- **Machine Learning With Python**

- **Welcome**
- Video 3 min
- In this course you'll learn how Machine Learning is used in many key fields and industries
- Bankers : Loan application approval , probability of default
- Bankers : Customer segmentation
- Websites , E-commerce , YouTube recommendation system

Regression
Classification
Scikit learn
SciPy
Cancer detection.
Predict customer churn
telecommunications unsubscribe

- Scikit learn library to estimate CO2 emission for car
- Customer churn
- Using jupyter lab

- **Introduction to Machine Learning**

Human cell
Thickness , shape , clump
Benign or malignant (cancer)
Cl

Steps
Clean data
Select proper algothim for building prediction model
Train models to understand patterns of benign or malignant
And can predict , new case

Make doctor
Task

Machine learning is sub field of computer science that gives computers ability to learn without being explicitly programmed

Cats Dogs image set ,
feature extraction
Machine learning model
Output classification

Learn understand differentiate process

Object recognition summarisation recommendation

Major machine learning technique

- **Regression / estimation**
  - Predicting continuous values
    - Price of house ,co2 estimates
- **Classification**
  - Predicting item class / category of a case
  - Ex : cancer cell benign or malignant , customer churn
- **Clustering**
  - Find the structure of data , summarisation
  - Customer segmentation in banking
  - Confine similar patients

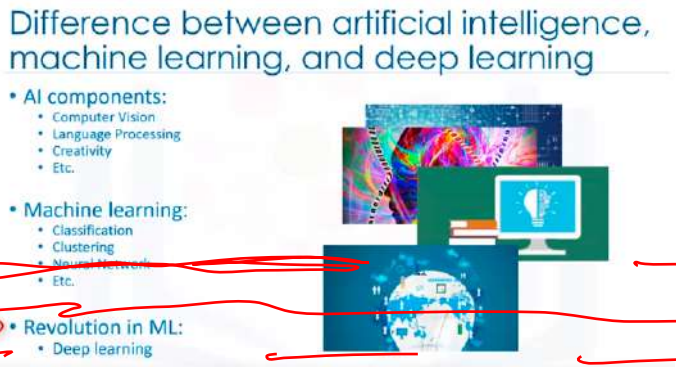
Which Machine Learning technique is proper for grouping of similar cases in a dataset, for example to find similar patients, or for customers

segmentation in a bank?

Clustering

- **Associations**
  - Associating frequent Co occurring items / events
  - Groceries items are bought together by a particular customer
- **Anomaly detection**
- Discover abnormal and unusual cases
- Credit card fraud detection
- **Sequence Mining**
- Predicting next events
- Clickstream in websites
- **Dimension Reduction**
- Reducing size of data
- **Recommendation systems**
- People preferences with others who have similar tastes
- Books and movies

AI
under that ML
under ML , DL



SciPy - numerical & Signal processing , high computation
Matplotlib
Visualisation library
Pandas - data manipulation analysis , data operations
Scikit learn - collection of algorithm for ML
FREE ML library
Have classification

Why Scikit is a proper library for Machine Learning (select all the options that are correct)?

Scikit-learn is a free machine learning library

that works with Numpy and Scipy.

Scikit-learn has most of machine learning algorithms.

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving

- **Python for Machine Learning**

Python preferred for DS
modules in python for ML
NumPy - math library ndim array
Images

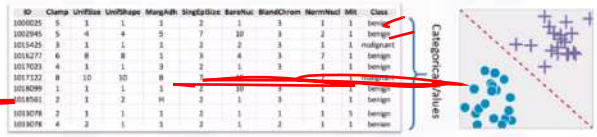
Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost.

scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)

clf.fit(X_train, y_train)

clf.predict(X_test)

from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))

import pickle
s = pickle.dumps(clf)
```

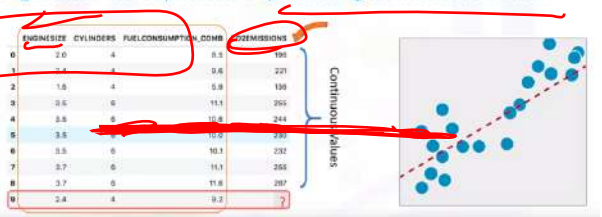
What is classification?

Classification is the process of predicting discrete class labels or categories.



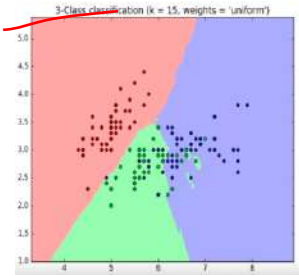
What is regression?

Regression is the process of predicting continuous values.



- **Supervised Vs Unsupervised Learning**
-

What is supervised learning?

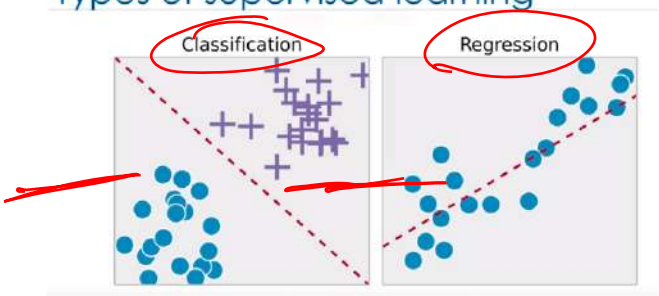


We "teach the model," then with that knowledge, it can predict unknown or future instances.

Teaching the model with labeled data

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Types of supervised learning



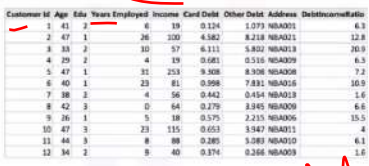
Which technique/s is/are considered as Supervised learning?

- [ ] Classification

Correct

- [ ] Regression

What is unsupervised learning?



Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering ★

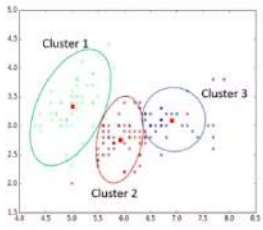
ALL OF THIS DATA IS UNLABELED

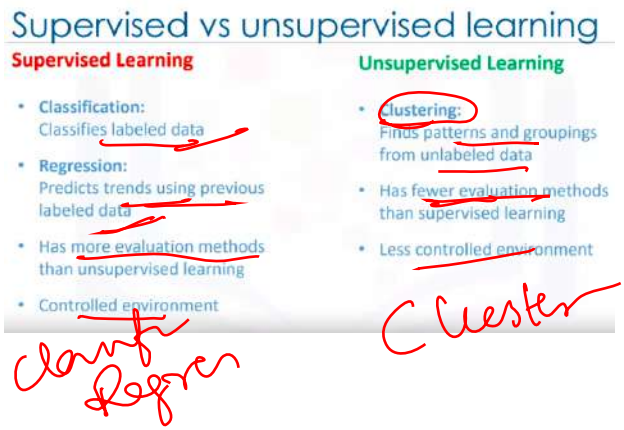
The model works on its own to discover information.

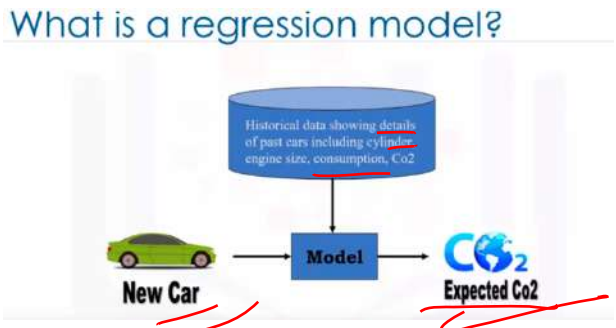
What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:

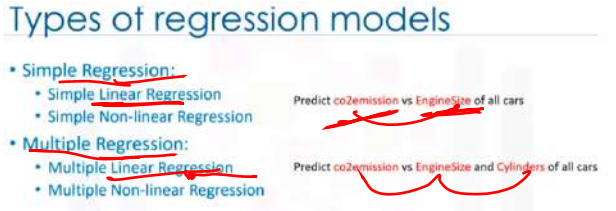
- Discovering structure
- Summarization
- Anomaly detection ★

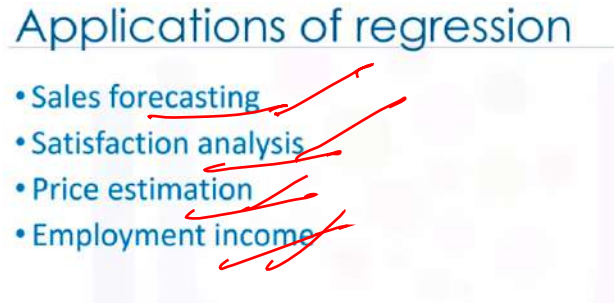






- **Quiz : Intro to Machine Learning**
- Supervised learning deals with unlabeled data, while unsupervised learning deals with labelled data.
- False
- The "Regression" technique in Machine Learning is a group of algorithms that are used for
- Predicting a continuous value; for example predicting the price of a house based on its characteristics.





- When comparing Supervised with Unsupervised learning, is this sentence True or False?
- In contrast to Supervised learning, Unsupervised learning has more models and more evaluation methods that can be used in order to ensure the outcome of the model is accurate
- False

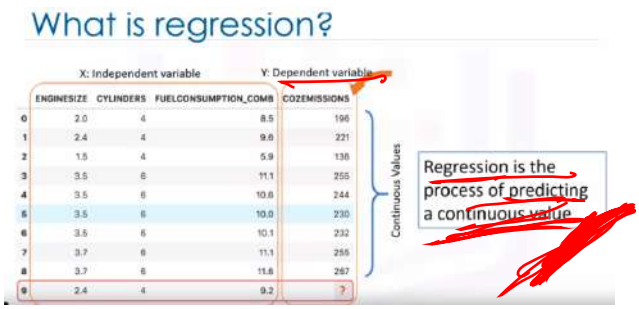
sample application of regression?

Forecasting rainfall amount for next day.

- **Linear Regression**

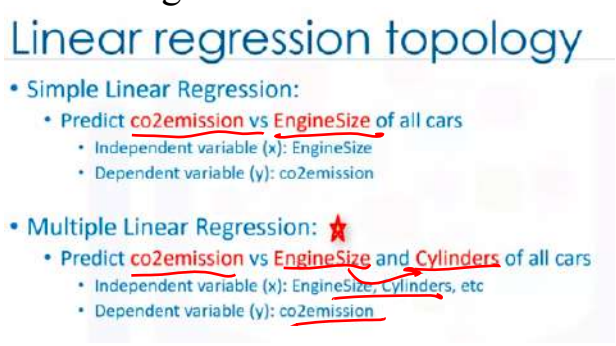
Linear or non linear based on relation between dependent and independent variable

- **Introduction to Regression**

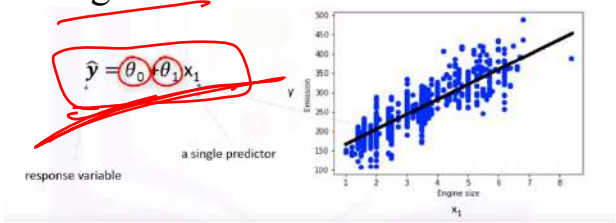


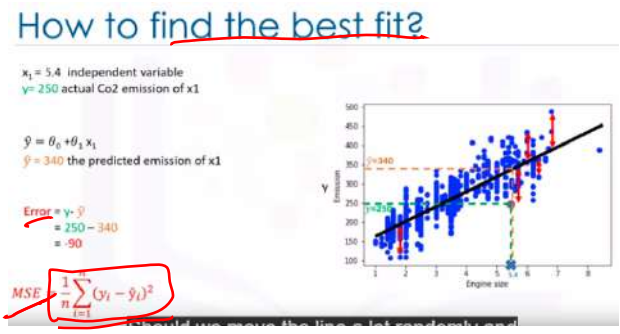
- **Simple Linear Regression**

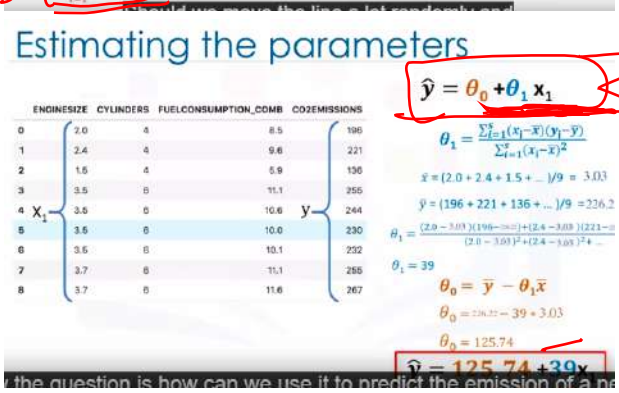
- In this video, we'll be covering linear regression. You don't need to know any linear algebra to understand topics in linear regression. This high-level introduction will give you enough background information on linear regression to be able to use it effectively on your own problems. So let's get started

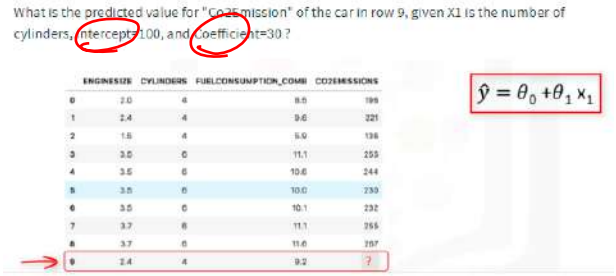


Fitting line

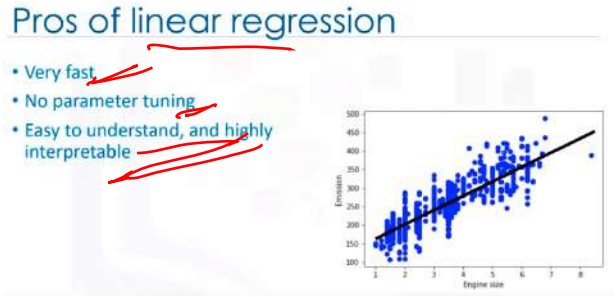




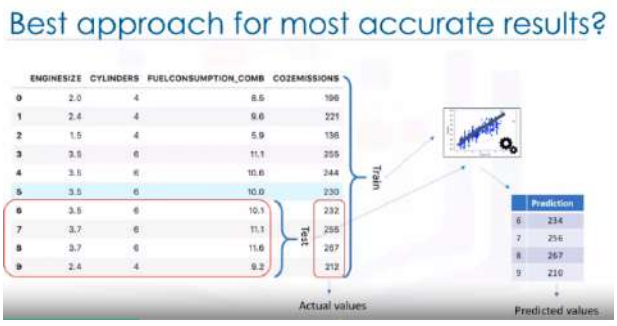




220

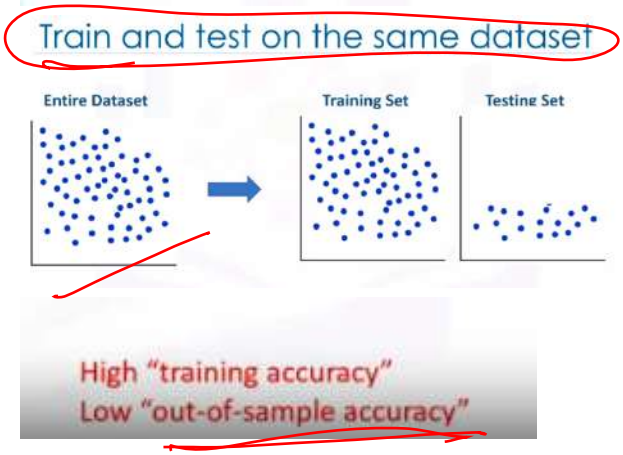


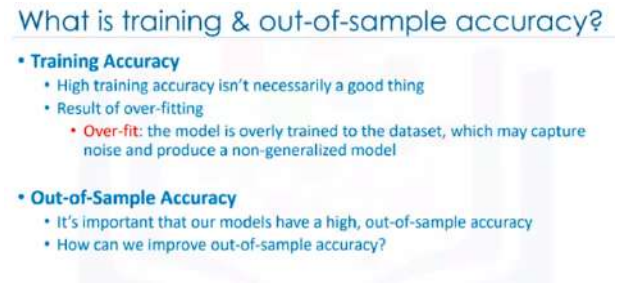
- **Model evaluation in Regression Models**



$$Error = \frac{(232 - 234) + (255 - 256) + \ldots}{4}$$

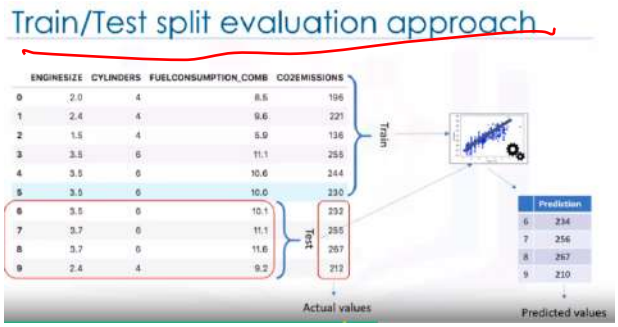
$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

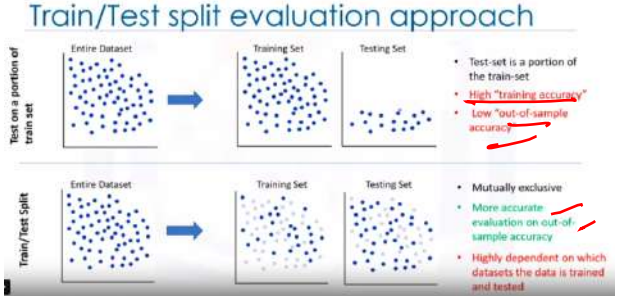


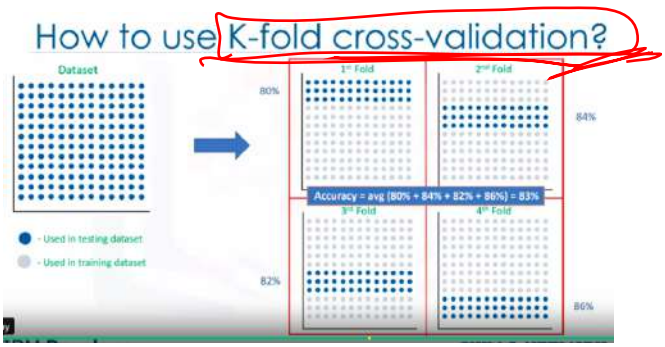


Having a high training accuracy may result in an 'over-fit' of the data.

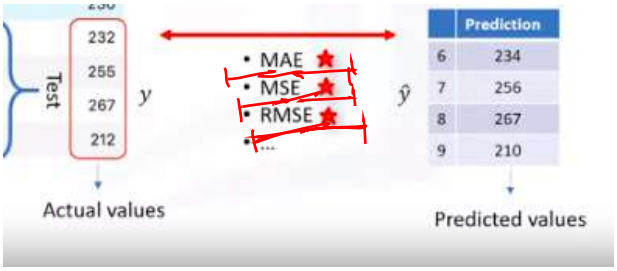
If a model is overly trained to the dataset, it may capture noise and produce a non-generalized model.





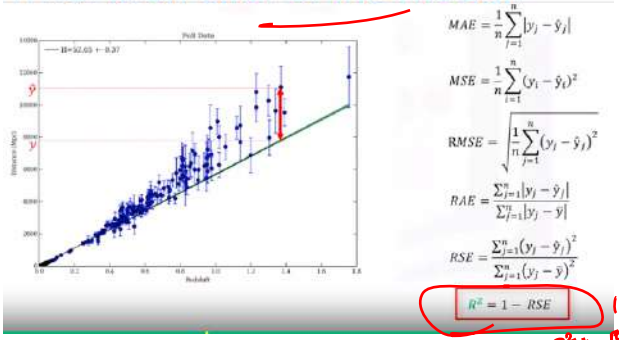


- **Evaluation Metrics in Regression Models**



Error of model

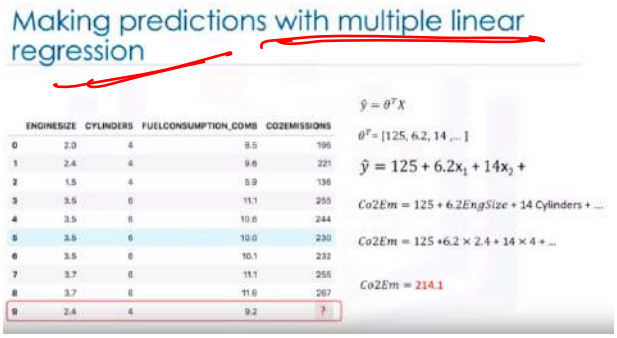
The difference between the data points and the trend line generated by the algorithm



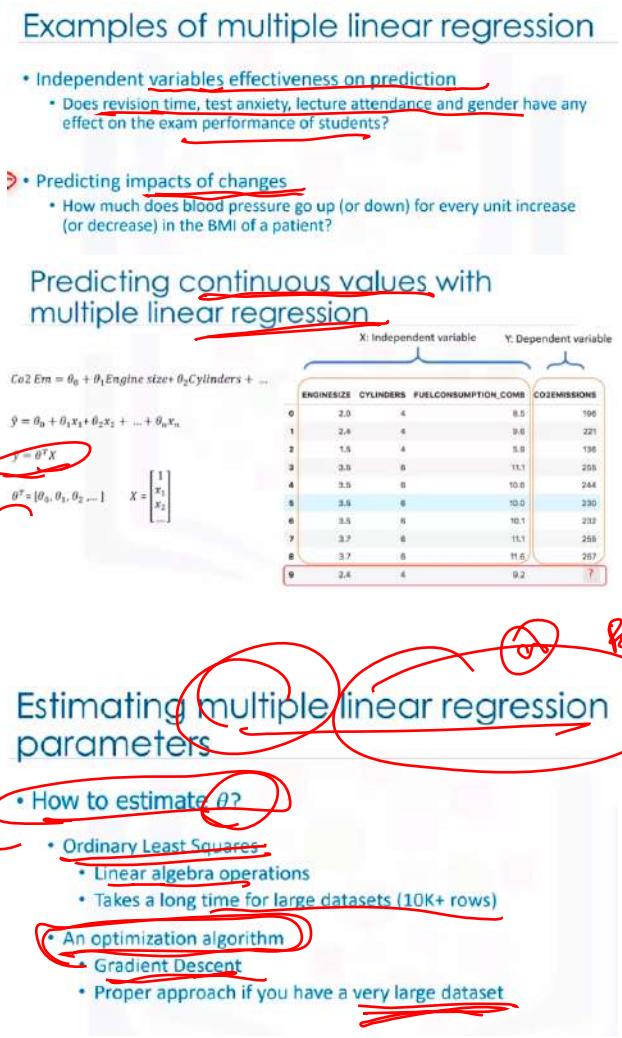
- **Lab : Simple Linear Regression**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%202/ML0101EN-Reg-Simple-Linear-Regression-Co2.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Reg-Simple-Linear-Regression-Co2%202.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Reg-Simple-Linear-Regression-Co2%202.ipynb



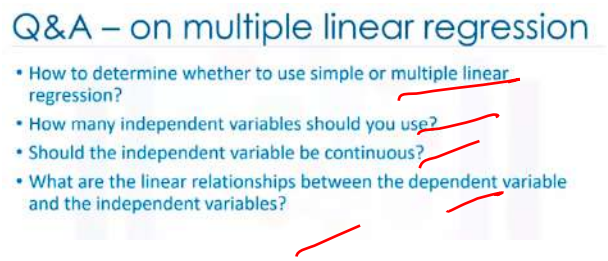
- **Multiple linear regression**



What is the best approach to find the parameter or coefficients for multiple linear regression, when we have very large dataset?

- [ ]
- [ ]

Using an optimization approach



- **Lab : Multiple Linear Regression**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%202/ML0101EN-Reg-Mulitple-Linear-Regression-Co2.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Reg-Mulitple-Linear-Regression-Co2%202.ipynb

- **Non Linear Regression**

NOT IMPORTANT , SO SKIPPED THIS SECTION VIDEO NOTES MAKING

- Video
- **Lab Polynomial Regression**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%202/ML0101EN-Reg-Polynomial-Regression-Co2.ipynb
- **Lab Non Linear Regression**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%202/ML0101EN-Reg-NoneLinearRegression.ipynb
- *Let's learn about non linear regressions and apply an example on python. In this notebook, we fit a* ***non-linear model to the data points corresponding to China's GDP from 1960 to 2014.***
- **Quiz : Regression**
- **Multiple Linear Regression is appropriate for:**
- Predicting tomorrow's rainfall amount based on the wind speed and temperature
- **Which of the following is the meaning of "Out of Sample Accuracy" in the context of evaluation of models?**
- "Out of Sample Accuracy" is the percentage of correct predictions that the model makes on data that the model has NOT been trained on
- **When should we use Multiple Linear Regression?**
- When there are multiple dependent variables
- When we would like to predict impacts of changes in independent variables on a dependent variable.
- **Which of the following statements are TRUE about Polynomial Regression?**
- Polynomial regression models can fit using the Least Squares method.
- Polynomial regression fits a curve line to your data.
- Polynomial regression can use the same mechanism as Multiple Linear Regression to find the parameters.
- **Which sentence isTRUE about Non-linear Regression?**
- Nonlinear regression is a method to model non linear relationship between the dependent variable and a set of independent variables.

- For a model to be considered non-linear, y must be a non-linear function of the parameters.
- **Non-linear regression must have more than one dependent variable (FALSE)**

- **WEEK 3 : CLASSIFICATION**

In this module, you will learn about classification technique. You practice with different classification algorithms, such as KNN, Decision Trees, Logistic Regression and SVM. Also, you learn about pros and cons of each method, and different classification accuracy metrics

- **K Nearest Neighbours**
- **Introduction to Classification**

What is classification?

- A supervised learning approach
- Categorizing some unknown items into a discrete set of categories or "classes"
- The target attribute is a categorical variable

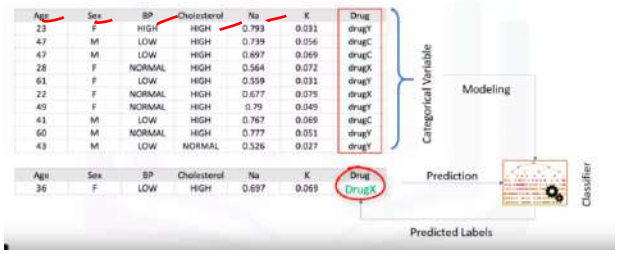
How does classification work?

Classification determines the class label for an unlabeled test case.

multi-class classifier?

A classifier that can predict a field with multiple discrete values, such as "DrugA", "DrugX" or "DrugY".

Example of multi-class classification



Classification use cases

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

- Which category a customer belongs to?
- Whether a customer switches to another provider/brand?
- Whether a customer responds to a particular advertising campaign?

Classification applications

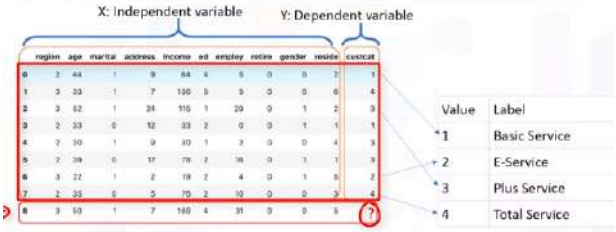
\

Classification algorithms in machine learning

- Decision Trees (ID3, C4.5, C5.0)
- Naïve Bayes
- Linear Discriminant Analysis
- *k*-Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM)

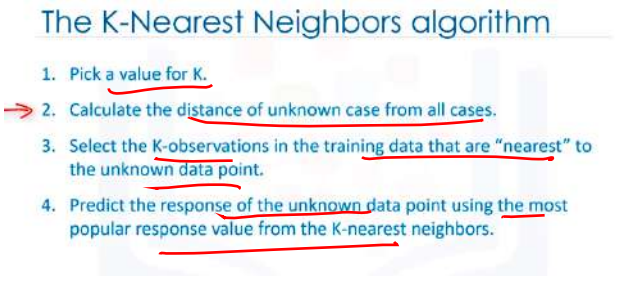
- **K Nearest Neighbours**

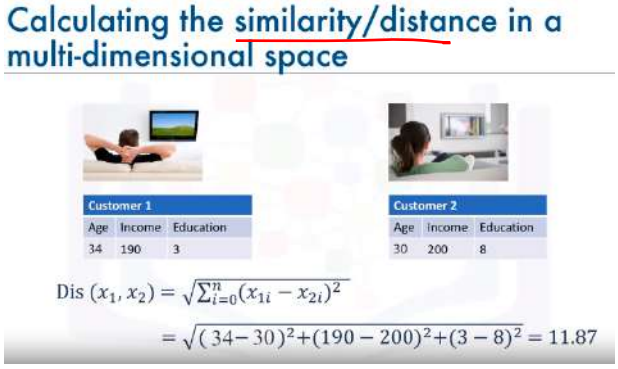
Intro to KNN



What is K-Nearest Neighbor (or KNN)?

- A method for classifying cases based on their similarity to other cases
- Cases that are near each other are said to be "neighbors"
- Based on similar cases with same class labels are near each other

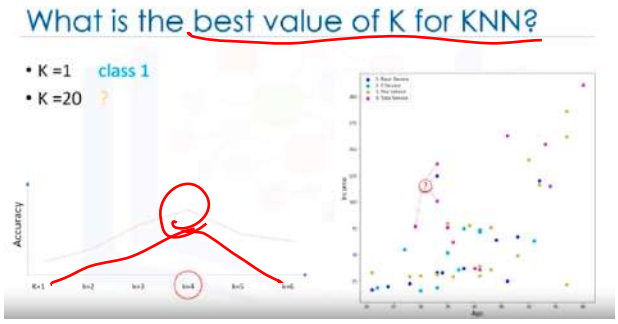


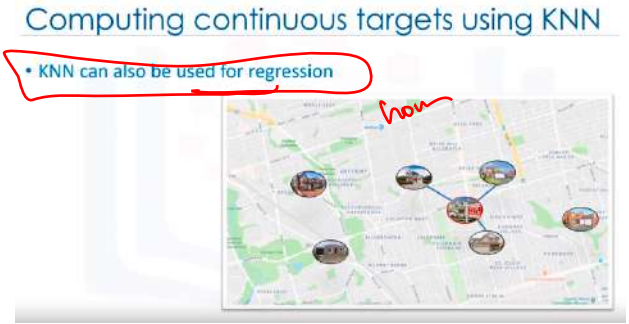


The kNN algorithm is a classification algorithm.

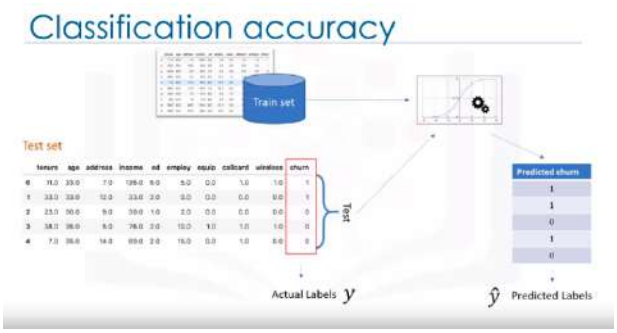
Correct

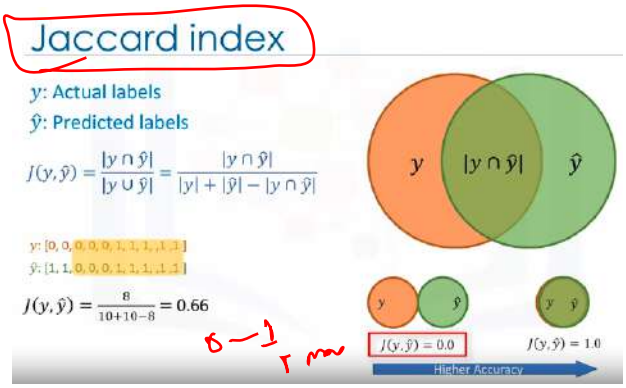
The kNN algorithm classify cases based on their similarity to other cases.

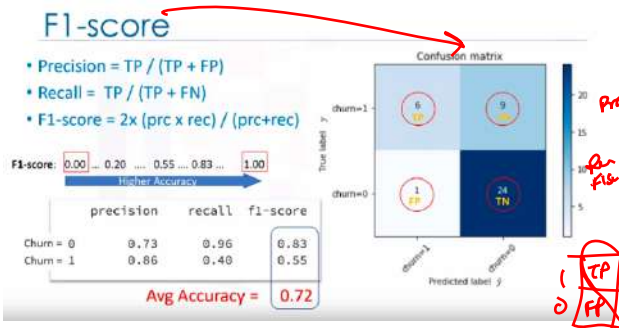




- **Evaluation metrics in classification**

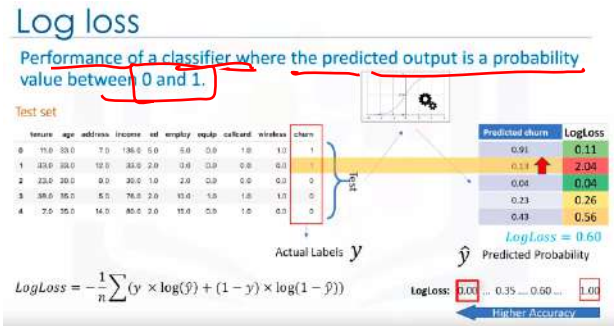






Which one is the ideal classifier?

The classifier with F1-score close to one.



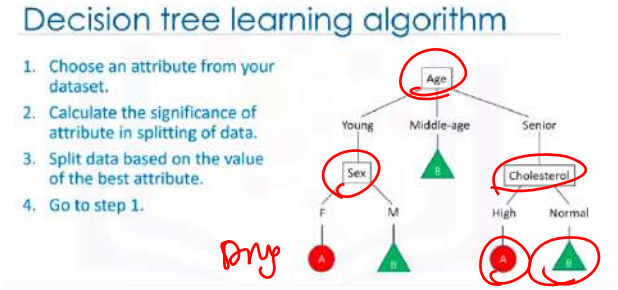
- **Lab : KNN**
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/

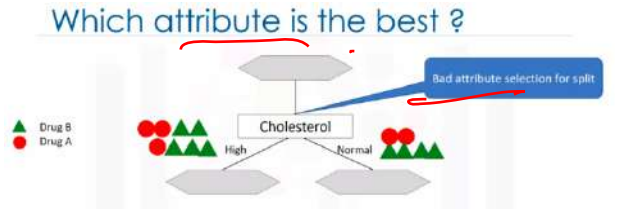
Module%203/ML0101EN-Clas-K-Nearest-neighbors-CustCat.ipynb
https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Clas-K-Nearest-neighbors-CustCat%202.ipynb

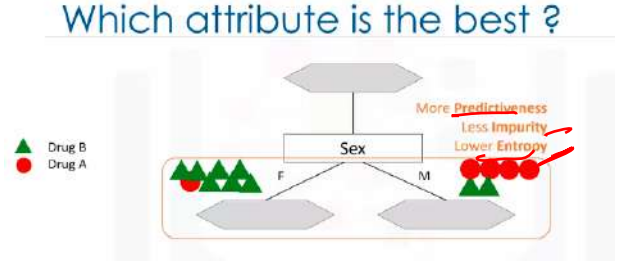
- **Decision Trees**
- **Introduction to Decision Trees**
-

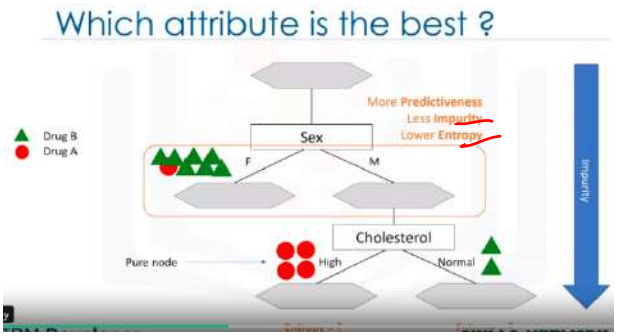
Decision Trees are built by splitting the training set into distinct nodes

One node in a Decision Tree contains all of or most of, one category of the data

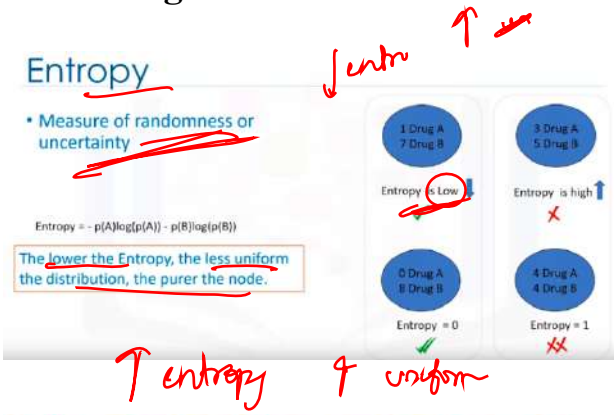


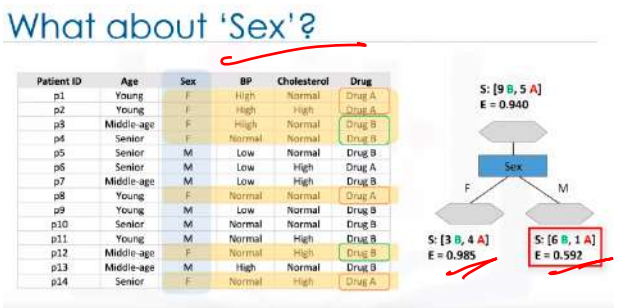


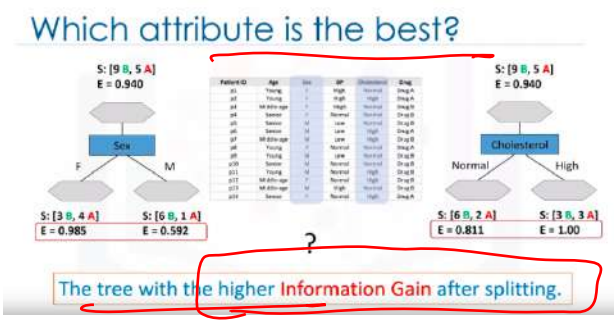




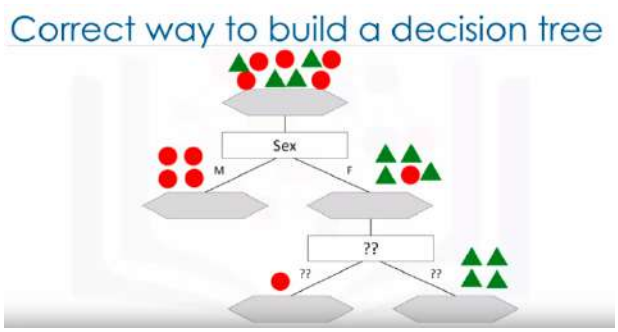
- **Building Decision Trees**







The entropy in a node is the amount of information disorder calculated in each node.



- **Lab : Decision Trees**
- In this lab exercise, you will learn a popular machine learning algorithm, Decision Tree. You will use this classification algorithm to **build a model from historical data of patients, and their respond to different medications. Then you use the trained decision tree to predict the class of a unknown patient, or to find a proper drug for a new patient.** Click HERE to download the lab notebook (.ipynb)

ML0101EN-Clas-Decision-Trees-drug 3.ipynb

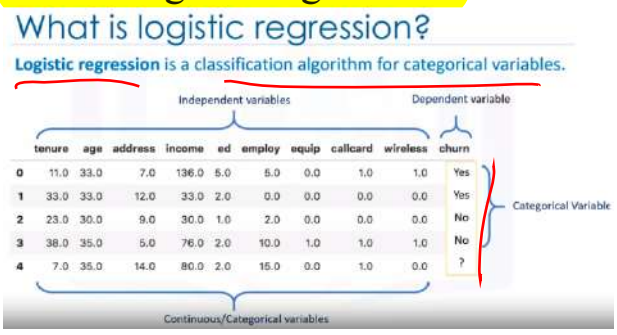
- https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Clas-Decision-Trees-drug%203.ipynb
-

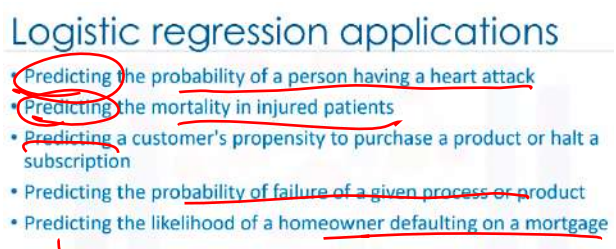
- **READ RANDOM FOREST AFTER DECISION TREES WHICH NOT COVERED IN IBM**

Refer 96-108 from Sir Machine Learning notes
9 - Machine learning.pdf

- **Logistic Regression**
- Intro to Logistic Regression

In this video, we'll learn a machine learning method called **Logistic Regression which is used for classification**. In examining this method, we'll specifically answer these three questions. What is logistic regression? What kind of problems can be solved by logistic regression? In which situations do we use logistic regression?





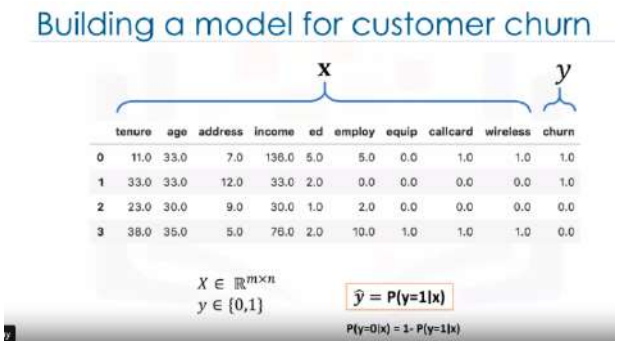
Logistic regression is analogous to linear regression but takes a categorical/discrete target field instead of a numeric one.

Correct

Logistic Regression measures the probability of a case belonging to a specific class.

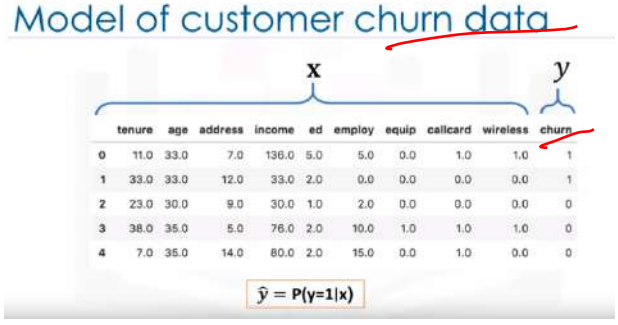
Correct

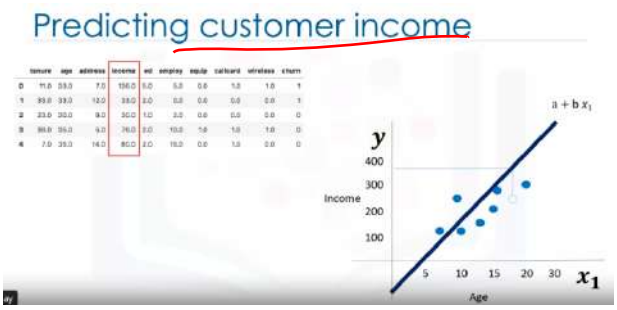
Logistic Regression can be used to understand the impact of a feature on a dependent variable.

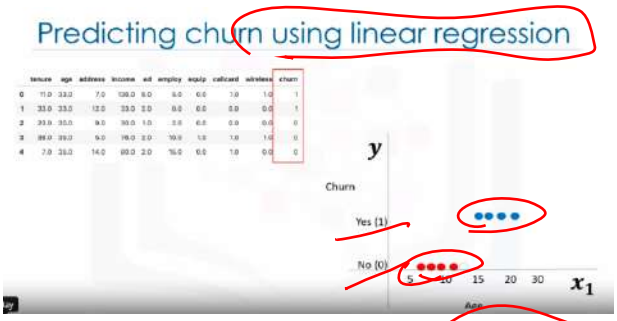


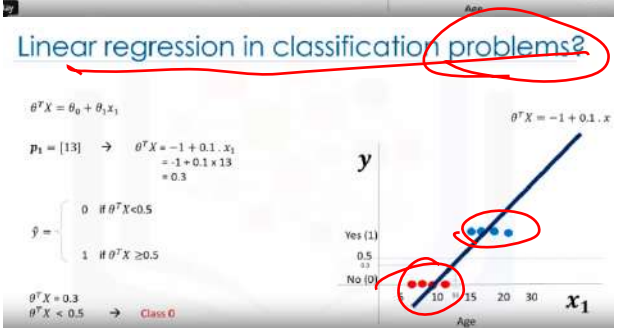
- **Logistic Regression Vs Linear Regression**

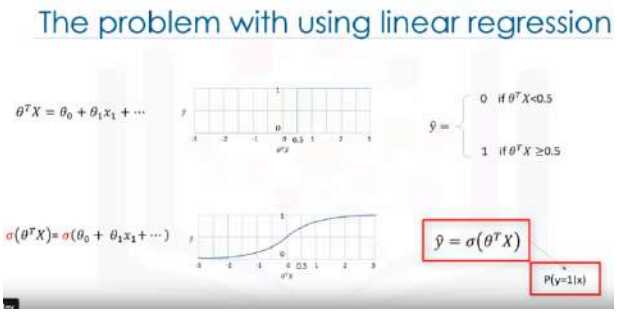
we will learn the difference between linear regression and logistic regression. We go over linear regression and see why it cannot be used properly for some binary classification problems. We also look at the sigmoid function, which is the main part of logistic regression.

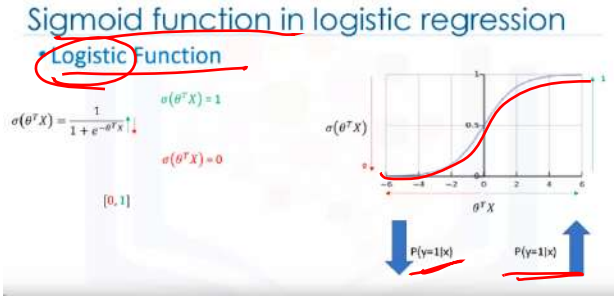


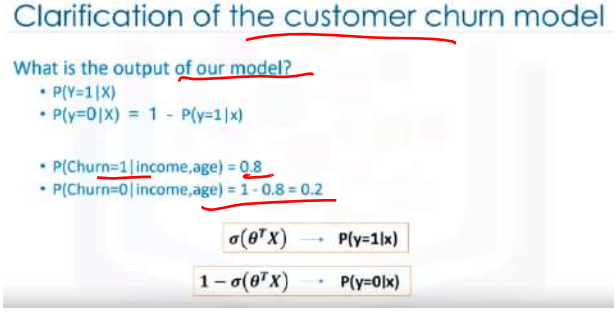


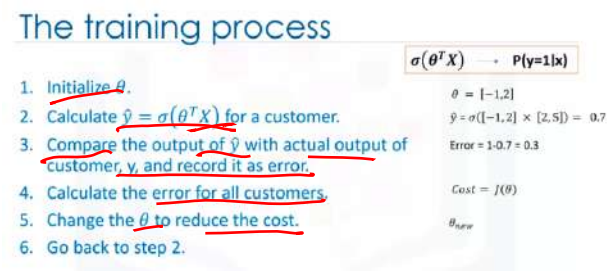


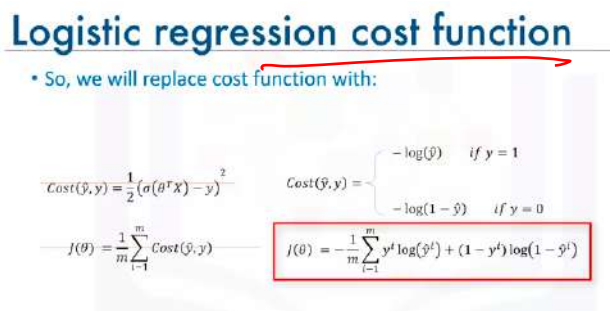


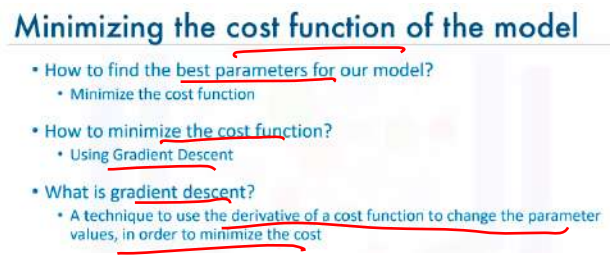












difference between Linear Regression vs Logistic Regression, in solving a classification problem?

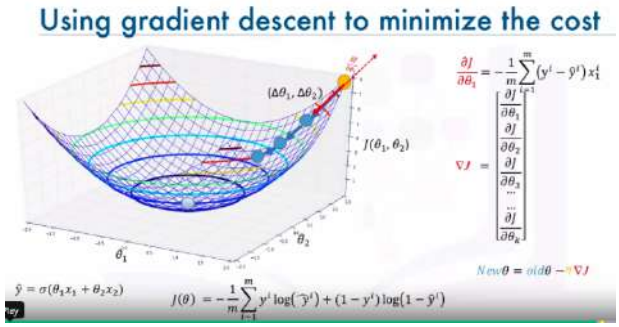
Linear Regression cannot properly measure the probability of a case belonging to a class

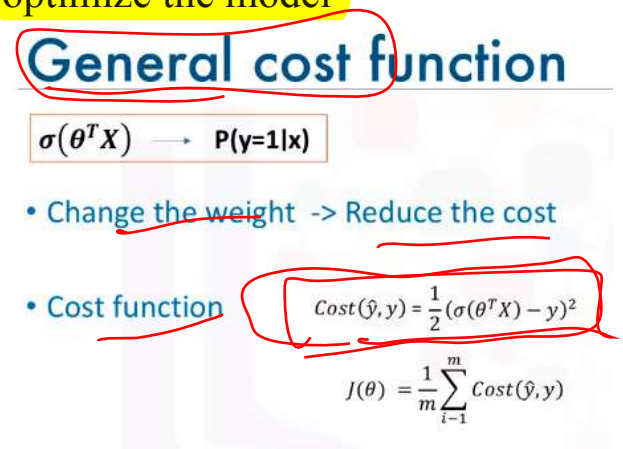
- **Logistic Regression Training**

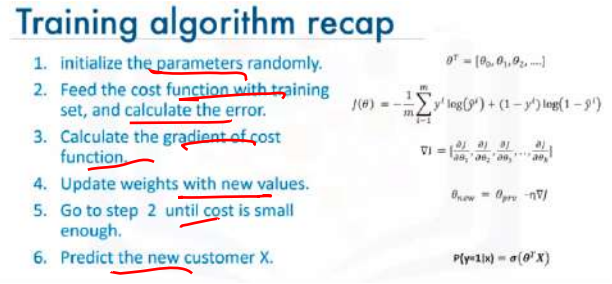
we will learn more about training a logistic regression model. Also, we will be discussing how to change the parameters of the model to better estimate the outcome. Finally, we talk about the cost function and gradient descent in logistic regression as a way to optimize the model

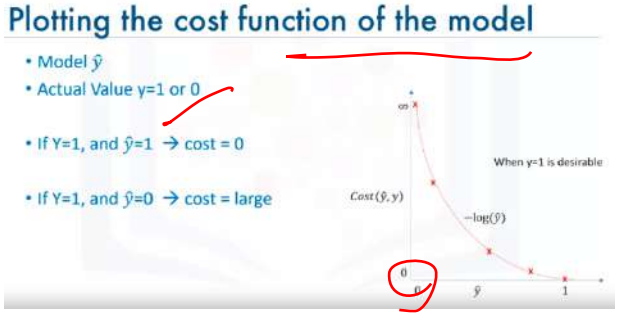
"**gradient descent**" in training process?

A technique to use derivative of a cost function to change the parameter values, to minimize the cost.









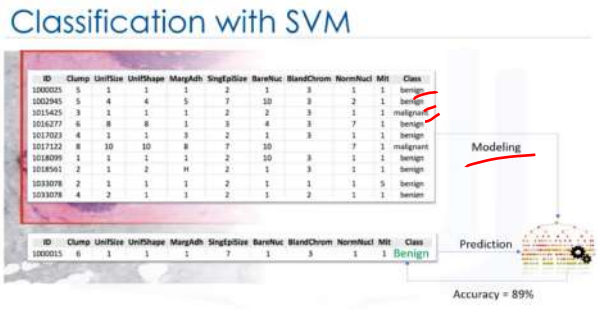
- **Lab : Logistic Regression**
- In this notebook, you will learn Logistic Regression, and then, you'll create a model with telecommunications data to predict when its customers will leave for a competitor, so that you can take

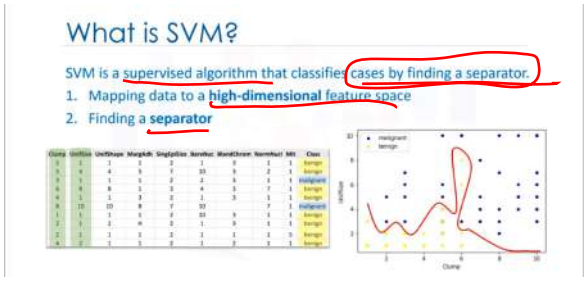
some action to retain the customer. Click HERE to download the lab notebook (.ipynb)

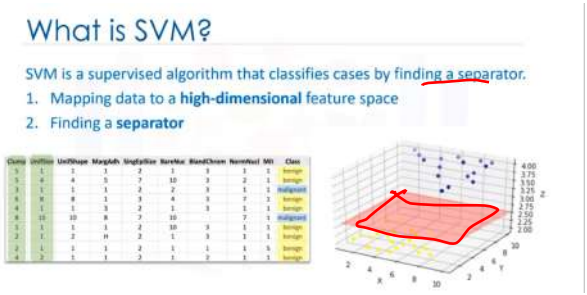
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%203/ML0101EN-Clas-Logistic-Reg-churn.ipynb

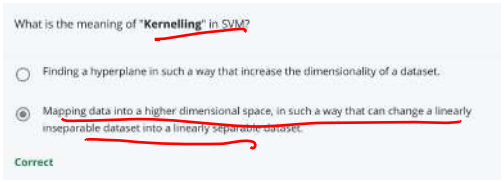
ML0101EN_Clas_Logistic_Reg_churn.ipynb https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN_Clas_Logistic_Reg_churn.ipynb

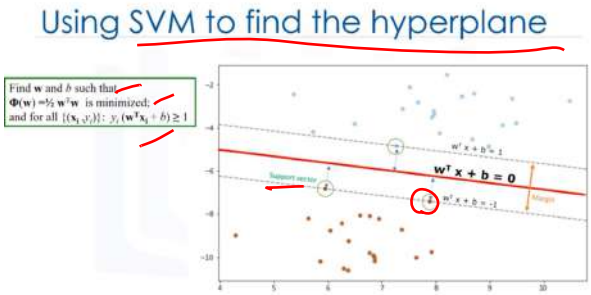
- **Support Vector Machine**
- Video











Hello and welcome. In this video, we will learn a machine learning method called, Support Vector Machine, or SVM, which is used for classification.

In this notebook, you will use SVM (Support Vector Machines) to build and train a model using human cell records, and classify cells to whether the samples are benign or malignant. Click HERE to download the lab notebook (.ipynb)

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%203/ML0101EN-Clas-SVM-cancer.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Clas-SVM-cancer%202%202.ipynb

Week 3 : Quiz

Classification

Question 1: Which one IS NOT a sample of classification problem?

- To predict the category to which a customer belongs to.
- To predict whether a customer switches to another provider/ brand.
- **To predict the amount of money a customer will spend in one year.**
- To predict whether a customer responds to a particular advertising campaign or not.

Question 2: Which of the following statements are TRUE about Logistic Regression? (select all that apply)

- **Logistic regression can be used both for binary classification and multi-class classification**
- **Logistic regression is analogous to linear regression but takes a categorical/ discrete target field instead of a numeric one.**
- **In logistic regression, the dependent variable is binary.**

Which of the following examples is/are a sample application of Logistic Regression? (select all that apply)

- **The probability that a person has a heart attack within a specified time period using person's age and sex.**
- **Customer's propensity to purchase a product or halt a subscription in marketing applications.**
- **Likelihood of a homeowner defaulting on a mortgage.**

- Estimating the blood pressure of a patient based on her symptoms and biographical data.

Which one is TRUE about the kNN algorithm?

- kNN is a classification algorithm that takes a bunch of unlabelled points and uses them to learn how to label other points.
- **kNN algorithm can be used to estimate values for a continuous target.**

What is "information gain" in decision trees?

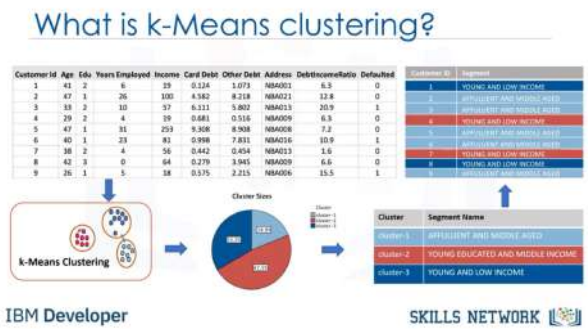
- It is the information that can decrease the level of certainty after splitting in each node.
- **It is the entropy of a tree before split minus weighted entropy after split by an attribute.**
- It is the amount of information disorder, or the amount of randomness in each node.

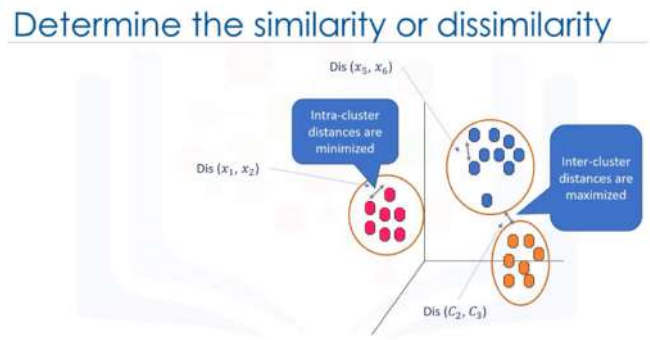
End of week 3

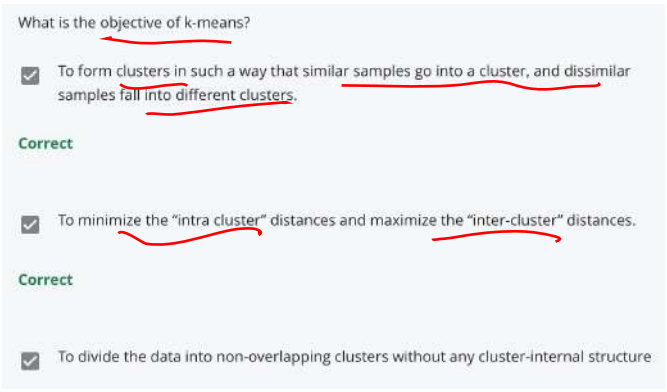
- **Week 4 : Clustering**
- **k-Means Clustering**
-
- **Intro to Clustering**

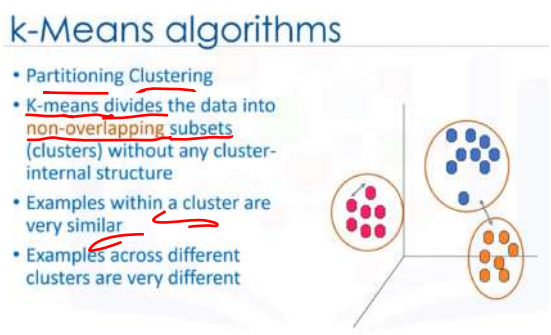
In this video we'll give you a high level introduction to clustering, its applications, and different types of clustering algorithms **this has been covered after k means (Read after)**

- **Intro to k means**









1-dimensional similarity/distance

Customer 1
Age
54

Customer 2
Age
50

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

$$\text{Dis}(x_1, x_2) = \sqrt{(54-50)^2} = 4$$

Multi-dimentional similarity/distance

Customer 1		
Age	Income	education
54	190	3

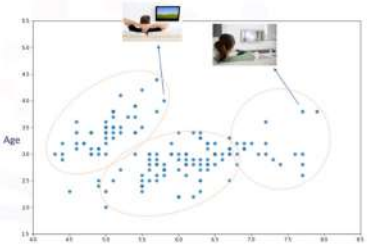
Customer 2		
Age	Income	education
50	200	8

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

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

How does k-Means clustering work?

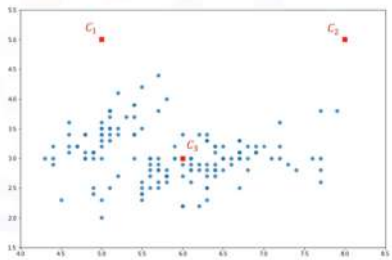
Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...	...	...



k-Means clustering – initialize k

1) Initialize k=3 centroids randomly

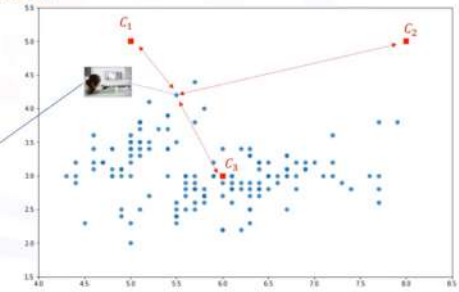
C_1 = [8., 5.]
C_2 = [5., 5.]
C_3 = [6., 3.]



K-Means clustering – calculate the distance

2) Distance calculation

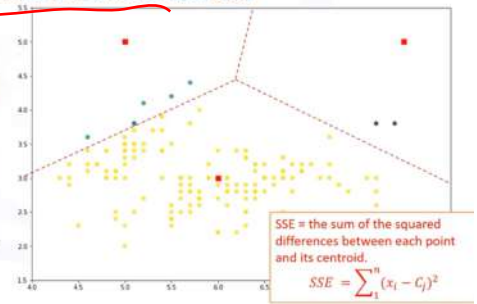
C_1	C_2	C_3
$d(p1,c1)$	$d(p1,c2)$	$d(p1,c3)$
$d(p2,c1)$	$d(p2,c2)$	$d(p2,c3)$
$d(p3,c1)$	$d(p3,c2)$	$d(p3,c3)$
$d(p4,c1)$	$d(p4,c2)$	$d(p4,c3)$
$d(p\ldots,c1)$	$d(p\ldots,c2)$	$d(p\ldots,c3)$
$d(pm,c1)$	$d(pm,c2)$	$d(pm,c3)$



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

C_1	C_2	C_3
$d(p1,c1)$	$d(p1,c2)$	$d(p1,c3)$
$d(p2,c1)$	$d(p2,c2)$	$d(p2,c3)$
$d(p3,c1)$	$d(p3,c2)$	$d(p3,c3)$
$d(p4,c1)$	$d(p4,c2)$	$d(p4,c3)$
$d(p\ldots,c1)$	$d(p\ldots,c2)$	$d(p\ldots,c3)$
$d(pm,c1)$	$d(pm,c2)$	$d(pm,c3)$



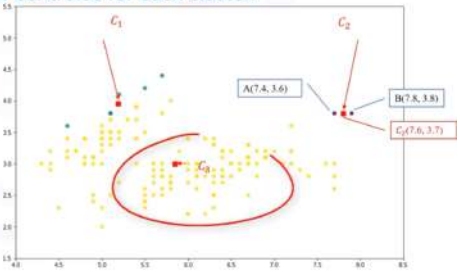
SSE = the sum of the squared differences between each point and its centroid.

$$SSE = \sum_{1}^{n}(x_i - C_j)^2$$

k-Means clustering – compute new centroids

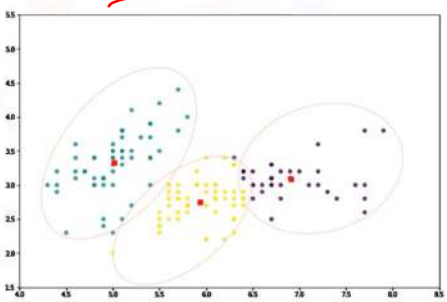
4) Compute the new centroids for each cluster.

C_1	C_2	C_3
$d(p1,c1)$	$d(p1,c2)$	$d(p1,c3)$
$d(p2,c1)$	$d(p2,c2)$	$d(p2,c3)$
$d(p3,c1)$	$d(p3,c2)$	$d(p3,c3)$
$d(p4,c1)$	$d(p4,c2)$	$d(p4,c3)$
$d(p\ldots,c1)$	$d(p\ldots,c2)$	$d(p\ldots,c3)$
$d(pm,c1)$	$d(pm,c2)$	$d(pm,c3)$



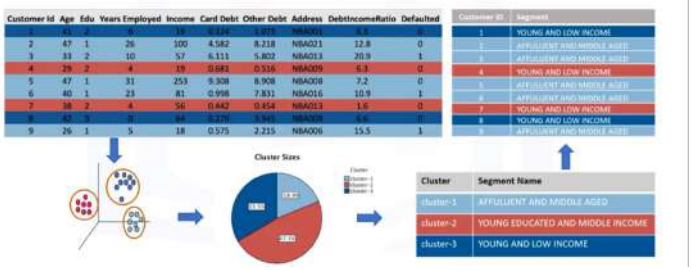
k-Means clustering – repeat

5) Repeat until there are no more changes.

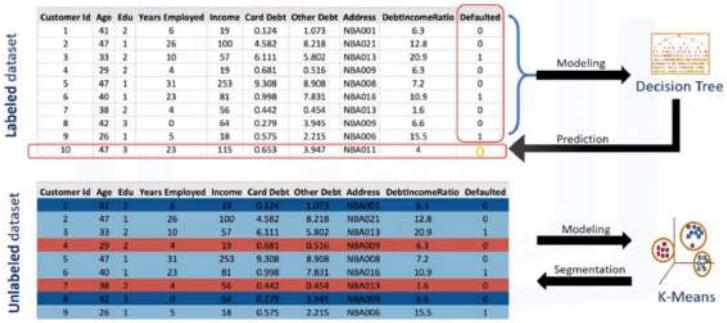


Intro to clustering (before k means

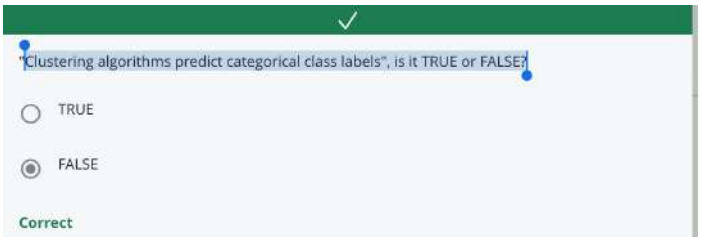
Clustering for segmentation



Clustering Vs. classification



)



Clustering applications

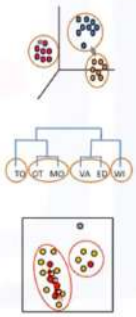
- PUBLICATION:
  - Auto-categorizing news based on their content
  - Recommending similar news articles
- MEDICINE:
  - Characterizing patient behavior
- BIOLOGY:
  - Clustering genetic markers to identify family ties

Why clustering?

- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

Clustering algorithms

- Partitioned-based Clustering
  - Relatively efficient
  - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
  - Produces trees of clusters
  - E.g. Agglomerative, Divisive
- Density-based Clustering
  - Produces arbitrary shaped clusters
  - E.g. DBSCAN



Clustering applications

- RETAIL/MARKETING:
  - Identifying buying patterns of customers
  - Recommending new books or movies to new customers
- BANKING:
  - Fraud detection in credit card use
  - Identifying clusters of customers (e.g., loyal)
- INSURANCE:
  - Fraud detection in claims analysis
  - Insurance risk of customers

- **More on K means**

In this video, we'll look at k-Means accuracy and characteristics.

k-Means clustering algorithm

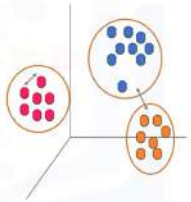
1. Randomly placing *k* centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the *k* centroids.

k-Means clustering algorithm

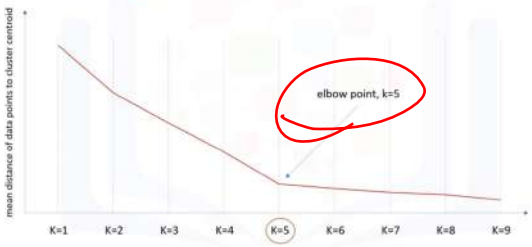
1. Randomly placing *k* centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the *k* centroids.
5. Repeat the steps 2-4, until the centroids no longer move.

k-Means accuracy

- External approach
  - Compare the clusters with the ground truth, if it is available.
- Internal approach
  - Average the distance between data points within a cluster.



Choosing k



In clustering evaluation process, "**elbow point**" is where the rate of accuracy increase sharply, when we run clustering multiple times, increasing k in each run.

- TRUE
- FALSE

k-Means recap

- Med and Large sized databases (*Relatively efficient*)
- Produces sphere-like clusters
- Needs number of clusters (k)

- **Lab : k means**

Despite its simplicity, the K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data. In this notebook, you learn how to use k-Means for customer segmentation. Click HERE to download the lab notebook (.ipynb)

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%204/ML0101EN-Clus-K-Means-Customer-Seg.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-Clus-K-Means-Customer-Seg%202.ipynb

Hierarchical clustering

Not important primarily , read Notemaking later

Intro to Hierarchical clustering

More on hierarchical clustering

Hello and welcome. In this video, we'll be covering more details about hierarchical clustering. Let's get started. Let's look at agglomerative algorithm for hierarchical clustering.

lets compare hierarchical clustering with K-means. K-means is more efficient for large data sets. In contrast to K-means, hierarchical clustering does not require the number of cluster to be specified. Hierarchical clustering gives more than one partitioning depending on the resolution or as K-means gives only one partitioning of the data. Hierarchical clustering always generates the same clusters, in contrast with K-means, that returns different clusters each time it is run, due to random initialization of centroids.

Lab : Agglomerative Clustering

In this lab, we will be looking at Agglomerative clustering, which is more popular than Divisive clustering. We will also be using Complete Linkage as the Linkage Criteria. Click HERE to download the lab notebook (.ipynb)

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%204/ML0101EN-Clus-Hierarchical-Cars.ipynb

- **Density Based Clustering**
-
- **Not important primarily , read Notemaking later**
-
- DBSCAN
- 7 min

Hello and welcome. In this video, we'll be covering DB scan. A density-based clustering algorithm which is appropriate to use when examining spatial data.

- **Lab : DBSCAN Clustering**

Density-based Clustering locates regions of high density that are separated from one another by regions of low density. Density, in this context, is defined as the number of points within a specified radius.

In this section, the main focus will be manipulating the data and properties of DBSCAN and observing the resulting clustering. Click HERE to download the lab notebook (.ipynb)

- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%204/ML0101EN-Clus-DBSCN-weather.ipynb
- Quiz : Clustering

Machine Learning with Python Coursera Quiz Answers Week 4

Question 1: Which statement is NOT TRUE about k-means clustering?

- k-means divides the data into non-overlapping clusters without any cluster-internal structure.
- The objective of k-means, is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.
- **As k-means is an iterative algorithm, it guarantees that it will always converge to the global optimum.**

Question 2: Which of the following are characteristics of DBSCAN? Select all that apply.

- **DBSCAN can find arbitrarily shaped clusters.**
- **DBSCAN can find a cluster completely surrounded by a different cluster.**
- **DBSCAN has a notion of noise, and is robust to outliers.**
- **DBSCAN does not require one to specify the number of clusters such as k in k-means**

Question 3: Which of the following is an application of clustering?

- Customer churn prediction
- Price estimation
- **Customer segmentation**
- Sales prediction

Question 4: Which approach can be used to calculate dissimilarity of objects in clustering?

- **Minkowski distance**
- **Euclidian distance**
- **Cosine similarity**
- **All of the above**

Question 5: How is a center point (centroid) picked for each cluster in k-means?

- **We can randomly choose some observations out of the data set and use these observations as the initial means.**
- **We can create some random points as centroids of the clusters.**
- We can select it through correlation analysis.
- **End of week 4**

- **Week 5**
-
- **Content Based Recommendation Engines**
- **Intro to Recommender systems**

What are recommender systems?

Recommender systems capture the pattern of peoples' behavior and use it **to predict what else they might want or like.**

Applications

- What to buy?
  - E-commerce, books, movies, beer, shoes
- Where to eat?
- Which job to apply to?
- Who you should be friends with?
  - LinkedIn, Facebook, ...
- Personalize your experience on the web
  - News platforms, news personalization

NETFLIX amazon

Advantages of recommender systems

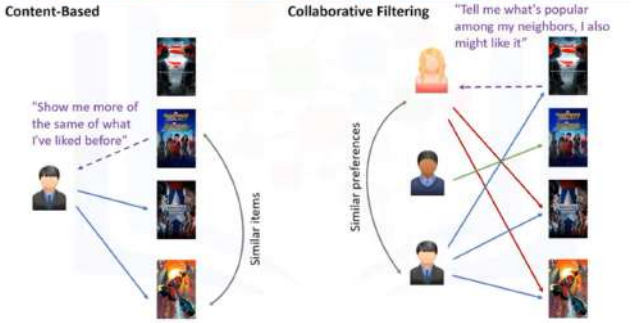
- Broader exposure
- Possibility of continual usage or purchase of products
- Provides better experience

What is a "**Content-based**" recommender system?

- ○ Content-based technique attempts to figure out what's popular among the neighbours, and recommend it.
- ◉ Content-based technique tries to figure out what a user's favourite aspects of an item is, and then recommends items that present those aspects.

Correct

Two types of recommender systems



-
- **Content based recommender systems**

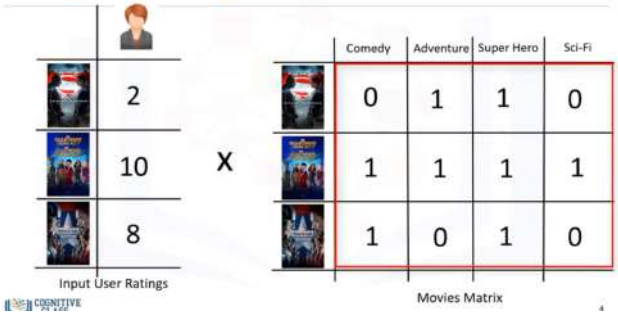
Content-based recommender systems

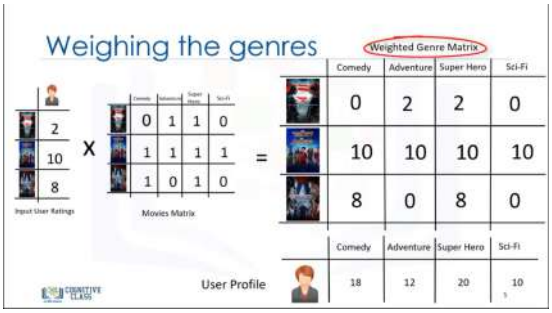
Content-based recommender systems

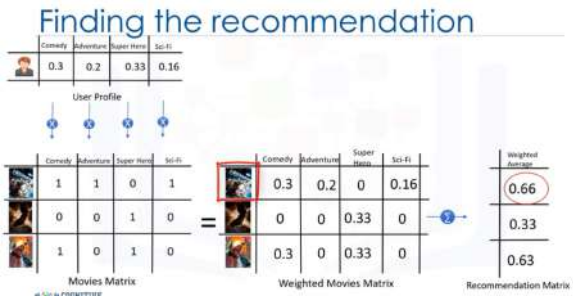
Which one is **TRUE** about Content-based recommendation systems?

- ☑ Content-based recommendation system tries to recommend items to the users based on their profile.
- ☑ In content-based approach, the recommendation process is based on similarity of users.
- ☑ In content-based recommender systems, similarity of users should be measured based on the similarity of the actions of users.

Weighing the genres







- **Lab : Content based Recommendation systems**

Recommendation systems are a collection of algorithms used to recommend items to users based on information taken from the user. These systems have become ubiquitous can be commonly seen in online stores, movies databases and job finders. In this notebook, we will explore Content-based recommendation systems and implement a simple version of one using Python and the Pandas library. Click HERE to download the lab notebook (.ipynb)

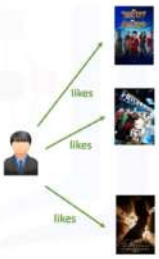
https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%205/ML0101EN-RecSys-Content-Based-movies.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-RecSys-Content-Based-movies%202.ipynb

- **Collaborative filtering**
- Video
- In this video, we'll be covering a recommender system technique called collaborative filtering

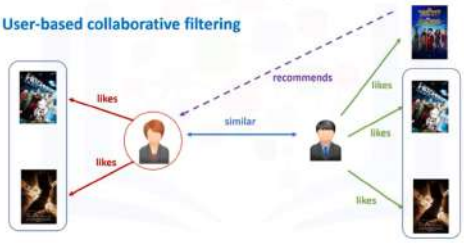
Collaborative filtering

- **User-based collaborative filtering**
  - Based on users' neighborhood
- **Item-based collaborative filtering**
  - Based on items' similarity

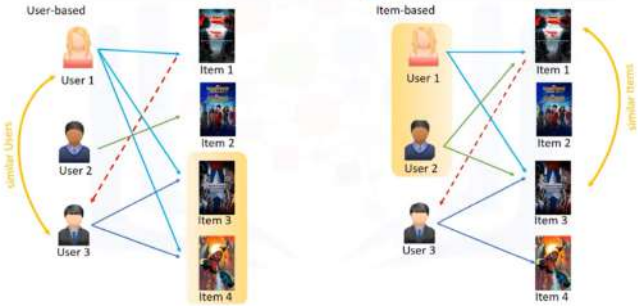


Collaborative filtering

- **User-based collaborative filtering**



Collaborative filtering



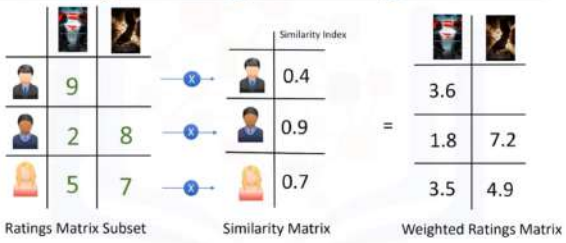
Which one is correct about user-based and item-based collaborative filtering?

- In **user-based** approach, the recommendation is based on users of the same neighborhood, with whom he/she shares common preferences.

Correct

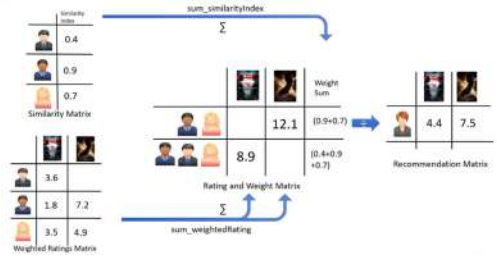
Learning the similarity weights

Creating the weighted ratings matrix



Challenges of collaborative filtering

- **Data Sparsity**
  - Users in general rate only a limited number of items
- **Cold start**
  - Difficulty in recommendation to new users or new items
- **Scalability**
  - Increase in number of users or items



Lab: Collaborative Filtering on Movies

Recommendation systems are a collection of algorithms used to recommend items to users based on information taken from the user. These systems have become ubiquitous can be commonly seen in online stores, movies databases and job finders. In this notebook, we will explore recommendation systems based on Collaborative Filtering and implement simple version of one using Python and the Pandas library. Click HERE to download the lab notebook (.ipynb)

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%205/ML0101EN-RecSys-Collaborative-Filtering-movies.ipynb

https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/ML0101EN-RecSys-Collaborative-Filtering-movies%202.ipynb

Quiz : Recommender System

Question 1: What is/are the advantage/s of Recommender Systems ?

- Recommender Systems provide a better experience for the users by giving them a broader exposure to many different products they might be interested in.
- Recommender Systems encourage users towards continual usage or purchase of their product
- Recommender Systems benefit the service provider by increasing potential revenue and better security for its consumers.

What is a content-based recommendation system?

- **Content-based recommendation system tries to recommend items to the users based on their profile built upon their preferences and taste.**

What is the meaning of "Cold start" in collaborative filtering?

The difficulty in recommendation when we have new user, and we cannot make a profile for him, or when we have a new item, which has not got any rating yet.

What is a "Memory-based" recommender system?

In memory based approach, we use the entire user-item dataset to generate a recommendation system.

What is the shortcoming of content-based recommender systems?

- **Users will only get recommendations related to their preferences in their profile, and recommender engine may never recommend any item with other characteristics.**

End of Week 5

Week 6

Final Project

Reading : How to do final project? PDF

Reading: Instructions for Final Peer Graded Assignment PDF

This final project will be graded by your peers who are completing this course during the same session. This project is worth 25 marks of your total grade, broken down as follows:

1. Building model using KNN, finding the best k and accuracy evaluation (7 marks)
2. Building model using Decision Tree and find the accuracy evaluation (6 marks)
3. Building model using SVM and find the accuracy evaluation (6 marks)
4. Building model using Logistic Regression and find the accuracy evaluation (6 marks)

Peer-graded Assignment: The best classifier

Rishab Rawat(Best Classifier).ipynb

Final Exam

What is the **subfield of computer science that gives "computers the ability to learn without being explicitly programmed."?**

Machine Learning

Regression/Estimation, Classification, Clustering, and Associations are all examples of what

ML Techniques

Which type of regression model can by transformed into a linear regression model using the Least Squares method?

Polynomial regression a

Which one IS a sample of classification problem?

To predict whether a customer responds to a particular advertising campaign or not.

To predict the category to which a customer belongs to.

To predict whether a customer switches to another provider/brand.

~~To predict the amount of money a customer will spend in one year.~~

Which of the following statements are TRUE about Logistic Regression? (select all that apply)

1 / 1 point

Logistic regression can be used both for binary classification and multi-class classification

Logistic regression is analogous to linear regression but takes a categorical/ discrete target field instead of a numeric one.

In logistic regression, the dependent variable is binary.

TRUE about k-means clustering?

The objective of k-means, is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.

k-means divides the data into non-overlapping clusters without any cluster-internal structure.

~~As k-means is an iterative algorithm, it guarantees that it will always converge to the global optimum.~~

characteristics of DBSCAN

DBSCAN can find arbitrarily shaped clusters.

DBSCAN can find a cluster completely surrounded by a different cluster.

DBSCAN has a notion of noise, and is robust to outliers.

DBSCAN does not require one to specify the number of clusters such as k in k-means

A Recommender system provides a better experience for the user by giving them a broader exposure to many different products they might be interested in.

Question 17) What is a content-based recommendation system?

- **Content-based recommendation system tries**

to recommend items to the users based on their profile built upon their preferences and taste.

When we should use Multiple Linear Regression?

- **When we would like to identify the strength of the effect that the independent variables have on a dependent variable.**

Ideapoke Soundarya Questions

- **Types of ML algorithm**
- **Supervised**
  - Regression
  - Classification
    - Naive Bayes Classifier
  - Forecasting
- **Semi supervised**
- **Unsupervised**
  - Clustering
  - Dimension reduction
- **Reinforcement**
- We do Trail & error , learns from past experiences
- **Normalisation & standardisation and it's range**

What is Normalization?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

What is Standardization?

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ

- **Project you done in Supervised ML Week 1**

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

load a dataset, clean it, and find out interesting insights from it.

Data collection
Using API

& web scraping

Week 2 :

Exploratory data analysis

Using pandas .
,Seaborn , Matplotlib

Visual analytics & dashboard
Data Visualization with folium
Dashboard with Plotly dash

Predictive analysis (Classfication)
https://github.com/SWAROOPNC/Machine-Learning-With-Python/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Logistic Regression

Standard scalar
Train test split

Logistic regression
Grid search
SVM
Decision tree classifier
KNN

https://github.com/rodrigoalvamat/ibm-data-science-capstone/blob/master/notebooks/machine-learning-prediction.ipynb

Logistic regression was better in predicting scale while KNN just Labelled

What is a neural network?

A **neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain**.

It is a **type of machine learning process, called deep learning**, that **uses interconnected nodes or neurons in a layered structure that resembles the human brain.**

How a Neural Network Works?

A neural network has many layers. Each layer performs a specific function, and the complex the network is, the more the layers are. That's why a neural network is also called a multi-layer perceptron.

The purest form of a neural network has three layers:

1. The input layer
2. The hidden layer
3. The output layer

As the names suggest, each of these layers has a specific purpose. These layers are made up of nodes. There can be multiple hidden layers in a neural network according to the requirements. The input layer picks up the input signals and transfers them to the next layer. It gathers the data from the outside world.

The hidden layer performs all the back-end tasks of calculation. A network can even have zero hidden layers. However, a neural network has at least one hidden layer. The output layer transmits the final result of the hidden layer's calculation.

Like other machine learning applications, you will have to train

a neural network with some training data as well, before you provide it with a particular problem. But before we go more in-depth of how a neural network solves a problem, you should know about the working of perceptron layers first:

Artificial Neural Network (ANN):

Artificial Neural Network (ANN), is a group of multiple perceptrons or neurons at each layer. ANN is also known as a Feed-Forward Neural network because inputs are processed only in the forward direction.

Convolutional Neural Network (CNN):

Convolutional neural networks (CNN) are one of the most popular models used today. This neural network computational model **uses a variation of multilayer perceptrons and contains one or more convolutional layers that can be either entirely connected or pooled**

Recurrent Neural Network (RNN):

Recurrent neural networks **(RNN) are more complex.** They **save the output of processing nodes and feed the result back into the model (they did not pass the information in one direction only)**. This is how the model is said to learn to predict the outcome of a layer. Each node in the RNN model acts as a memory cell, continuing the computation and implementation of operations. **If the network's prediction is incorrect, then the system self-learns and continues working towards the correct prediction during backpropagation.**

Refer ANN CNN RNN pdf in PI GitHub

Bi gram Trigram N gram

Language modeling is the way of determining the probability of any sequence of words.

N-gram

N-gram can be defined as the contiguous sequence of n items from a given sample of text or speech. The items can be letters, words, or base pairs according to the application. The N-grams typically are collected from a text or speech corpus (A long text dataset).

An N-gram language model predicts the probability of a given N-gram within any sequence of words in the language. A good N-gram model can predict the next word in the sentence i.e the value of p(w|h)

NLP

Natural language processing (NLP) is a subfield of Artificial Intelligence (AI).

This is a widely used technology for personal assistants that are used in various business fields/areas. This technology works on the speech provided by the user, breaks it down for proper understanding and processes accordingly.

The field is divided into three different parts:

1. Speech Recognition — The translation of spoken language into text.
2. Natural Language Understanding (NLU) — The computer's ability to understand what we say.
3. Natural Language Generation (NLG) — The generation of natural language by a computer.

Technologies related to NLP
Machine Translation
Chatterbots
AI software

Applications
Chatbits
Spam filters
Answering questions

TF-IDF or (Term Frequency(TF) — Inverse Dense Frequency(IDF))is a technique which is used to find meaning of sentences consisting of words and cancels out the incapabilities of Bag of Words technique which is

good for text classification or for helping a machine read words in numbers

quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document

TF-IDF is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. We can then remove the words that are less important for analysis

In **CountVectorizer** we **only count the number of times a word appears in the document** which results in biasing in favour of most frequent words. this ends up in ignoring rare words which could have helped is in processing our data more efficiently.

NLP steps

1. Step #1: Sentence Segmentation. ...
2. Step #2: Word Tokenization. ...
3. Step #3: Predicting Parts of Speech for each token. ...
4. Step #4: Lemmatization. ...
5. Step #5: Identifying stop words. ...
6. Step 6.1: Dependency Parsing.

Narendra Modi is PM of India , How do you process name of PM in NLP in 2 lines

✗

activation function is **a function that is added into an artificial neural network in order to help the network learn complex patterns in the data**. When comparing with a neuron-based model that is in our brains, the activation function is at the end deciding what is to be fired to the next neuron.

You can use relu function as activation in the final layer.

Range of relu , 0 to infinity

Sigmoid function
range (0,1)
NLp usage ,
Rating prediction
using reviews