Overview of Data Analytical Algorithms

-Sohrab Ardeshar Vakharia

What is Data Analytics?

- As the process of analyzing raw data to find trends and answer questions, the definition of data analytics captures its broad scope of the field. However, it includes many techniques with many different goals.
- Data analytics is the science of analyzing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

Data analytics algorithms

- Classification
- Clustering
- Decision Trees
- K-nearest neighbour
- K-means clustering
- Regression

- And many such..

Classification

- Classification could be one of the biggest problem to a dataset. Unless we make it a boon to ourselves. Data is vast in our context and in all different forms. I would like the term "Big Data" for it. This dataset needs cleansing, a lot of it. Where things can be classified to make it simpler to understand and work. To understand classification better we need to understand machine learning and how it works.
- when the classes are defined and we need to segregate the data into the defined classes. *For example,* the output is categorized as different colors "Red", "Blue", "Green" or different categories of vehicles such as "Bikes", "Cars", "Trucks".

Classification (Regression)

- It is when the outcome is based on real values such as “Rupees”, “Years”, “Weight”.
- For which an online definition says “A technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, change in one or more independent variables.”

Regression Analysis

Regression analysis attempts to explain the influence that a set of variables has on the outcome of another variable of interest. Often, the outcome variable is called a *dependent variable* because the outcome depends on the other variables. These additional variables are sometimes called the *input variables* or the *independent variables*. Regression analysis is useful for answering the following kinds of questions:

- What is a person's expected income?
- What is the probability that an applicant will default on a loan?

Linear regression is a useful tool for answering the first question, and logistic regression is a popular method for addressing the second.

Regression Analysis

- Regression analysis is a useful explanatory tool that can identify the input variables that have the greatest statistical influence on the outcome. With such knowledge and insight, environmental changes can be attempted to produce more favorable values of the input variables. For example, if it is found that the reading level of 10-year-old students is an excellent predictor of the students' success in high school and a factor in their attending college, then additional emphasis on reading can be considered, implemented, and evaluated to improve students' reading levels at a younger age.

Learning Decision Trees

A decision tree represents classification. Decision tree learning is the most promising technique for supervised classification learning. Since it is a decision tree it is mend to take decision and being a learning decision tree it trains itself and learns from the experience of set of input iterations. These input iterations are also well known as “input training sets” or “training set data”.

Decision trees predict the future based on the previous learning and input rule sets. It taken multiple input values and returns back the probable output with the single value which is considered as a decision. The input/output could be continuous as well as discrete. A decision tree takes its decision based on the defined algorithms and the rule sets.

Learning Decision Trees Example

For example you want to take a decision to buy a pair of shoes. We start with few set of questions:

- 1. Do we need one?
- 2. What would be the budget?
- 3. Formal or informal?
- 4. Is it for a special occasion?
- 5. Which colour suits me better?
- 6. Which would be the most durable brand?
- 7. Shall we wait for some special sale or just buy one since its needed?

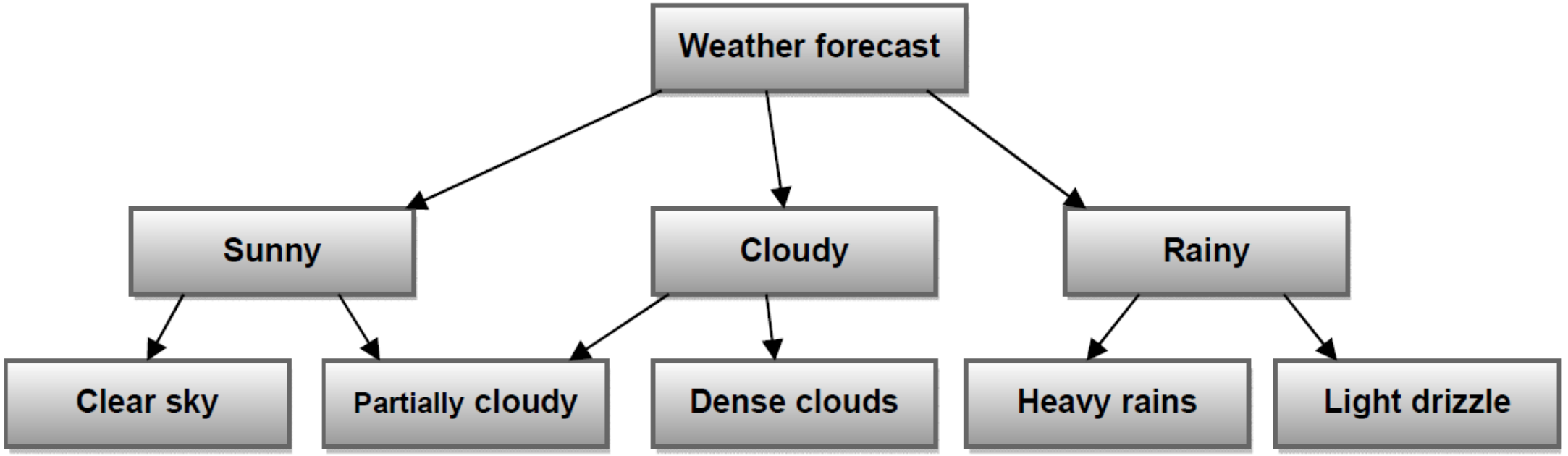


Fig. 6.1: Example Showing the Decision Tree of Weather Forecast

The above figure shows how a decision needs to be taken in a weather forecast scenario where the day is specified as Sunny, Cloudy or Rainy. Depending upon the metrics received by an algorithm it will take the decision. The metrics could be humidity, sky visibility and others. We can also see the cloudy situation having two possibilities of having partially cloudy and dense clouds, wherein having partial clouds is also a subset of a Sunny day. Such occurrences make decision tree bivalence.

Evaluation of classification Models

We have 4 known categories of the classification models:

1. **Heuristic models**
2. **Separation Models (divide and rule)**
3. **Regression models**
4. **Probabilistic models**

Best Classification Model:

While considering one best fit model for our classification we need to consider the following mention points:

Speed: Amount computing time taken to get the solution.

Accuracy: Correctness in the model in regards to the outcome.

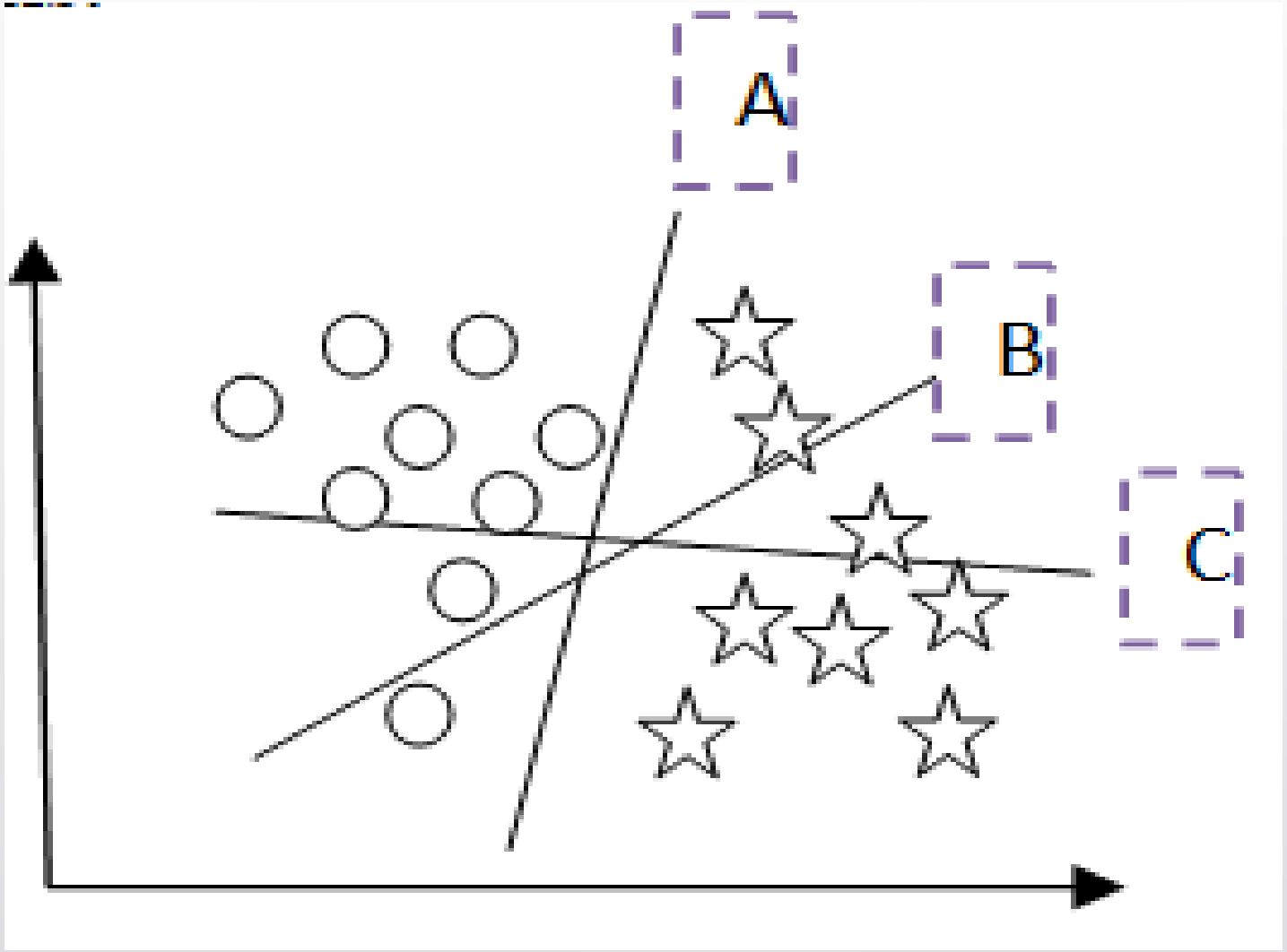
Scalability: It should be able to consider the various considerations in the large datasets.

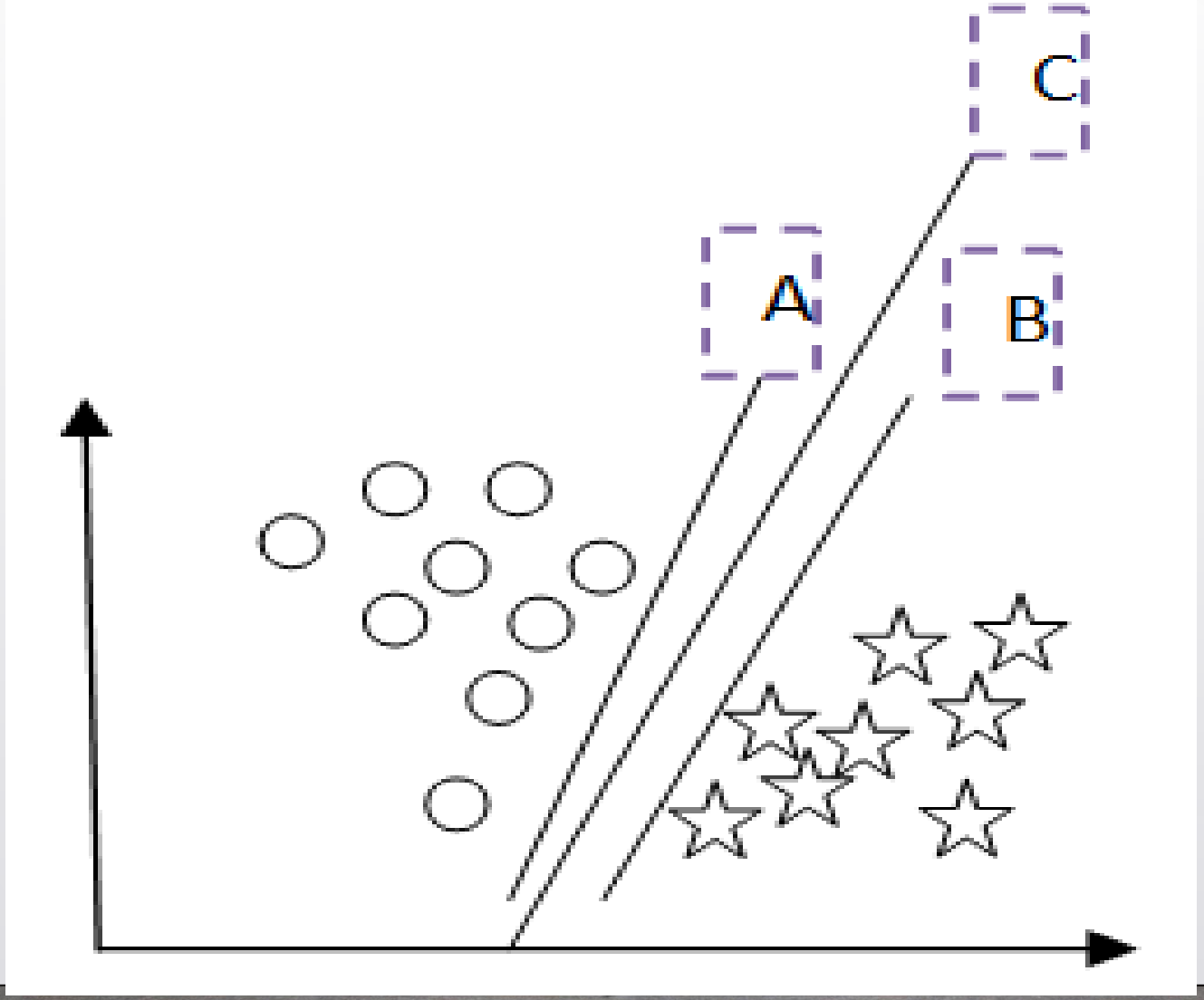
Interpretability: Easier for its users to understand and work.

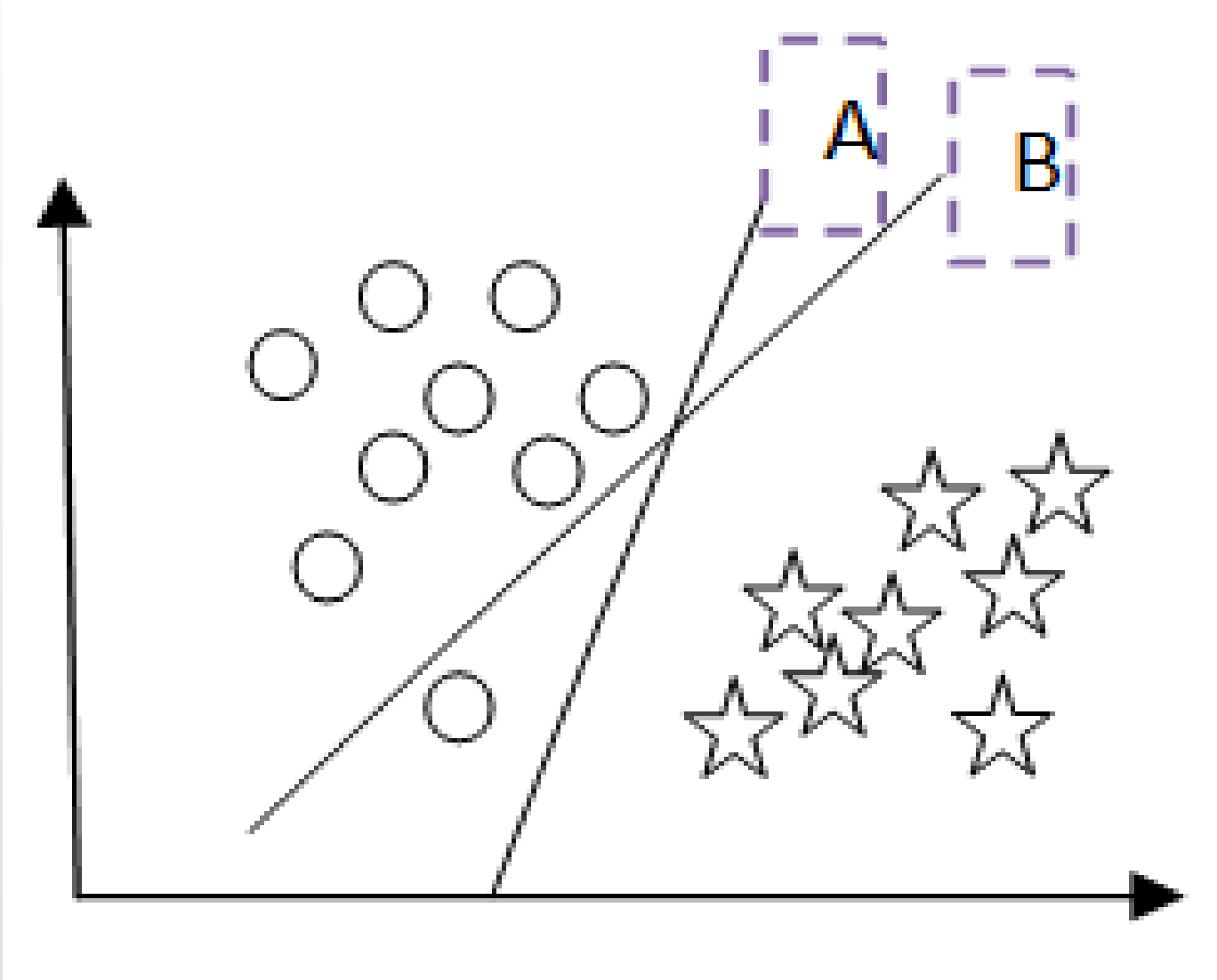
Robustness: With less loopholes/errors/bugs. As nothing is perfect but it should be next to perfect.

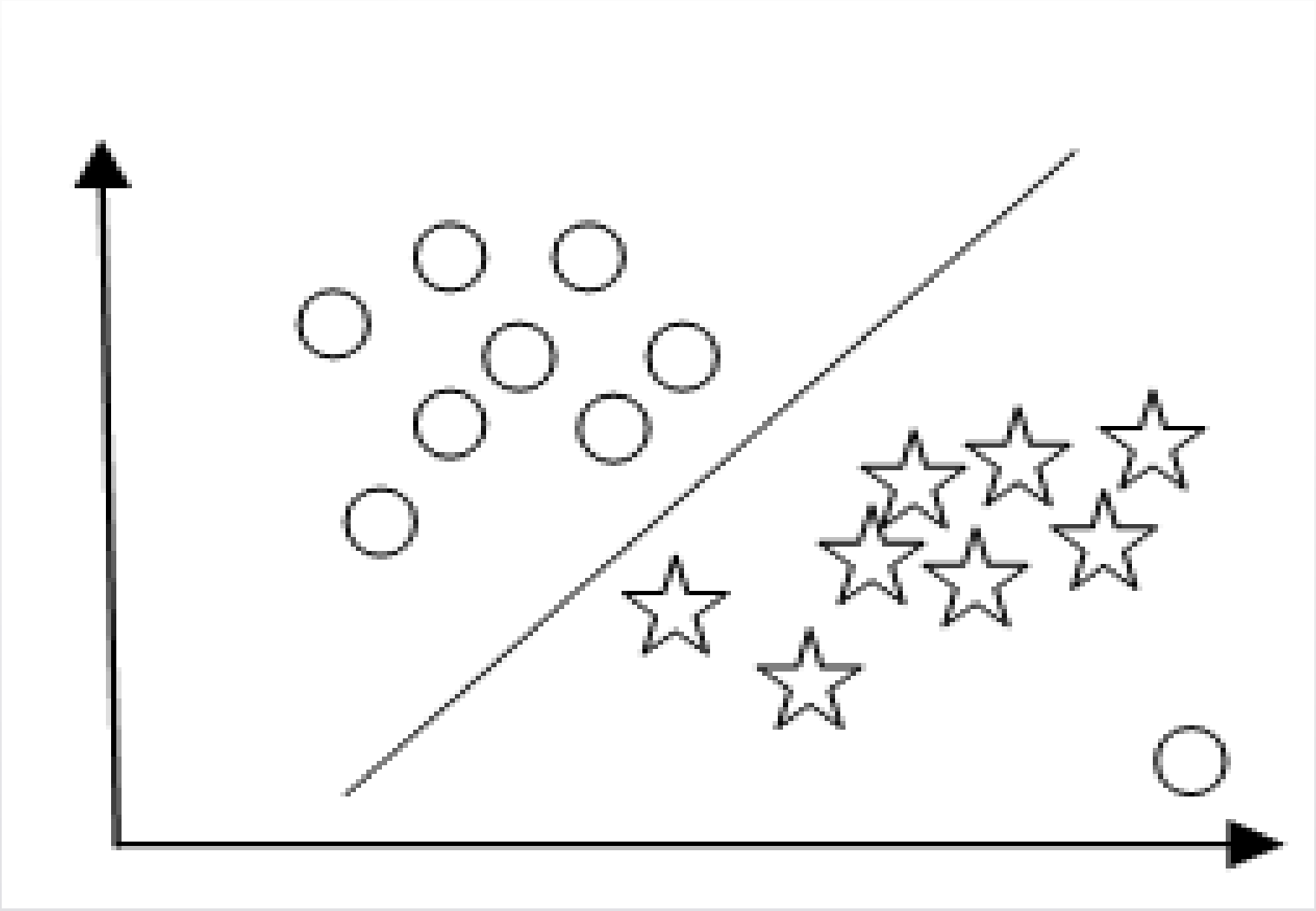
Support vector Machine

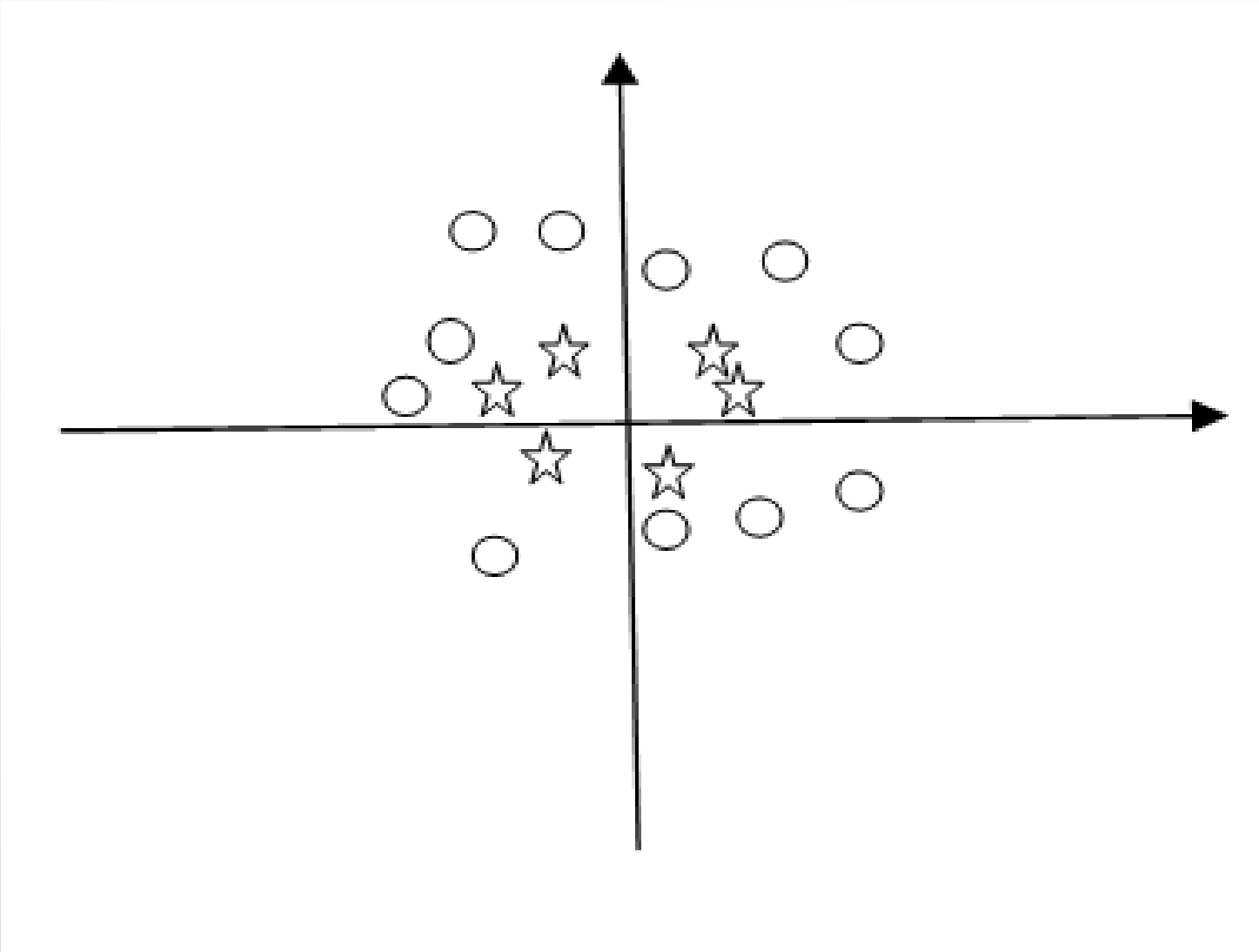
- Support vector machine is an algorithm which is used for classification in a supervised learning algorithm example. It does classification of the inputs received on the basis of the rule-set. It also works on Regression problems.

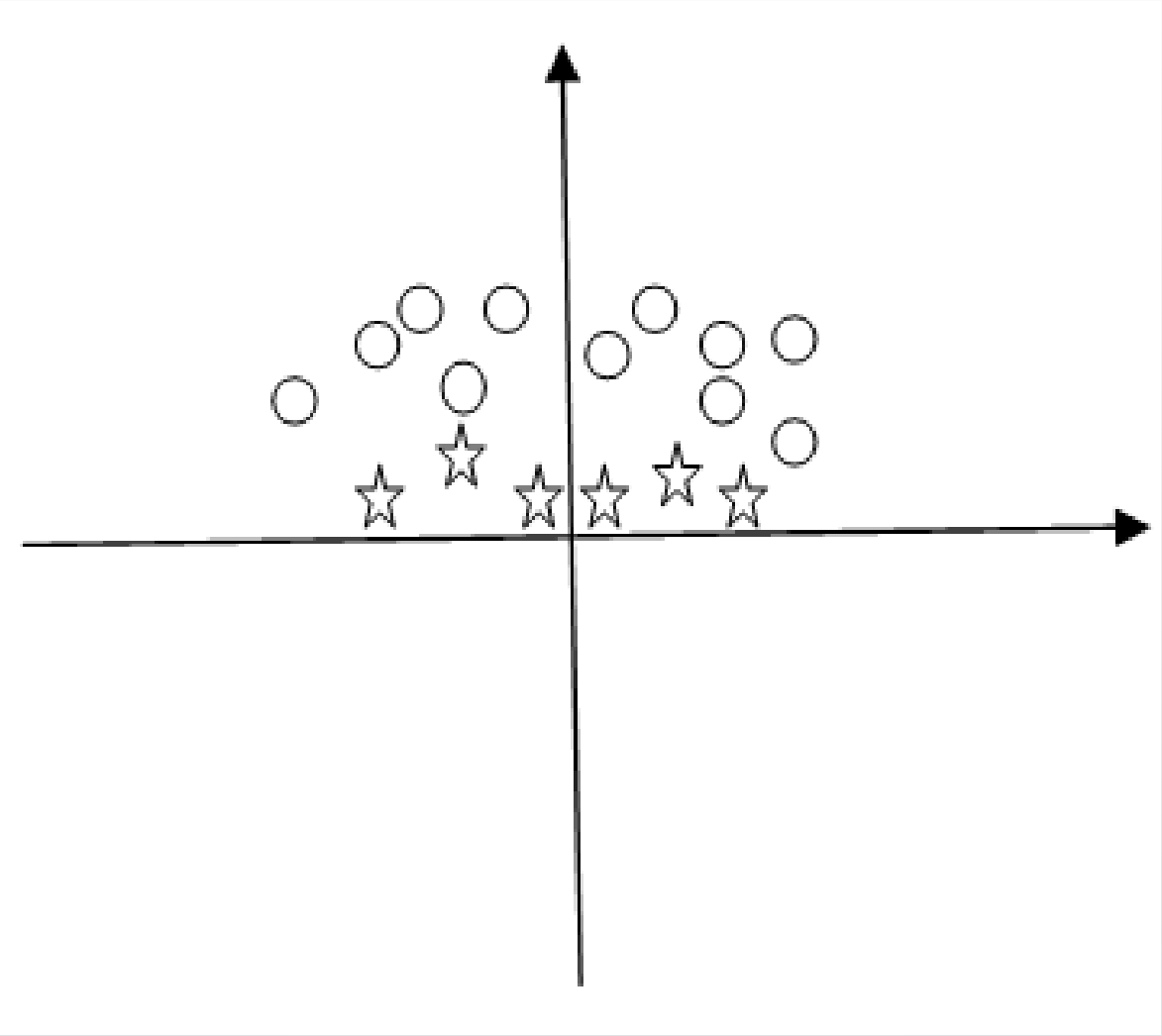

A
B
C


C
A
B


A
B







Clustering

Clustering is an unsupervised learning model, similar to classification, it helps creating different set of classes together. It groups the similar types together by creating/identifying the clusters of the similar types. Clustering is a task of dividing homogeneous data types or population or groups. It does so by identifying the similar data types or nearby data elements over graph. In classification the classes are defined with the help algorithms or predefined classes are used and then the data inputs is considered, while in clustering the algorithm inputs itself decides with the help of inputs, the number of clusters depending upon it similarity traits. These similar set of inputs forms a group and called as clusters. Clustering is more dynamic model in terms of grouping.

Types of Clustering

Soft clustering and hard clustering.

Let me give one example to explain the same. For an instance we are developing a website for writing blogs. Now your blog belongs to a particular category such as: Science, Technology, Arts, Fiction etc. It might be possible that the article which is written could belong or relate 2 or more categories. Now, in this case if we restrict our blogger to choose one of the category then we would call this as “hard or strict clustering method”, where a user can remain in any one of the category. Let say this work is done automated by our piece of code and it chooses categories on the basis of the blog content. If my algorithm chooses any one of the given cluster for the blog then it would be called as “hard or strict clustering”. In contradiction to this if my algorithm chooses to select more than one cluster for the blog content then it would be called as “ soft or loose clustering” method.

Clustering Algorithms and Methods

- **Connectivity Clustering Method:** This model is based on the connectivity between the data points. These models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away.
- **Clustering Partition Method:** It works on divisions method, where the division or partition between data set is created. These partitions are predefined non-empty sets. This is suitable for a small dataset.

Clustering Algorithms and Methods:

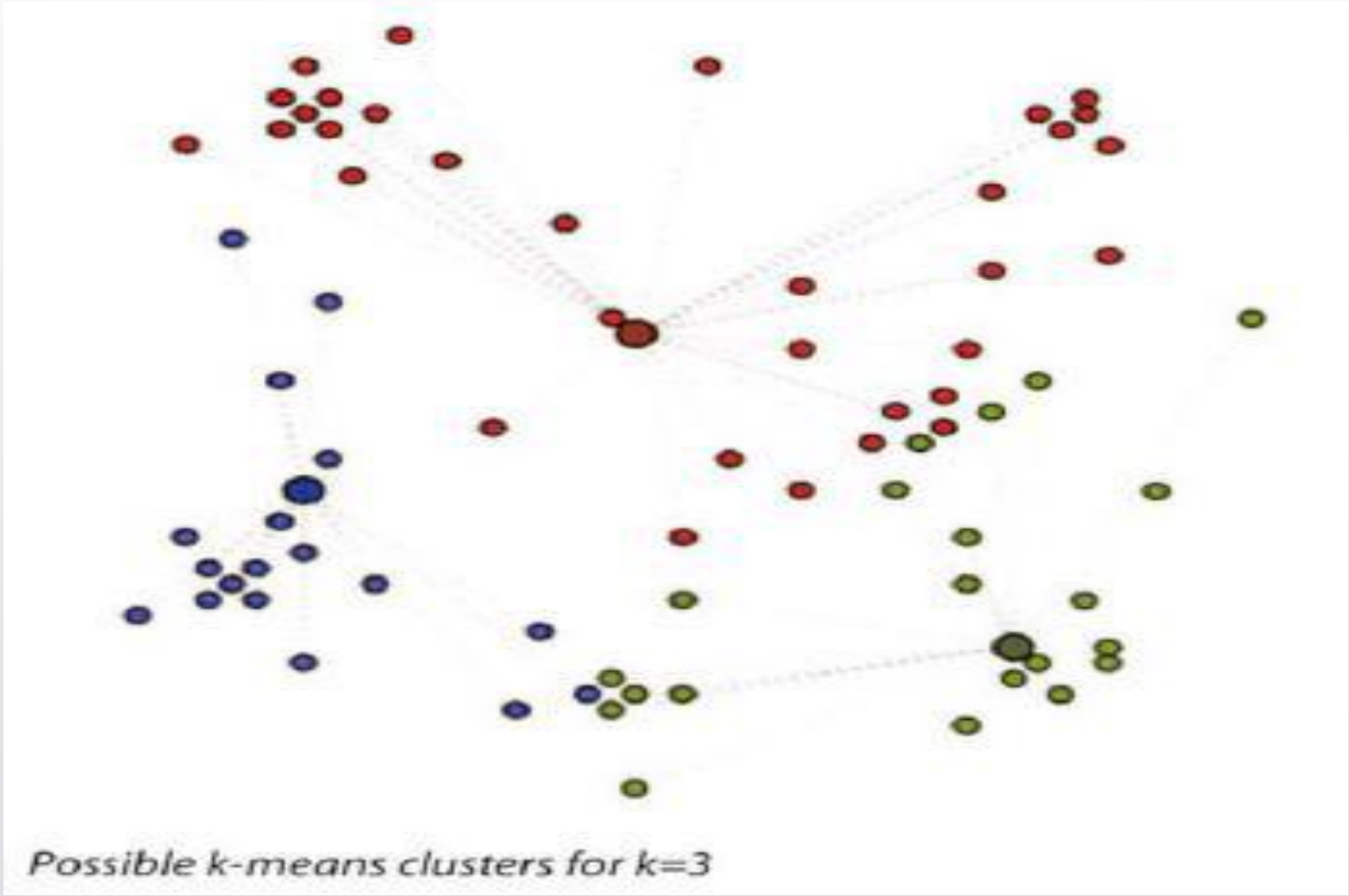
- **Centroid Cluster Method:** This model revolve around the center element of the dataset. The closest data points to the center data point (centroid) in the dataset is considered to form a cluster. K-Means clustering algorithm is the best fit example of such model.
- **Hierarchical clustering Method:** This method describes the tree based structure (nested clusters) of the clusters. In this method we have clusters based on the divisions and their sub divisions in a hierarchy (nested clustering). The hierarchy can be pre-determined based upon user choice. Here number of clusters could remain dynamic and not needed to be predetermined as well.

Clustering Algorithms and Methods

- **Density-based Clustering Method:** In this method the density of the closest dataset is considered to form a cluster. The more number of closer data sets (denser the data inputs), the better the cluster formation. The problem here comes with outliers, which is handled in classifications (support vector machine) algorithm.

K-Means Clustering

- K-means is an unsupervised method which works iteratively to assign each data point to one of the group of “K” groups based on the features.
- Data points are clustered based on the similarity of the features where the centroids of the K clusters are used to cluster the datasets.
- The datasets nearby to the centroids are considered in this scenario.



Possible k-means clusters for k=3

Association Rules

- An unsupervised learning method called association rules. This is a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset. The disclosed relationships can be represented as rules or frequent item sets. Association rules are commonly used for mining transactions in databases.

Here are some possible questions that association rules can answer:

• Which products tend to be purchased together?

• Of those customers who are similar to this person, what products do they tend to buy?

• Of those customers who have purchased this product, what other similar products do they tend to view or purchase?

Thank You

-Sohrab Ardeshar Vakharia