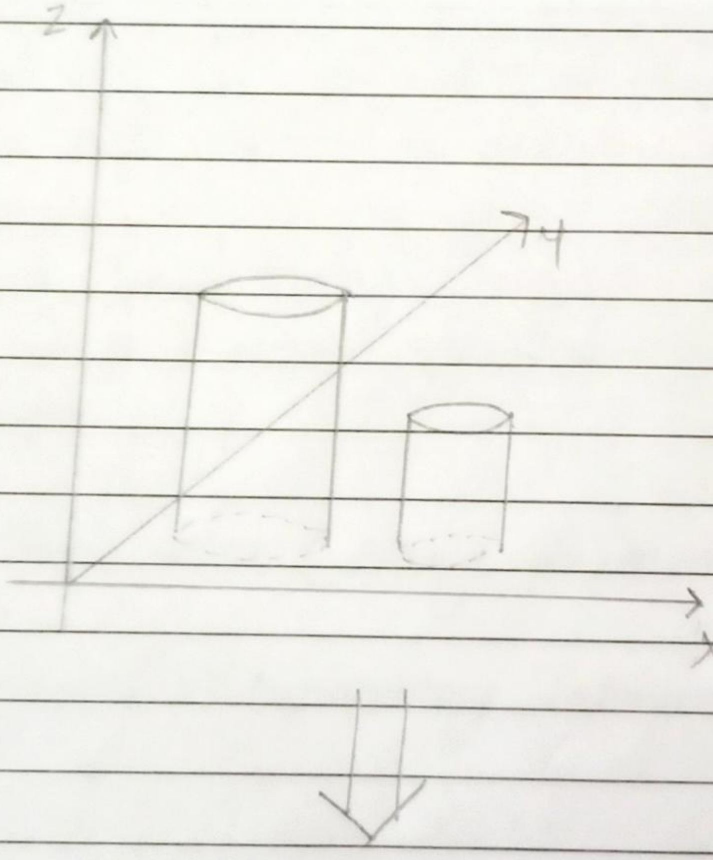
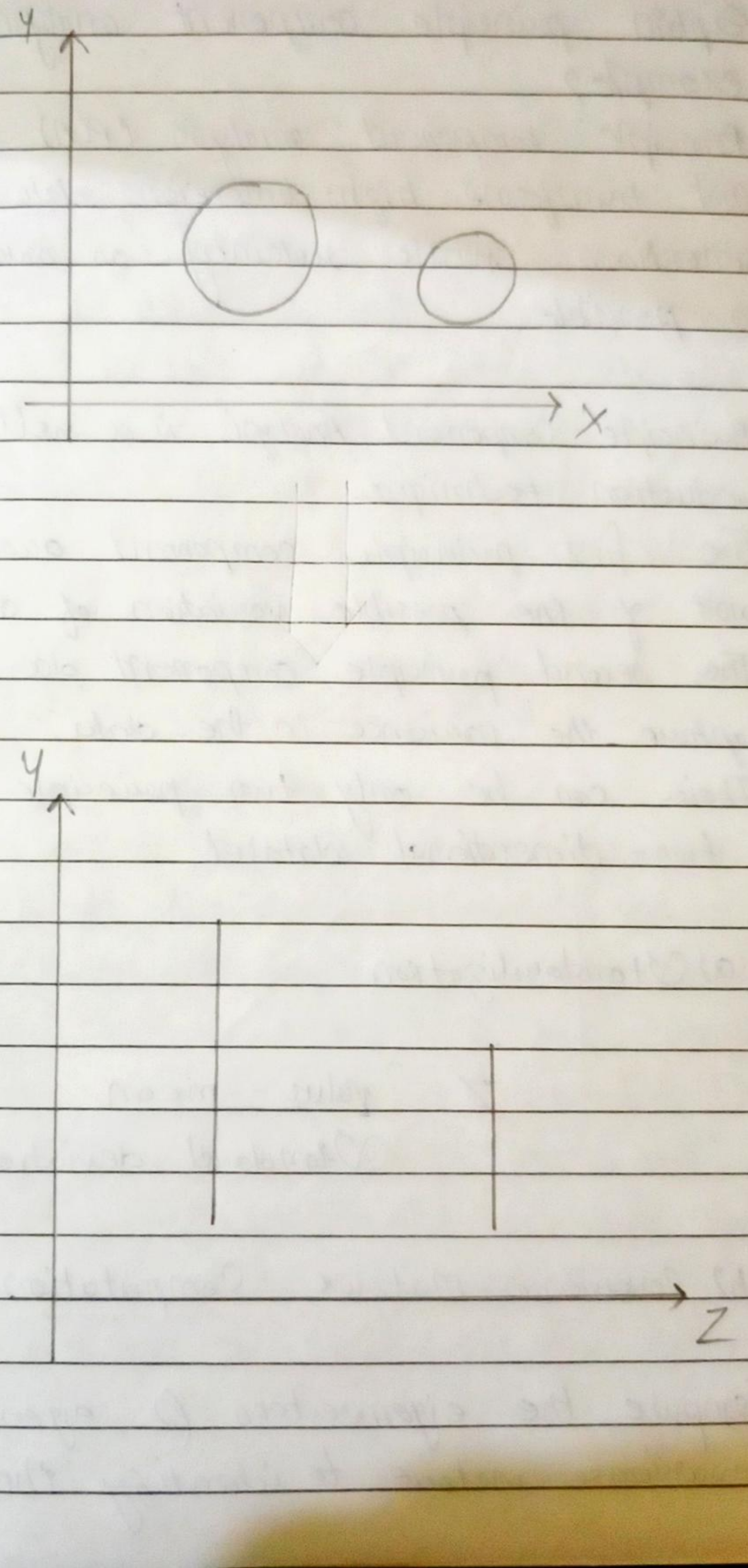
Assignment 2

1 Why is dimentionality reduction important?

Ans Dimensionality reduction can be discussed through a simple e-mail classification problem, where we need to classify whether the e-mail is spam or not. In a classification problem that relies on both humidity & rainfall can be collapsed into just one underlying features, since both of the dimentioned features are correlated to a high degree. Hence, we can reduce the number of features in such problems.




Y
X
Y
Z

2 Explain principle component analysis with an example?

Ans Principle component analysis (PCA) is a technique that transforms high-dimensions data into lower-dimentions while retaining as much information as possible.

- Principle Component Analysis is a well known dimension reduction technique.
- The first principle component accounts for the most of the possible variation of original data.
- The second principle component does its best to capture the variance in the data.
- Their can be only two principle components for a two-dimentional dataset.

a) Standardization

$$Z = \frac{\text{value} - \text{mean}}{\text{Standard deviation}}$$

b) Covarience matrix Computation

Compute the eigenvectors & eigenvalues of the covariance matrix to identify the principle components

3 What is clustering? Explain the k-means clustering algorithm?

Ans Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Clustering is a process of partitioning a set of data in set of meaningful sub-classes, called clusters.

A cluster is therefore a collection of objects which are similar between them & are dissimilar to objects belonging to other clusters

- K-means algorithm -

K means clustering is an algorithm to classify or to group the different object based on attributes or features into k number of groups. k is a positive integer number. Group the elements into the clusters which are nearer to the centaroid of that cluster. Follow the same methods & group the elements based on new centroid. Do the same process till no element is moving from one cluster to another.

4 Write a short note on hierachical clustering & partitional clustering?

Ans Hierachical Clustering - A Hierachical clustering methods works via grouping data into a tree of clusters. Hierachical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequently takes steps Identify the two clusters which can be closest together, & merge the two maximum comparable clusters.

There are two types of hierachical cluster, Divisive & Agglomeration

Partitional Clustering - A data set into a sets of disjoint clusters. Given a dataset of N points, a partitioning method constructs K (N ≥ K) partitions of the data, with each partition representing a cluster.

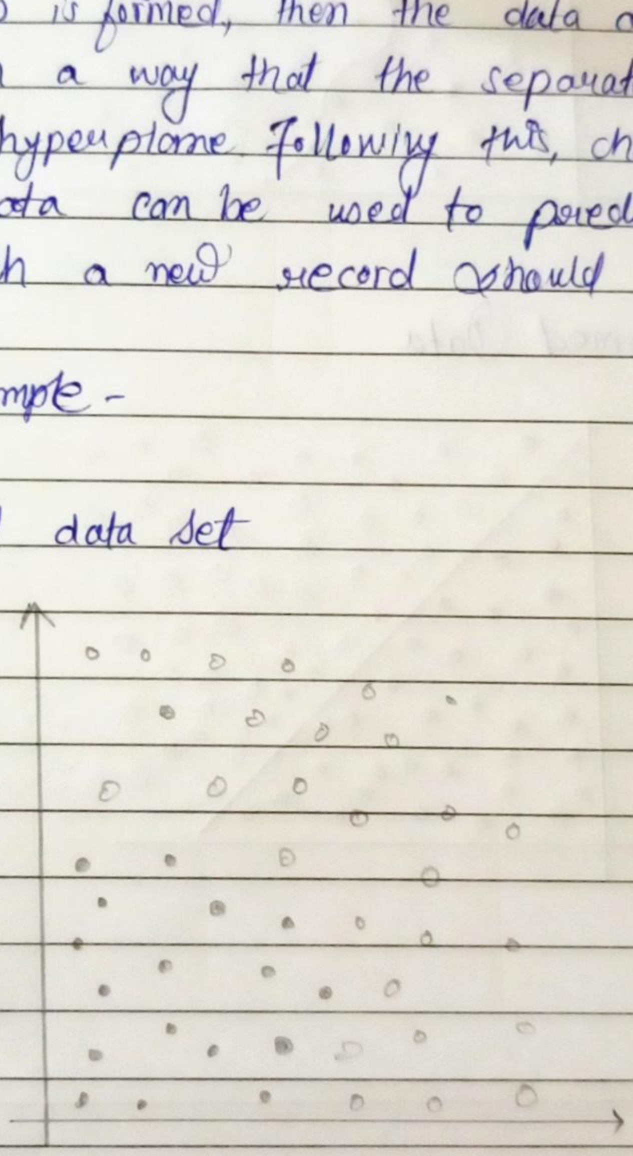
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other datapoints in the same group & dissimilar to the data points in other groups.

5 What is SVM ? Explain it ?

Ans Support vector Machine works by mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable A separator between the categories is formed, then the data are transformed in such a way that the separator could be drawn as a hyperplane Following this, characteristics of new data can be used to predict the group to which a new record should belong.

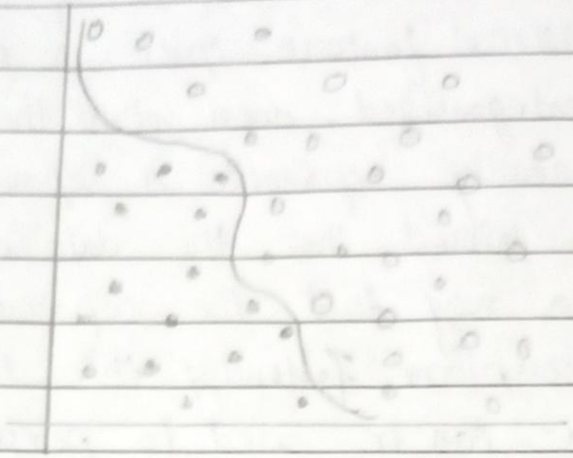
for example -

a) Original data set



The two categories can be separated

b) Data with separator added



c) Transformed Data