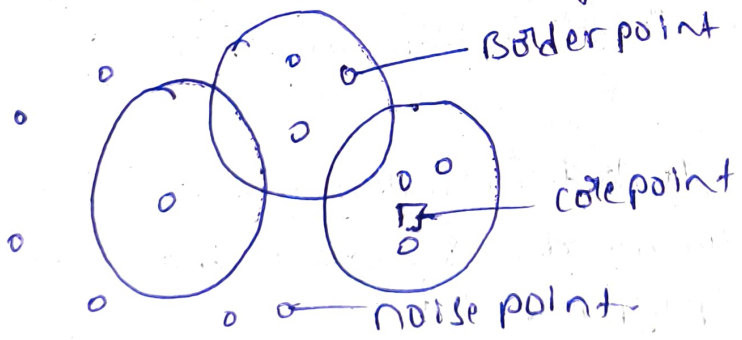
1. divisive clustering: It is top to down approach in which observation starts from large cluster and splits into smaller ones.
2. Agglomerative clustering: It is bottom to top approach in which observation start from many clusters and merged as one cluster in the end.

* density based clustering

→ It locates region of high density and separates outliers. So, the region of high density get separated from the region of low density.

what is DBSCAN

→ The most powerful attribute of DBSCAN algorithm is that it can find out any arbitary shap ecluster without getting effected by noise.



→ It depends on 2 parameters one is radius and other one is minimum points, which epsilon determined a specified ~~width~~ radius.

→ If that radius include enough points within it we can call it a dense area.

→ M determines the minimum no. of data points we want in a neighborhood to define a cluster.

→ the algorithm proceeds by arbitary picking up a point in the dataset.
→ If there are atleast m points within a radius of epsilon to the point, then we consider all these points to be part of the same cluster.

Advantages

→ It can discover arbitrarily shaped cluster,
→ find cluster surrounded by different cluster
→ Robust towards noise detection.

* Recommdation System

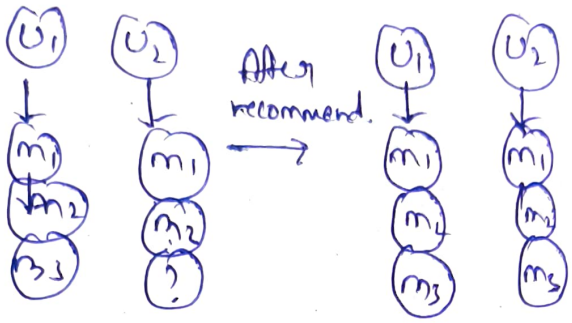
→ Recommender System algorithm aimed at suggesting relevent items to users.
Eg:- Items can be movies to watch, texts to read, products to buy or anything else depending on the industry.
→ The application of recommender system include movies, music, books, website and television program.
→ netflix, Amazon, LinkedIn, Twitter and facebook are using recommender systems in their software for different purpose.

Benefits

→ It increase customer engagement
→ Increased customer satisfaction by delivering relevant content.
→ Reduce the time of content searching

collaborative filtering Recommendation System

It predicts the interests of a user on a project by collecting preference.



Content based Recommender System

→ It utilizes Product features to recommend other product like what the user likes, based on the other user previous action or explicit feedback such as rating on products