1. **What do you mean by Cluster Analysis?**

A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects.

**2. What are the fields in which clustering techniques are used?**

Clustering is used in biology to develop new plants and animal taxonomies. Clustering is used in business to enable marketers to develop new distinct groups of their customers and characterize the customer group on basis of purchasing.

Eg: Clustering is used in the identification of groups of automobiles Insurance policy customer. Clustering is used in the identification of groups of houses in a city on the basis of house type, their cost and geographical location. Clustering is used to classify the document on the web for information discovery.

**3.What are the requirements of cluster analysis?**

The basic requirements of cluster analysis are

Dealing with different types of attributes.

Dealing with noisy data.

Constraints on clustering.

Dealing with arbitrary shapes.

High dimensionality

Ordering of input data Interpretability and usability

Determining input parameter and Scalability

**4.What are the different types of data used for cluster analysis?**

The different types of data used for cluster analysis are interval scaled, binary, nominal, ordinal and ratio scaled data.

**5. What are interval scaled variables?**

Interval scaled variables are continuous measurements of linear scale. For example, height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance

**6. Define Binary variables? And what are the two types of binary variables?**

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and

when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

**7. Define nominal.**

A nominal variable is a generalization of the binary variable. Nominal variable has more than

two states, For example, a nominal variable, color consists of four states, red, green, yellow, or black.

In Nominal variables the total number of states is N and it is denoted by letters, symbols or integers.

**8 ordinal**

An ordinal variable also has more than two states but all these states are ordered in a meaningful sequence.

**9 ratio scaled variables?**

A ratio scaled variable makes positive measurements on a non-linear scale, such as exponential scale, using the formula

AeBt or Ae-Bt

Where A and B are constants.

**10. What do u mean by partitioning method?**

In partitioning method, a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

11. Write formula for Precision.

* Precision = tp/(tp + fp)

where tp and fp are the numbers of true positive from the prediction **p** and false positive predictions for the considered class when the actual value in data set

12. Write formula for **F-measures.**

* **F-measures** = 2\*precision\*recall/(precision + recall)

where precision can be seen as a measure of exactness, whereas recall is a measure of completeness or quantity.

13. Define Confusion matrix.

* **confusion matrix**, the diagonal elements are the predicted data of class label as **class**.

14. What do you mean by Purity?

Purity is **a measure of the extent to which clusters contain a single class**. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster.

15. How entropy can measure in cluster?

The weight for each cluster is the probability of the cluster, which is just the number of tuples in the cluster divided by the total number of tuples.

16 marks

1. Explain briefly on partitioning method in cluster analysis. 8m

* clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. Its the data analysts to specify the number of clusters that has to be generated for the clustering methods.
* In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represent ts a cluster and a particular region.
* There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Mediods), CLARA algorithm (Clustering Large Applications) etc.

1. Illustrate K- Means Algorithm in clustering. 8m

* The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the inter-cluster similarity is low.
* Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster’s centroid or center of gravity.
* The k-means algorithm proceeds as follows
* First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.
* For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
* It then computes the new mean for each cluster.
* This process iterates until the criterion function converges.

1. Determine an Advantages, Issues and application of K- Means Cluster. 8m

Advantages

* Relatively simple to implement.
* Scales to large data sets.
* Guarantees convergence.
* Can warm-start the positions of centroids.
* Easily adapts to new examples.
* Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Issues

* It requires to specify the number of clusters (k) in advance.
* It can not handle noisy data and outliers.
* It is not suitable to identify clusters with non-convex shapes.

Application of K Means

Some common applications for clustering include the following:

* market segmentation
* social network analysis
* search result grouping
* medical imaging
* image segmentation
* anomaly detection

1. Illustrate on K- Medoids Clustering algorithm with suitable example. 8m

A medoid can be defined as the point in the cluster, whose dissimilarities with all

the other points in the cluster is minimum.

The dissimilarity of the medoid (Ci) and object (Pi) is calculated by using E = |Pi - Ci|.

1. Initialize*:* select k random points out of the n data points as the medoids.  
   2. Associate each data point to the closest medoid by using any common distance

metric methods.  
3. While the cost decreases:  
        For each medoid m, for each data o point which is not a medoid:  
        Swap m and o, associate each data point to the closest medoid, recompute thecost.  
      If the total cost is more than that in the previous step, undo the swap.

**Advantages:**

1. It is simple to understand and easy to implement.
2. K-Medoid Algorithm is fast and converges in a fixed number of steps.
3. PAM is less sensitive to outliers than other partitioning algorithms.
4. Explain about silhouette Method. 8 m

The**silhouette Method**is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters.

* The silhouette value is a measure of how similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Computing Silhoutte Coefficient:

Steps to find the silhouette coefficient of an i’th point:

1. Compute a(i): The average distance of that point with all other points in the same

clusters.

1. Compute b(i): The average distance of that point with all the points in the closest

cluster to its cluster.

1. Compute s(i) — silhouette coefficient or i’th point using below mentioned formula.



       After computing the silhouette coefficient of each point in the dataset, plot it to get a visual representation of how well the dataset is clustered into k clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like the number of clusters visually. This measure has a range of [-1, 1].

1. Solve the sample data of given series and form 2 cluster using K- Means technique.

S= {2, 3, 4, 10, 11, 12, 20, 25, 30} 16 m

Let consider K=2,

Iteration 1:

Choose two points randomly,

M1= 4 M2 =12

K1={2, 3, 4} K2= {10, 11, 12, 20, 25, 30}

Iteration 2:

Then find the mid value of these two cluster k1 and k2 as

M1= 3 and M2= 18

K1={ 2, 3 , 4, 10} K2= { 11, 12, 20, 25, 30}

Then find the mid value of these two cluster k1 and k2 as

M1= 5 and M2= 20

Iteration 3:

K1={ 2, 3 , 4, 10, 11, 12} K2= { 20, 25, 30}

Then find the mid value of these two cluster k1 and k2 as

M1= 7 and M2= 25

Iteration4:

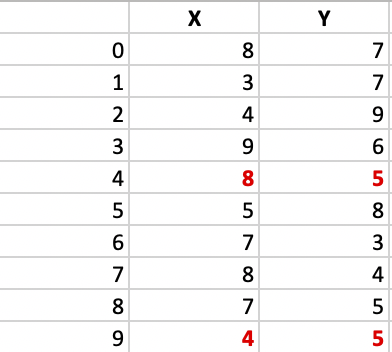
K1={ 2, 3 , 4, 10, 11, 12} K2= { 20, 25, 30}

Then find the mid value of these two cluster k1 and k2 as

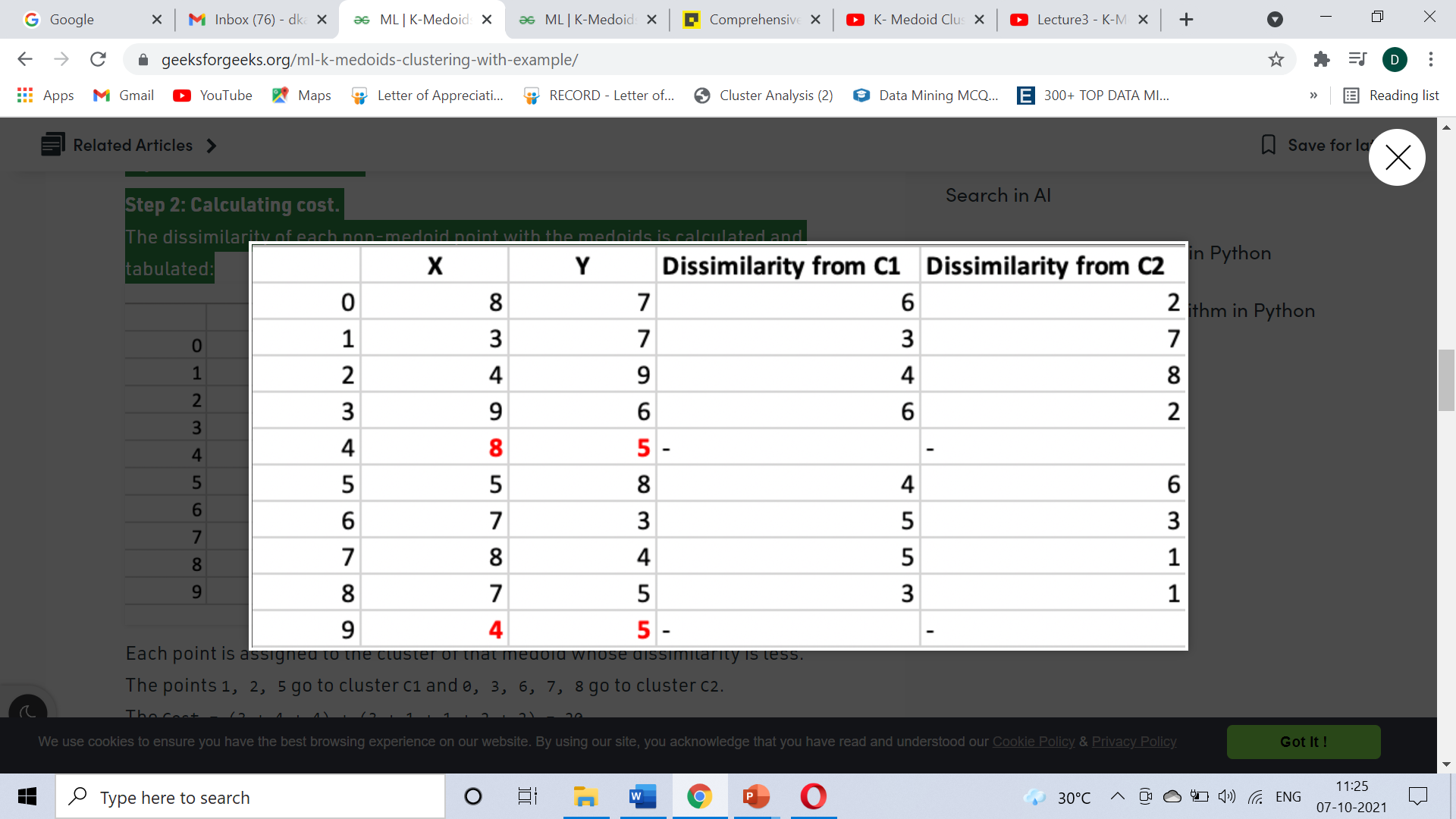
M1= 7 and M2= 25

Further we get mean value as similar as previous value. Now to stop the process and keep the value in cluster K1 an K2 respectively.

1. To cluster the following dataset using K- Medoids technique. Initially can take two values K1= 8 and 5 K2= 4 and 5.



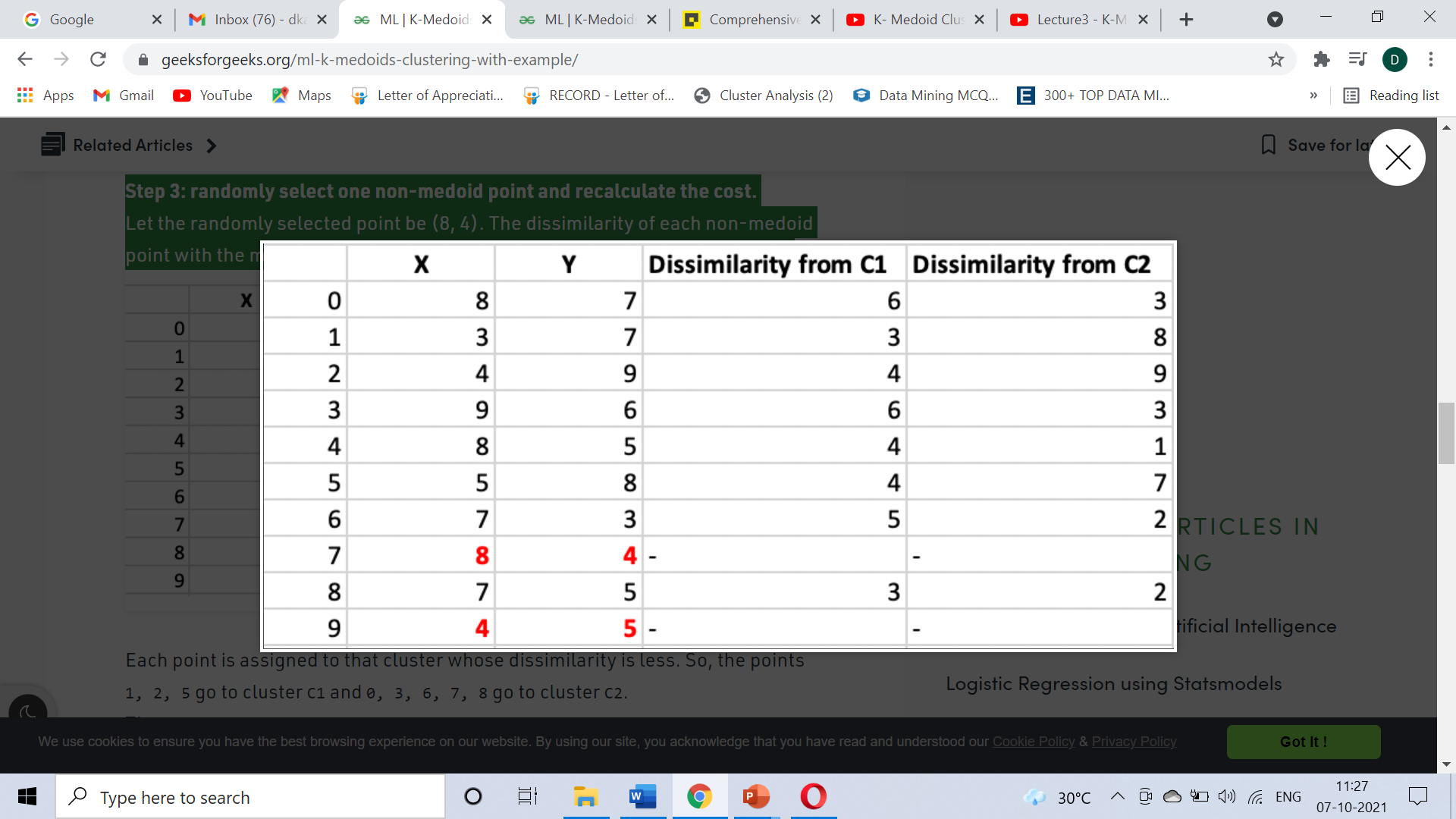
* **Step 1:**  
  Let the randomly selected 2 medoids, so select k = 2 and let **C1 -(4, 5)** and **C2 -(8, 5)** are the two medoids.
* **Step 2: Calculating cost.**  
  The dissimilarity of each non-medoid point with the medoids is calculated and tabulated:



Each point is assigned to the cluster of that medoid whose dissimilarity is less.  
The points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.  
The Cost = (3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20

**Step 3: randomly select one non-medoid point and recalculate the cost.**  
Let the randomly selected point be (8, 4). The dissimilarity of each non-medoid point

with the medoids – C1 (4, 5) and C2 (8, 4) is calculated and tabulated.



Each point is assigned to that cluster whose dissimilarity is less.

So, the points 1, 2, 5, 9 go to cluster C1 and 0, 3, 4, 6, 7, 8 go to cluster C2.  
The New cost = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22  
Swap Cost = New Cost – Previous Cost = 22 – 20 and **2 >0**

As the swap cost is not less than zero, we undo the swap.

Hence **(3, 4) and (8, 4)**are the final medoids.

Unit 4

1. Where are decision trees mainly used?

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

1. Define Data Classification.

Data classification is the process of analyzing structured or unstructured data and organizing it into categories based on file type, contents, and other metadata.

1. How is prediction different from classification?

Classification is a supervised learning approach, which predicts the ‘class label’ of a new data using model constructed from the existing data, for example, decision tree.

The task of predicting continuous values for a given input is termed as prediction. Regression – based methods are used for prediction, For example: multi – linear regression analysis.

1. How to evaluate the Classification and prediction methods?

Accuracy • Speed • Robustness • Scalability • Interpretability

1. List out some classification methods.

Genetic Algorithm, Fuzzy Classifier, Rough set, K- Nearest Neighbor Classifier.

1. What is the process involving in data preparation?

Data tuples – Irrelevant attributes- Require transformation

1. List out the two main strategies that can be used for testing and training.

Percentage split and K- fold cross validation.

1. How do you calculate Euclidean Distance Metrics?

This formula says the distance between two points (x1 , y1 ) and (x2 , y2  ) is **d = √[(x2 – x1)2 + (y2 – y1)2]**.

1. How is **Manhattan Distance** calculated?

The **Manhattan Distance** between two points **(X1, Y1)** and **(X2, Y2)** is given by **|X1 – X2| + |Y1 – Y2|**.

1. What is the difference between “supervised” and unsupervised” learning?

Supervised learning algorithms are trained using labeled data. Unsupervised learning algorithms **are trained using unlabeled 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.

1. What does the Bayesian classification provide?

Bayesian classification is based on Bayes Theorem. It provides the basis for probabilistic learning that accommodates prior knowledge and takes into account the observed data.

1. How will you solve a classification problem using decision trees?

Decision tree builds **classification or regression models in the form of a tree structure**. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Decision trees can handle both categorical and numerical data.

1. How does KNN determine k value?

The optimal K value usually found is **the square root of N**, where N is the total number of samples.

1. What is meant by KNN classifier?

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

1. Why is the KNN Algorithm known as Lazy Learner?

When the KNN algorithm gets the training data, it does not learn and make a model, it just stores the data. Instead of finding any discriminative function with the help of the training data, it follows **instance-based learning** and also uses the training data when it actually needs to do some prediction on the unseen datasets.

16 Marks

1. Decision Tree Induction 8 marks

Decision tree is a non-parametric supervised machine learning algorithm which is used mostly for classification and regression problems. In this scenario, we split the population or sample into two or more homogeneous sets (or subpopulations) based on the most significant splitter/differentiator in input variables.

A flow-chart-like tree structure

¬ Internal node denote on an attribute node (non leaf node) denote attribute ¬ Branch representation outcome of the test

¬ Leaf nodes represent class labels or class distribution (Terminal node)

The topmost node in a tree is the root node. Decision tree generation consists of two phases

¬ Tree construction

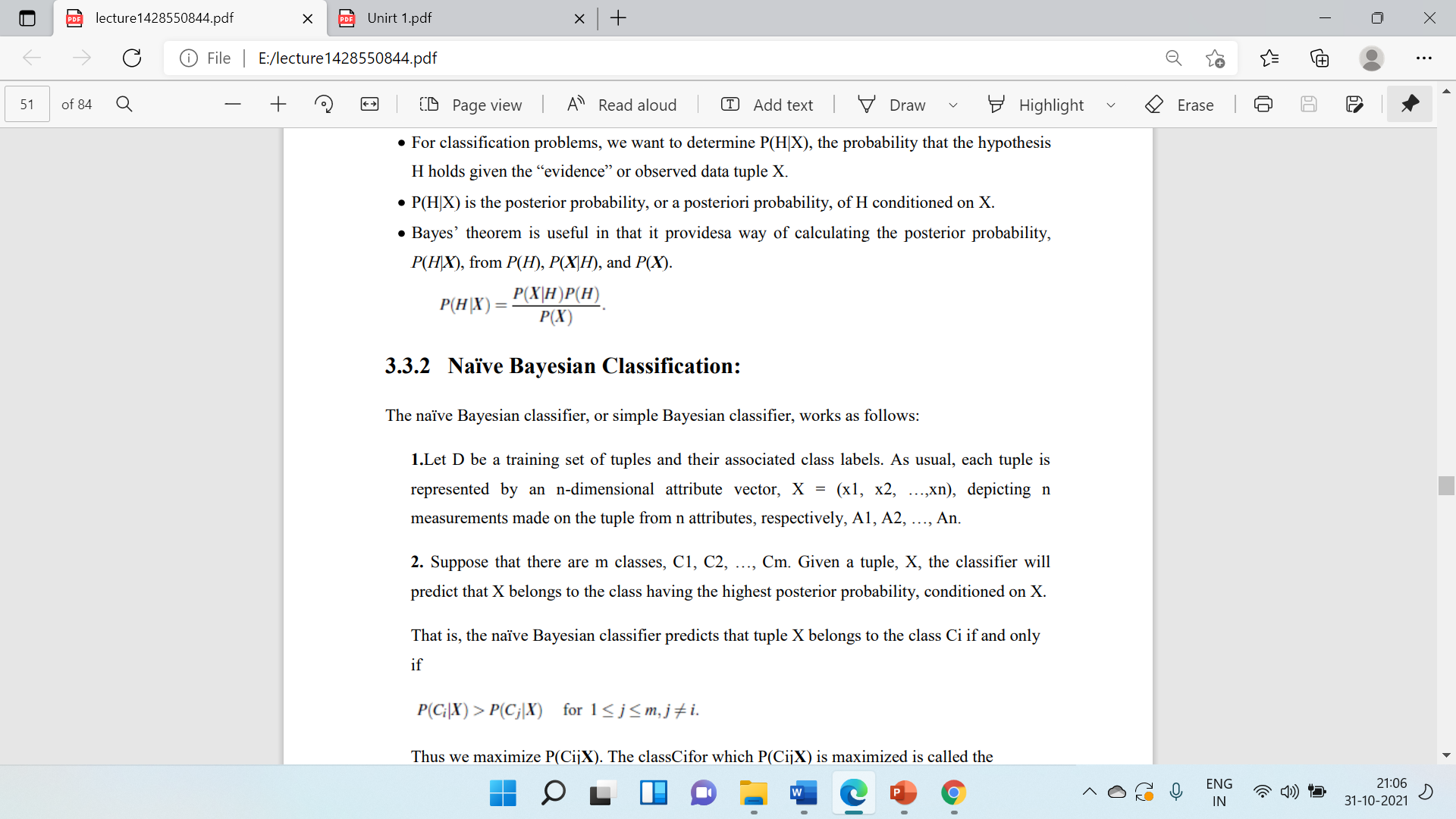
♣ At start, all the training examples are at the root ♣ Partition examples recursively based on selected attributes

¬ Tree pruning ♣ Identify and remove branches that reflect noise or outliers

1. Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs toa particular class. Bayesian classification is based on Bayes’ theorem.

Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the ―evidence‖ or observed data tuple X. P(H|X) is the posterior probability, or a posteriori probability, of H conditioned on X. Bayes’ theorem is useful in that it providesa way of calculating the posterior probability, P(H|X), from P(H), P(X|H), and P(X).



1. Illustrate on steps follow in K- Nearest Neighbour using R with suitable example.

16 m

* Step 1: Import the dataset
* Step 2: Data Normalization
* Step 3: Data Splicing
* Step 4: Building a Machine Learning model
* Step 5: Model Evaluation

Step 6: View in Plot or Cross Table

**Import the data**

Knn <-read.csv("E:\\RProgramming\\dataset1.csv")

print(Knn)

**#Normalization**

normal<- function(x)

{

return((x-min(x))/(max(x)-min(x)))}

#apply Normalization

knn.n <- as.data.frame(lapply(Knn[,1:7],normal))

print(knn.n)

nrow(knn.n)

**#Data Spacing - Partioning**

set.seed(123)

partion<- sample(1:nrow(knn.n),size=nrow(knn.n)\*0.8,replace=FALSE) #RANDOM SELECTION OF 80% DATA

train.knn<- Knn[partion,]

test.knn<- Knn[-partion,] #remaining 20% test data

print(train.knn)

print(test.knn)

train.label<-Knn[partion,1]

test.label<-Knn[-partion,1]

print(train.label)

print(test.label)

install.package('class')

library(class)

**# To finding the no. of Observation**

nrow(train.knn)

knn.6<-knn(train=train.knn,test=test.knn,cl=train.label,k=6)

knn.4<-knn(train=train.knn,test=test.knn,cl=train.label,k=4)

print(knn.6)

print(knn.4)

1. How does the KNN algorithm make the predictions on the unseen dataset? Explain. 8

The following operations have happened during each iteration of the algorithm. For each of the unseen or test data point, the kNN classifier must:

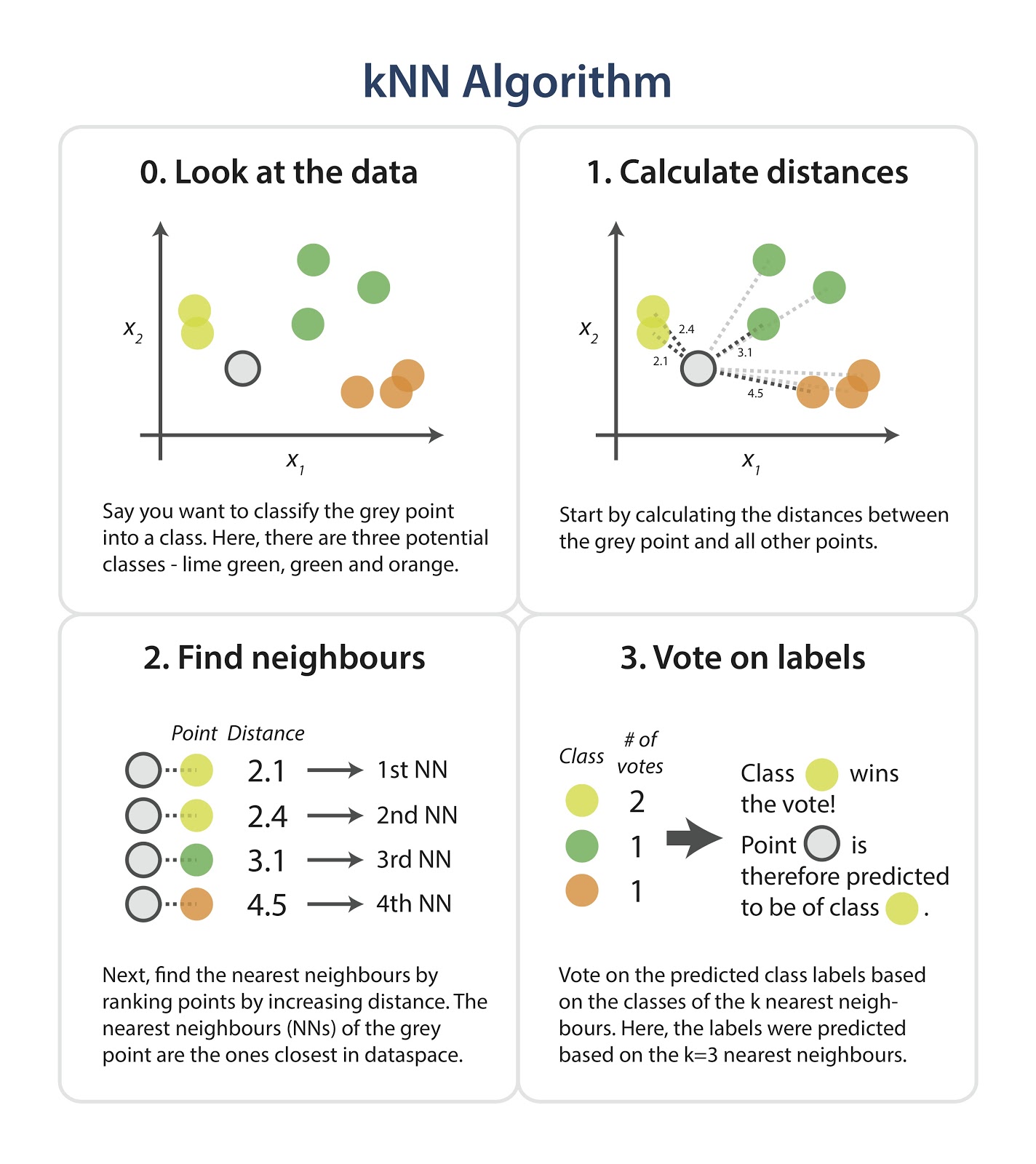
**Step-1:** Calculate the distances of test point to all points in the training set and store them

**Step-2:** Sort the calculated distances in increasing order

**Step-3:** Store the K nearest points from our training dataset

**Step-4:**Calculate the proportions of each class

**Step-5:**Assign the class with the highest proportion



1. Explain briefly on different phases in data classification. 8

* Training phase
* Testing phase

Process of classifier Data Creation

* Percentage Split
* K- Fold Cross Validation

1. Illustrate on data preparation and application of classification. 8

Data Tuples- Irrelevant attributes- Require Transformation

**Application**

Marketers- Meteorologist- Health care system- Financial

1. Illustrate on Nearest neighbour classifier. 8

* Intuitively, it is based on the idea that if x, y are close to each other, then probably they’ll have the same class.
* If x, y∈ R^d are two length-d vectors, define ||x − y|| to be the distance between x and y. For instance, we might use the Euclidean distance:

