Data Science and Machine Learning

Vasudevan T V

Course Contents

- Module 1 - Data Science, Data Visualisation
- Module 2 - Introduction to machine learning, Lazy learning, Probabilistic learning
- Module 3 - Decision tree learning, Classification rules learning, Regression methods
- Module 4 - Neural network learning, Support vector machines
- Module 5 - Clustering, Evaluating model performance, Improving model performance

Module 1
Introduction to Data Science

- Data science is a collection of techniques used to extract value from data
- It has become an essential tool for any organisation that collects, stores, and processes data as part of its operations
- Data science techniques find useful patterns, connections, and relationships within data
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining
- However, each of these terms have a different meaning depending on the context

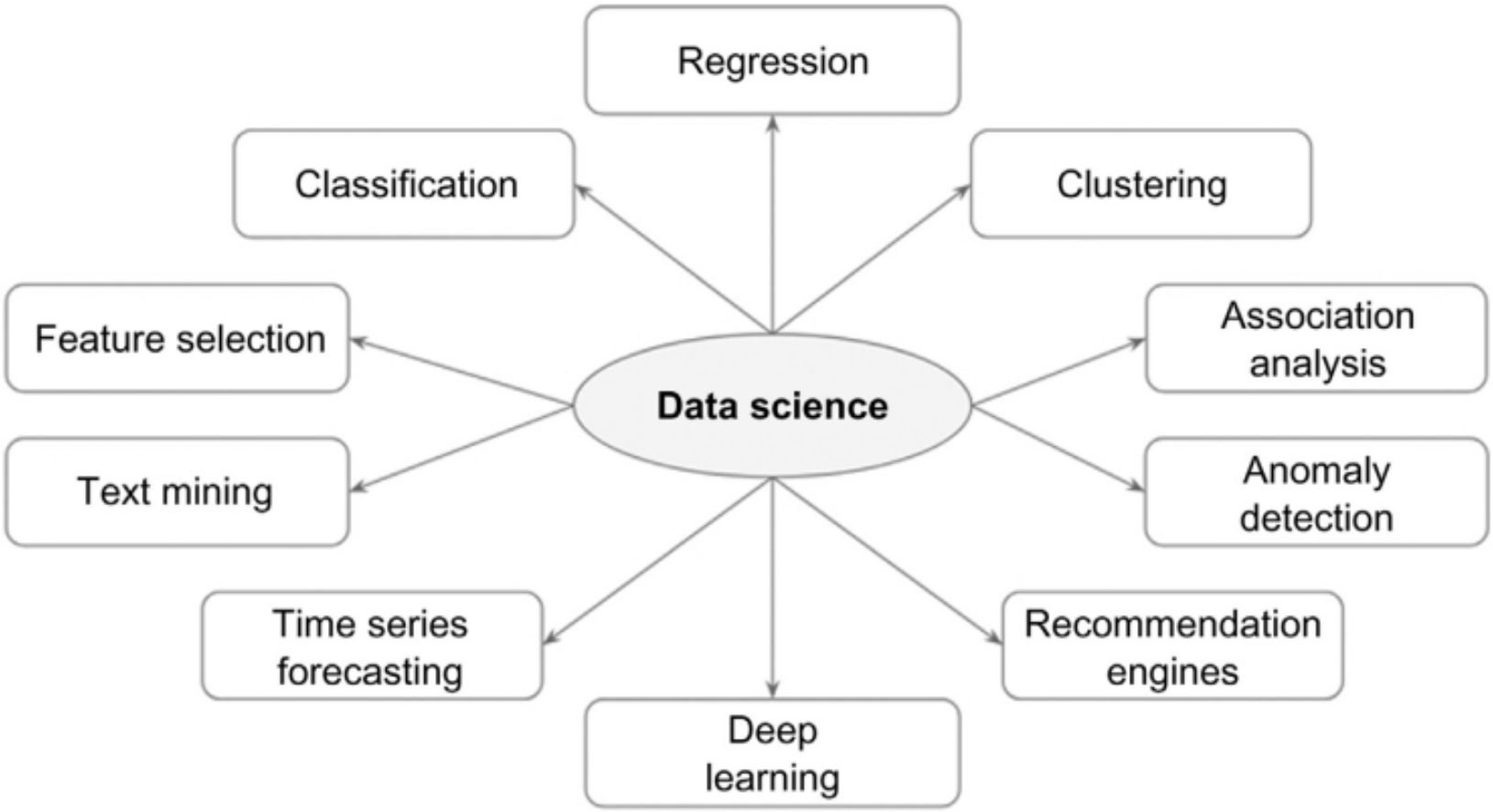
Data Science Classification

- Data Science problems are broadly classified into two types

1. Supervised Learning Model
   - In this model, training is given to the machine for solving problems
   - Simple Model
   - Highly Accurate
2. Unsupervised Learning Model
   - Here no training is given to the machine for solving problems
   - Complex Model
   - Less Accurate

Data Science Classification

- Data Science Tasks



Data Science Classification

- Classification and Regression techniques predict a target variable based on input variables
- The output variable which is predicted is called a target variable
- In classification, the target variable is a category such as 'yes', 'no', 'red', 'blue' etc.
- Example - Predicting whether monsoon will be normal this year
- In regression, the target variable is a numeric value
- Example - Predicting the age of a person
- Clustering is the process of identifying natural groupings within a data set
- Example - Grouping books in a library based on topics

Data Science Classification

- Association Analysis involves identifying associations or relationships within a data set
- Example - Finding which all items are bought together from a store
- Anomaly Detection is the process of identifying data points which are significantly different from other data points in a data set
- Example - Detecting fraudulent credit card transactions
- Recommendation Engines recommend items to users based on their preference / behaviour
- Example - Recommend items to users based on shopping behaviour
- Deep Learning is a machine learning model based on artificial neural networks, which are inspired by the functioning of human brain

Data Science Classification

- Time Series Forecasting involves predicting the future value of a variable based on its historical values
- Example - Predicting Temperature
- Text Mining is the process of retrieving information from text data such as documents, messages, email, web pages etc.
- Example - Web Search Engine
- Feature Selection is the process of selecting the most relevant features(attributes) to get the required output in the algorithm
- Example - Old Cars Data Set (Model, Year, Kms, Owner)
- Here we can discard 'Owner' details and select the other attributes for determining the cars to be crushed for spare parts

Data Science Process

- The method of discovering useful relationships and patterns in data is called the data science process
- Steps

1. Prior Knowledge
2. Data Preparation
3. Modelling
4. Application

Data Science Process

1. Prior Knowledge

- Here we define what problem is being solved
- We find out what data is needed for solving the problem
- Example - Consumer Loan Business
- Problem - If the interest rates and credit scores of past borrowers are known, can we predict the interest rate of a new borrower?
- Data - A sample data set of 10 data points with 3 attributes : borrower id, credit score, and interest rate

Data Science Process

1. Prior knowledge

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

- Credit Score is a measure of the ability of a borrower to repay the loan
- Larger the credit score, greater the ability to repay the loan

Data Science Process

1. Prior knowledge

- A data set is a collection of data with a well defined structure
- Here, table is the data set
- A data point is a single instance of the data set
- Here, each record in the table is a data point
- An attribute is a single property of the data set
- Here, each column in the table is an attribute
- An identifier is a special attribute used for locating data points in a data set
- Here, Borrower Id is the identifier

Data Science Process

1. Prior knowledge

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

- A label is the special attribute to be predicted based on all the input attributes
- Here interest rate of the new borrower is to be predicted

Data Science Process

2. Data Preparation

- In this stage we prepare the whole data set needed for the data science task
- Steps

2.1 Data Exploration

- This involves in depth analysis of data to gain better understanding about it
- For this, statistical analysis and visualisation tools are used

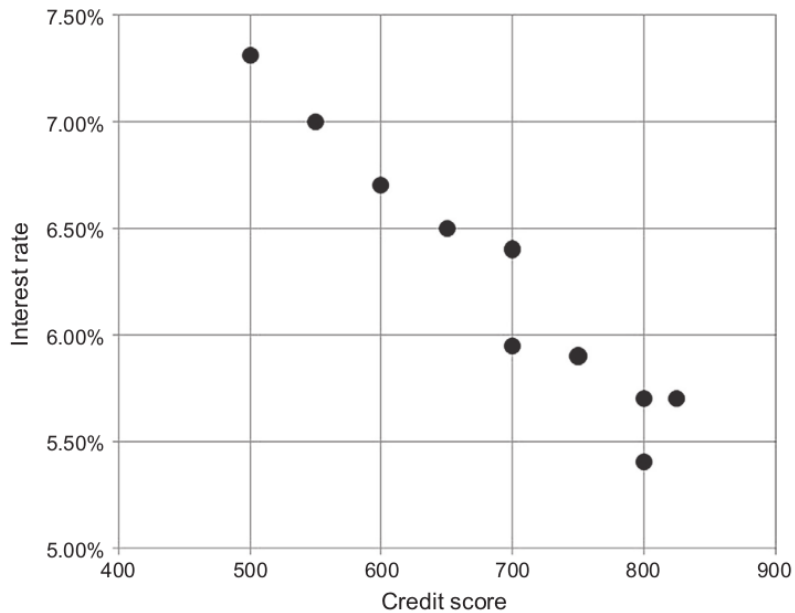


FIGURE 2.3
Scatterplot for interest rate dataset.

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

1. Elimination of Duplicate Records
2. Dealing with Missing Values

- Example - Missing Credit Score
- It can be replaced with a credit score derived from the data set(mean)
- Alternatively, we can eliminate records with missing values

3. Data Type Conversion

- Depending on the requirement, we convert data from one type to another
- This depends on the data science algorithm we are using
- We can convert credit score to categorical values such as poor = 400, good = 600, excellent = 800

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

4 Transformation of Attribute Ranges

- Different attributes have different ranges
- For example, range of income is larger compared to range of credit score
- For some data science algorithms, these ranges are normalised to a uniform scale from 0 to 1

5 Handling Outliers

- Outliers are anomalies in a data set
- Example - Human height as 1.73cm instead of 1.73m
- We need to correct these anomalies

Data Science Process

2.3 Feature Selection

- All the attributes in the data set may not be needed for solving the problem
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection
- This leads to a simplified model

2.4 Data Sampling

- It involves selecting a subset of the original data set for analysis
- It reduces the amount of data to be processed
- It can speed up the process of analysis

Data Science Process

3 Modelling

- A model is the abstract representation of the data and the relationships in a given data set
- There are 2 kinds of data sets associated with a model
- The data set used to create the model is called a training data set
- The data set used to validate the model is called a test data set
- The entire data set is split into training data set and test data set
- A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test data set

Data Science Process

Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Data Science Process

3 Modelling

- Now we will evaluate the model using the test data set
- We will be using simple linear regression technique for predicting interest rates of test data set

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)	Model Predicted (*Y*) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

4 Application

- In this stage we present our findings to the world
- We can build an application that automatically updates reports, spreadsheets and presentation slides

Data Exploration

- Data exploration is the detailed analysis of data to gain better understanding of it
- It can be classified into two types

1. Descriptive Statistics

- Here we summarise the data using various statistical measures such as mean, median, mode etc.

2. Data Visualisation

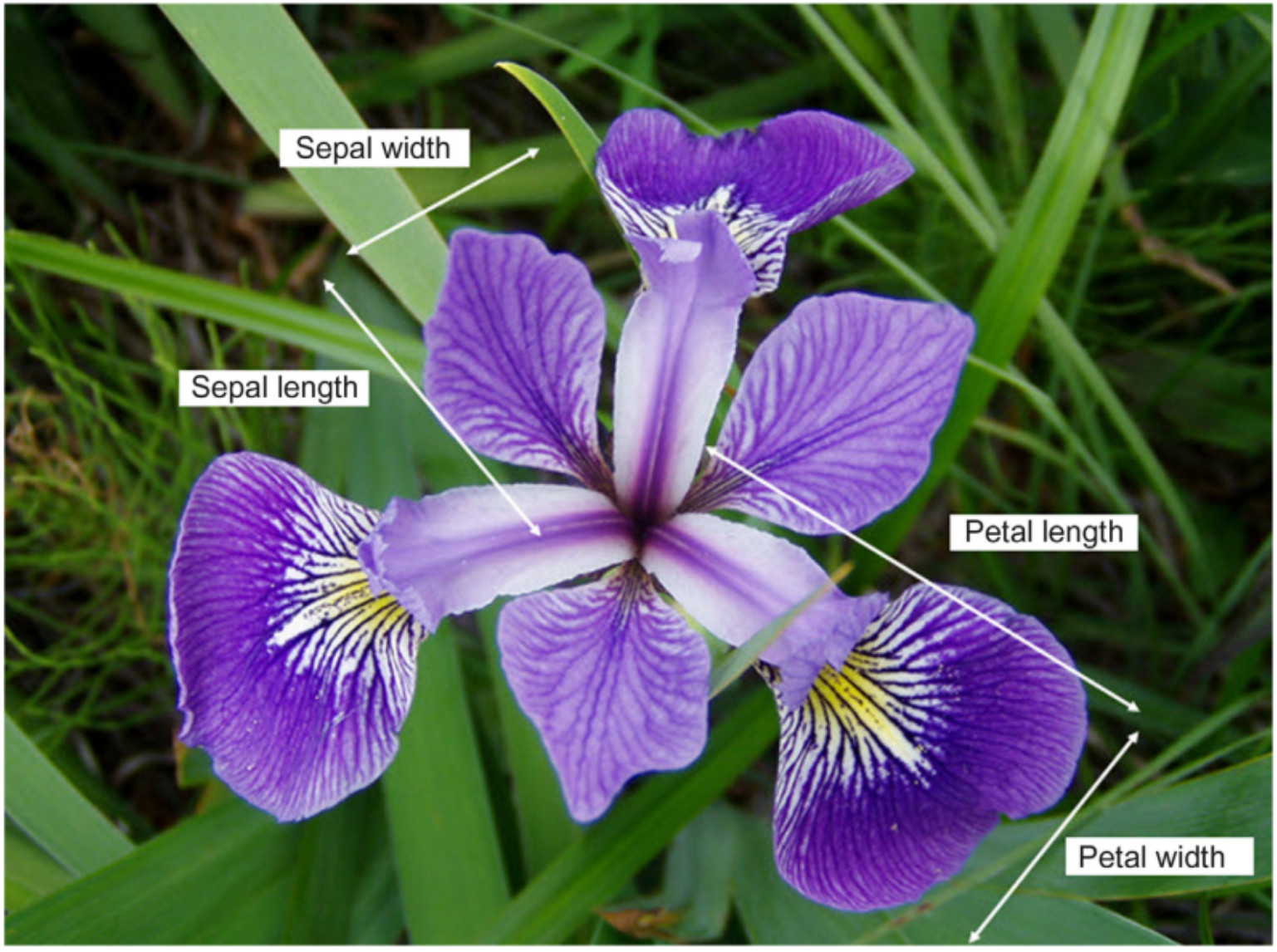
- Here we make visual representation of data using various charts like histogram, scatter plot, bubble chart etc.

Data Exploration

- Data Sets
- We will be considering a data set about Iris, a flowering plant
- Each observation has 4 attributes - sepal length, sepal width, petal length, petal width
- Sepal is the outer part of a flower
- Petal is the inner part of a flower
- Sepals support petals during their growth

Data Exploration

- Iris Flower



Data Exploration

Table 3.1 Iris Dataset and Descriptive Statistics (Fisher, 1936)

Observation	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.1	1.5	0.1
...	...	...	...	...
49	5	3.4	1.5	0.2
50	4.4	2.9	1.4	0.2
Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.006	3.418	1.464	0.244
Median	5.000	3.400	1.500	0.200
Mode	5.100	3.400	1.500	0.200
Range	1.500	2.100	0.900	0.500
Standard deviation	0.352	0.381	0.174	0.107
Variance	0.124	0.145	0.030	0.011

Data Exploration

- Descriptive Statistics can be classified into two types

1 Descriptive Statistics for Univariate Data

- Here analysis of one attribute is done at a time

a Mean - It is the average of all observations in the data set

b Median - It is the value of the central point in the distribution

c Mode - It is the most frequently occurring observation

d Range - It is the difference between the maximum value and minimum value of the attribute

e Variance - It is a measure of how data points differ from the mean

f Standard Deviation - It is the square root of variance

Data Exploration

- Descriptive Statistics can be classified into two types
2. Descriptive Statistics for Multivariate Data
- Here analysis of multiple attributes is done at a time

a. Correlation

- It measures the statistical relationship between attributes
- It measures the dependence of one attribute on another
- Example - There is correlation between average temperature of a day and ice cream sales
- Correlation between two attributes is commonly measured by the Pearson correlation coefficient
- It ranges from -1 to 1, where negative values indicate negative correlation, positive values indicate positive correlation and 0 indicates no correlation
- -1 and 1 indicate perfect correlation

Data Visualisation

- For better understanding, visual representation of data is done using various charts
- Data Visualisation can also be classified into two types

1 Univariate Visualisation

- Here visualisation of one attribute is done at a time
- Charts - Histogram, Quartile Plot, Distribution Chart

2 Multivariate Visualisation

- Here visualisation of multiple attributes is done at a time
- Charts - Scatter Plot, Bubble Chart, Density Chart

Histogram

- It plots the frequency of occurrence of data within different ranges

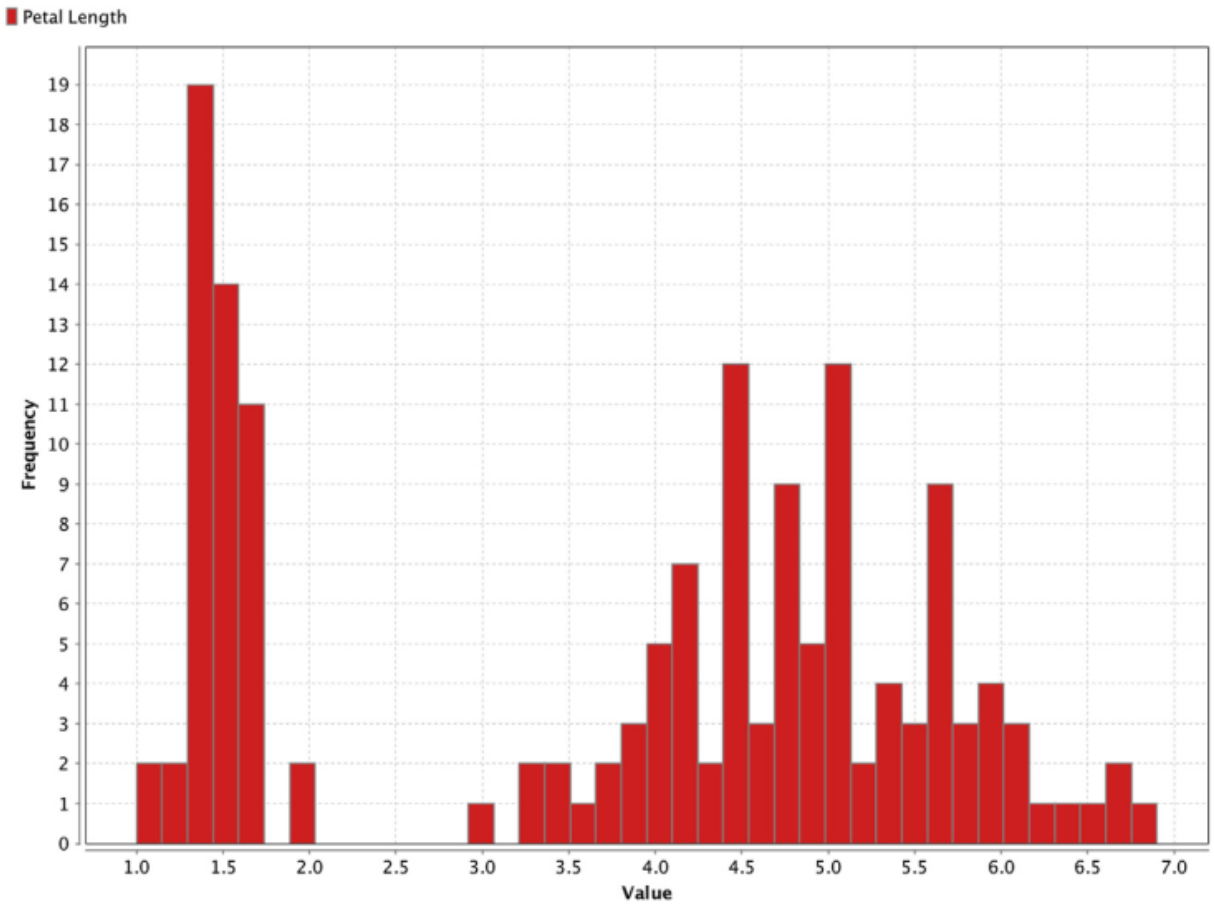


FIGURE 3.5

Histogram of petal length in Iris dataset.

Quartile Plot

- It plots the quartiles, outliers, mean and standard deviation
- The quartiles are denoted by Q1, Q2 and Q3
- In a distribution, 25% of the data points will be below Q1, 50% will be below Q2, and 75% will be below Q3
- The Q1 and Q3 points in a quartile plot are denoted by the edges of the box
- The Q2 point, the median of the distribution, is indicated by a cross line within the box
- The outliers are denoted by circles
- The mean point is denoted by a solid dot overlay and standard deviation as a line overlay.

Quartile Plot

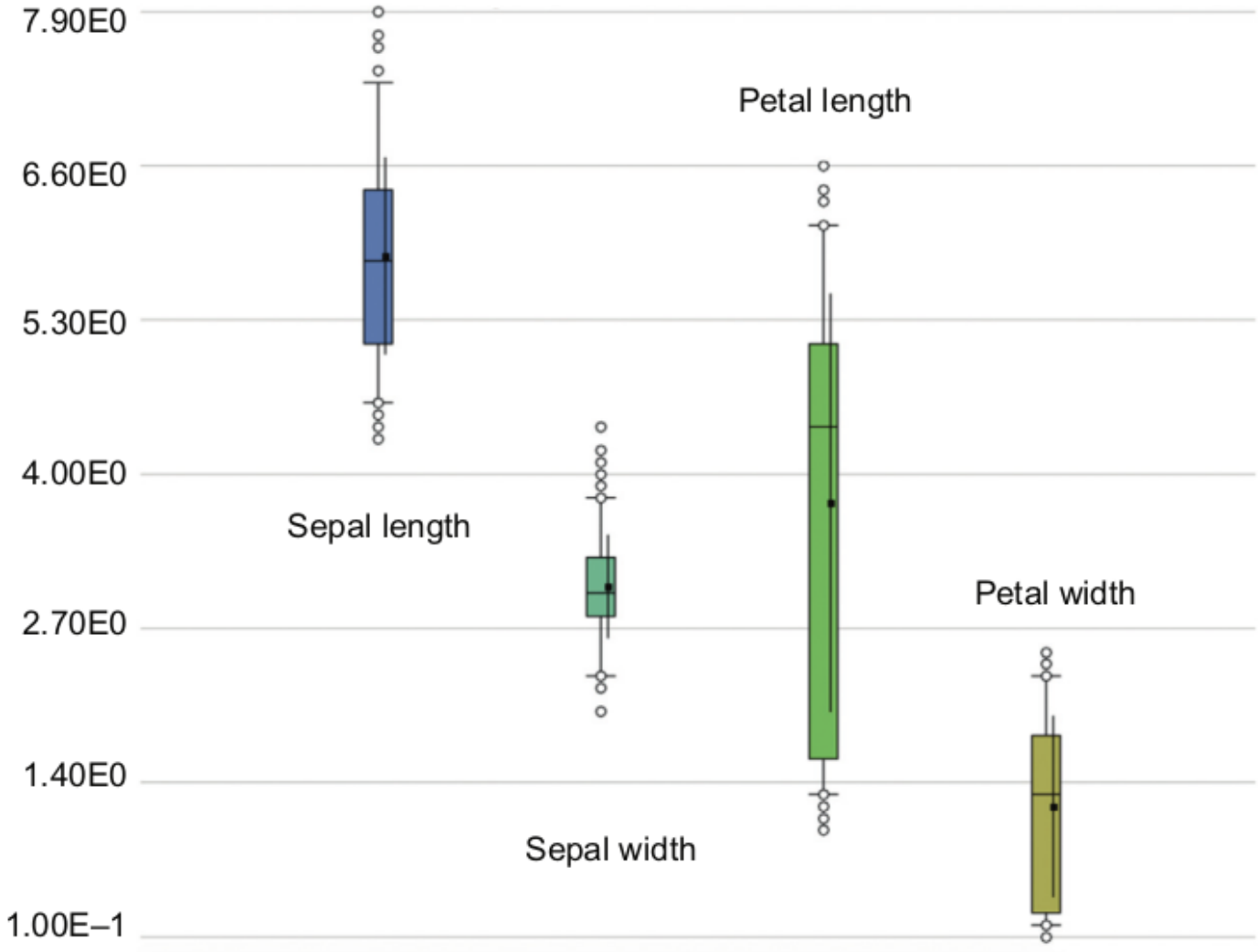


FIGURE 3.7
Quartile plot of Iris dataset.

Distribution Chart

- It shows the normal distribution function of the data
- It is also called the bell curve, due to its bell shape
- It shows the probability of occurrence of a data point within a range of values
- Example - Distribution charts of 3 different Iris species
- From the chart, we can predict that, an Iris flower of petal length 1.5 cm belongs to setosa species
- But we cannot predict the species of petal length 5 cm

Distribution Chart

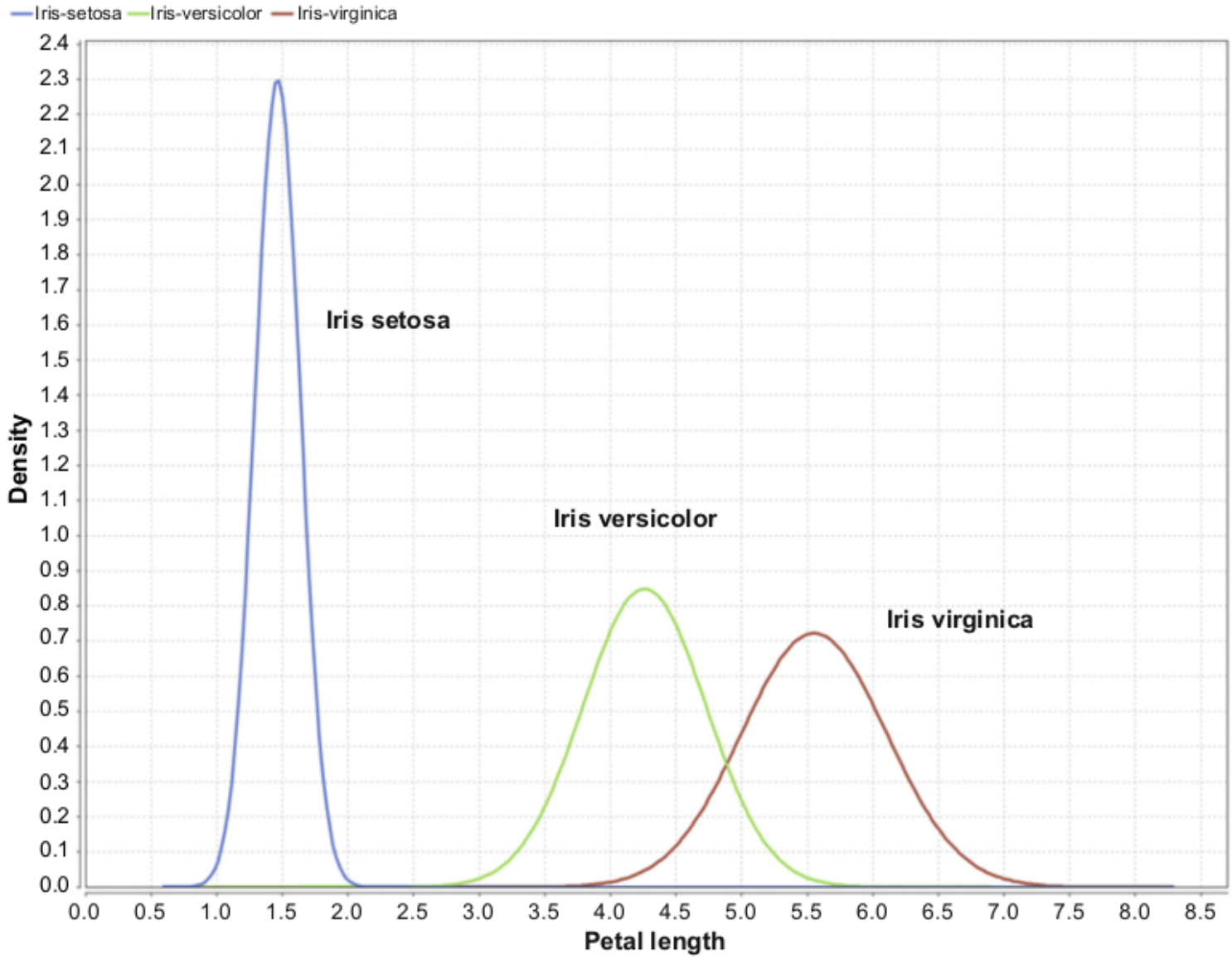


FIGURE 3.9

Distribution of petal length in Iris dataset.

Scatter Plot

- It shows the relationship between 2 attributes in the data set
- Each value in the scatter plot is represented using a dot
- Example - A scatter plot that shows the relationship between petal length and petal width of 3 different species of Iris data set

Scatter Plot

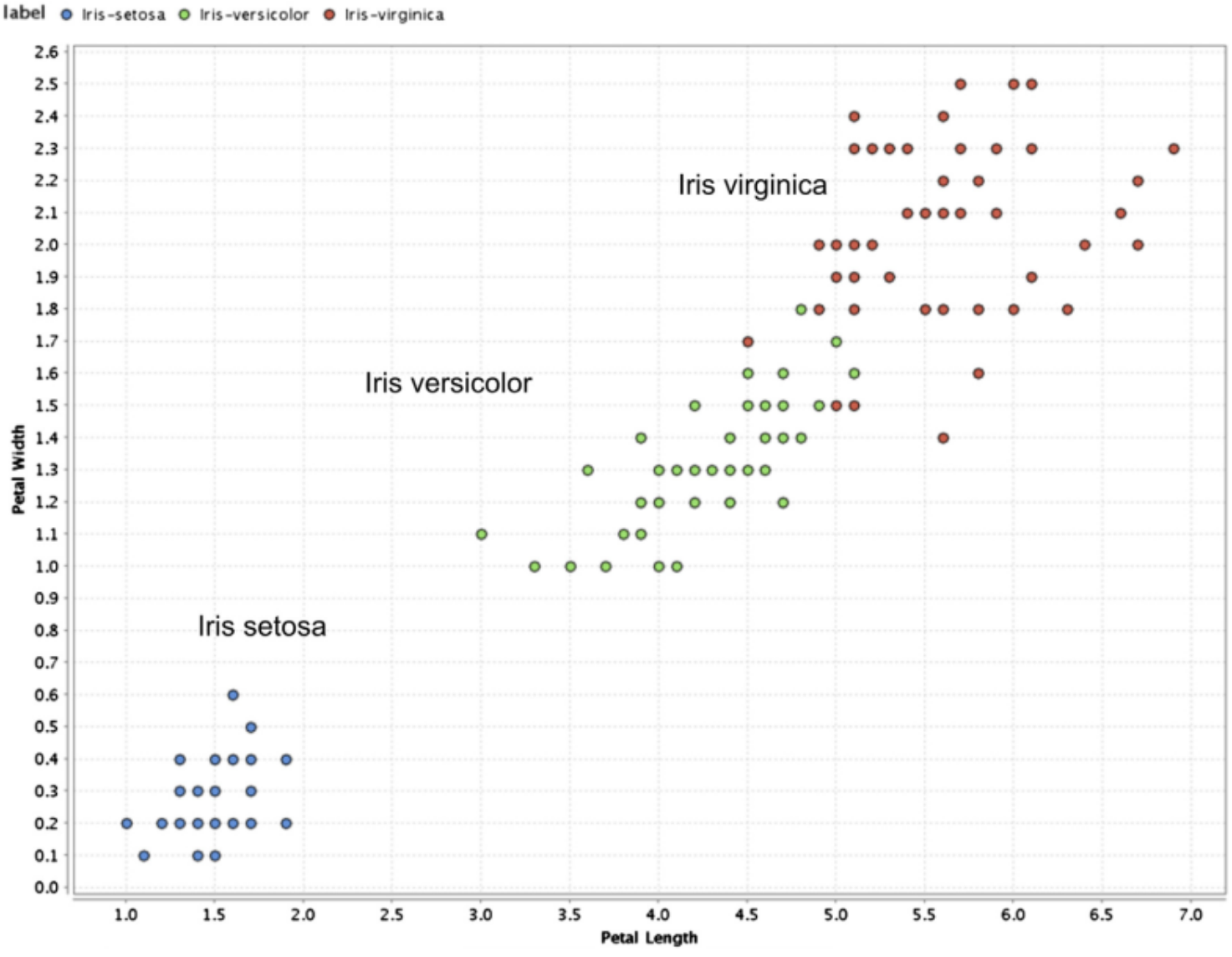


FIGURE 3.10
Scatterplot of Iris dataset.

Bubble Chart

- It also shows the relationship between 2 attributes in the data set
- Each value in the scatter plot is represented using a bubble
- Here, the size of the bubble is determined by a third attribute
- Example - A bubble chart that shows the relationship between petal length and petal width of 3 different species of Iris data set
- Here, the size of the bubble is determined by sepal width

Bubble Chart

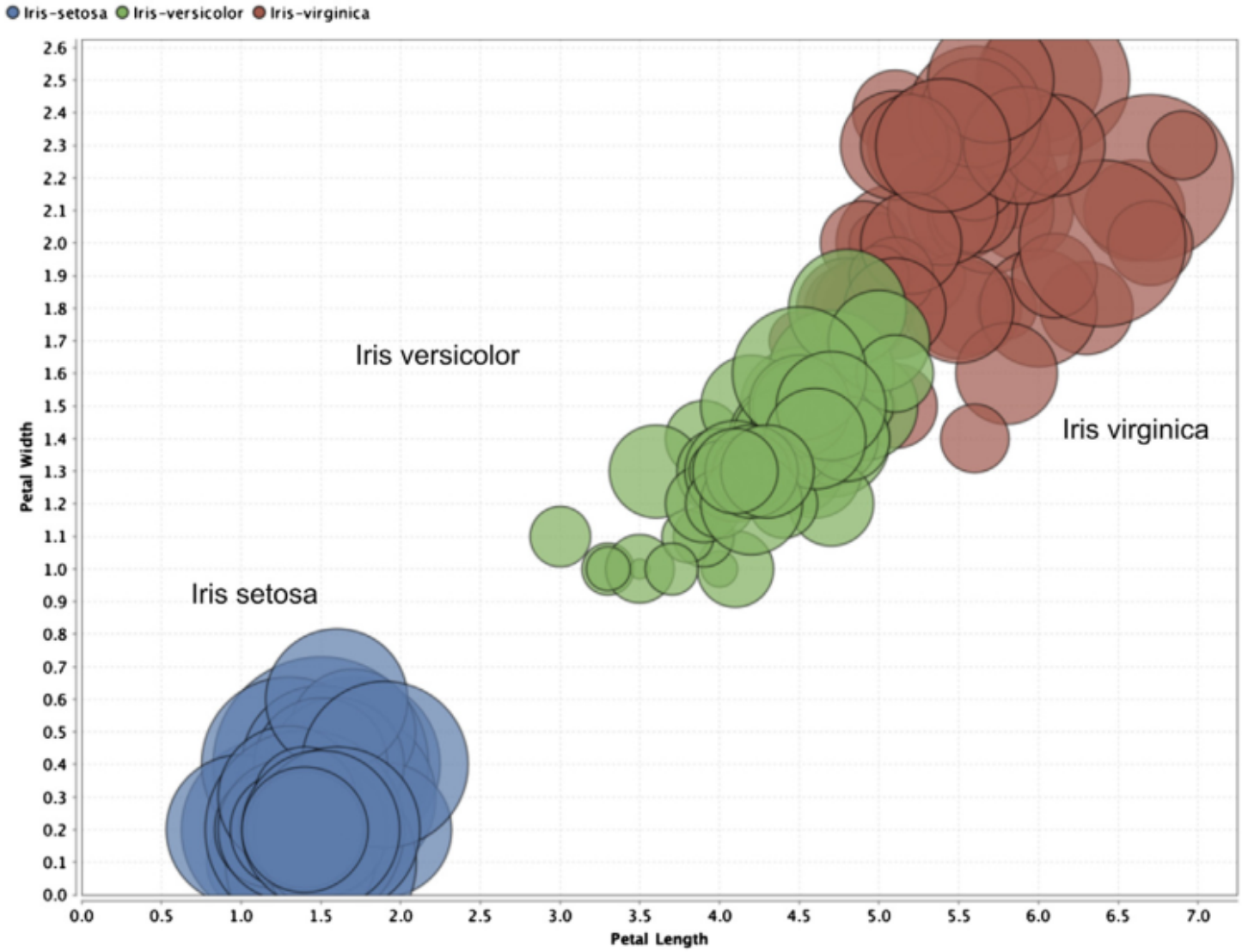


FIGURE 3.13

Bubble chart of Iris dataset.

Density Chart

- It also shows the relationship between 2 attributes in the data set
- Each value in the Density Chart is represented using a dot
- Here, the background colour is determined by a third attribute
- Example - A density chart that shows the relationship between petal length and petal width of 3 different species of Iris data set
- Here, the background colour is determined by sepal width

Density Chart

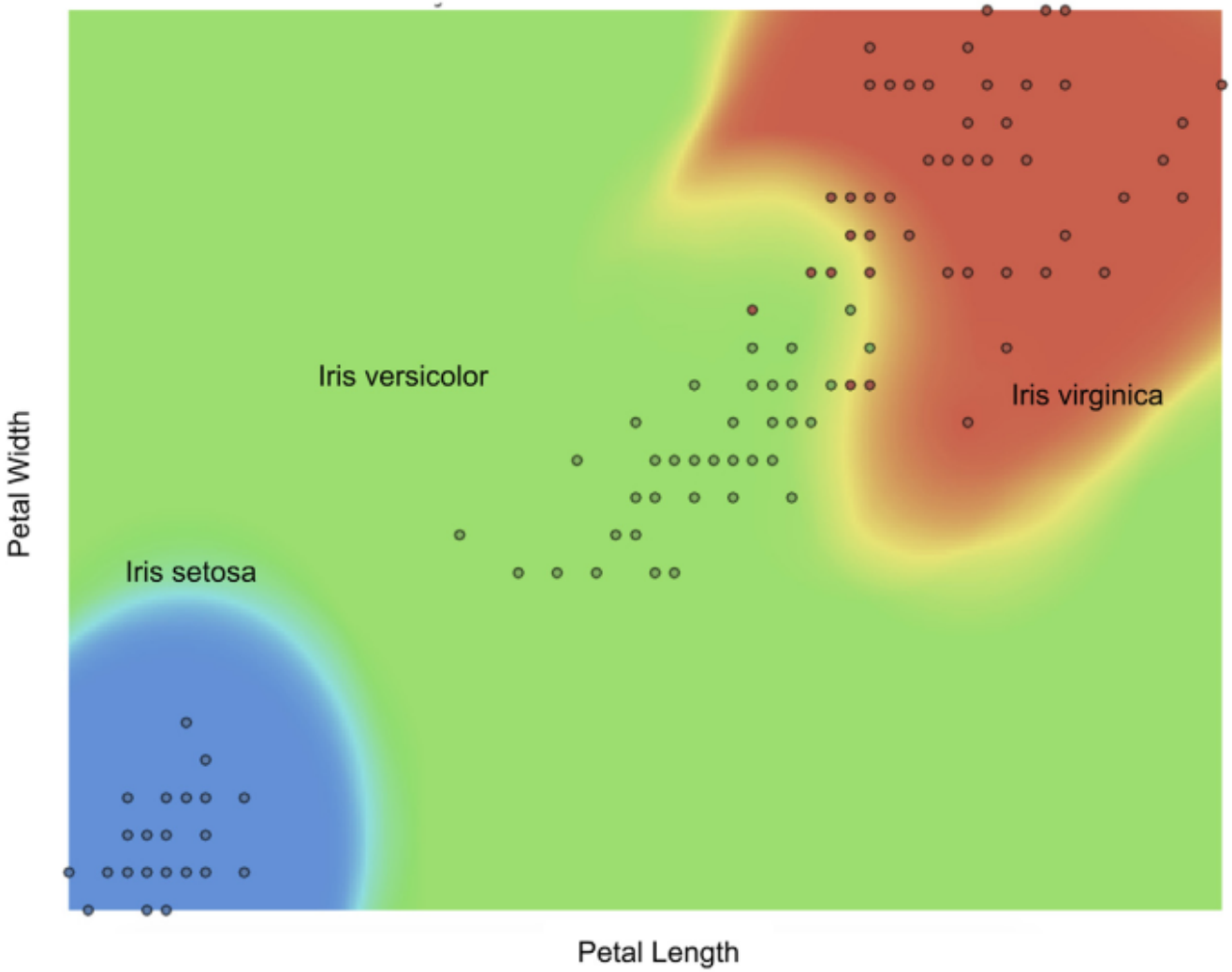
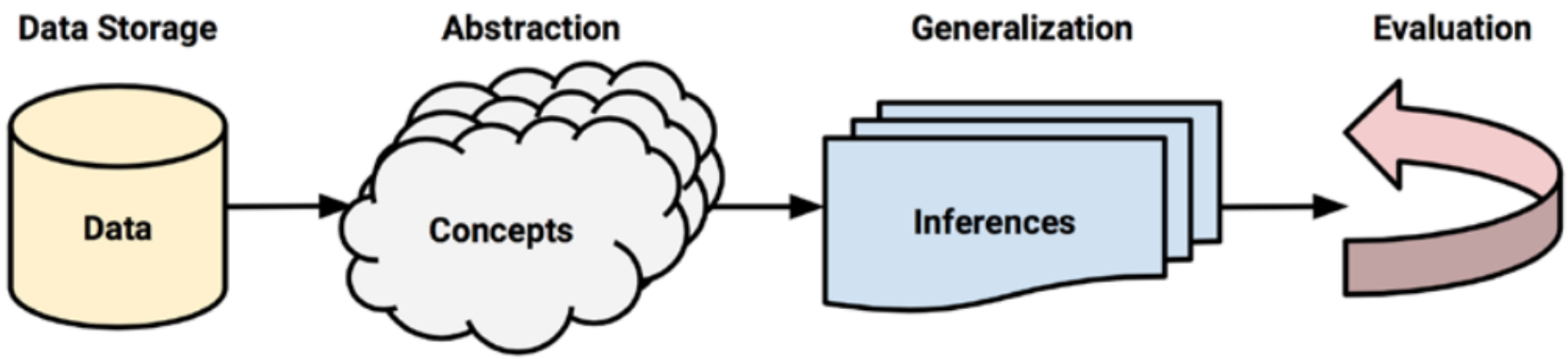


FIGURE 3.14
Density chart of a few attributes in the Iris dataset.

Module 2
Introduction to Machine Learning :

- How machines learn
- A machine is able to learn if it is able to utilise an experience so that its performance improves on similar experiences in future
- Theoretically, machine learning is divided into 4 related components

1. Data Storage
2. Abstraction
3. Generalisation
4. Evaluation



How Machines Learn

1. Data Storage
   - Here we identify the data set used for learning and store it in computers
2. Abstraction
   - In this stage, we assign meaning to stored data
   - In other words, we find out patterns in the data stored
   - Example - From a painting, we understand the object it represents, even though it is not the real object
   - These patterns are represented using a model
   - A model can be a mathematical equation, a tree , a graph or a cluster
   - The process of creating a model from the data set is called training
   - Data transforms to Knowledge after abstraction

How Machines Learn

- Isacc Newton's Model for Gravity of Earth

Observations → **Data** → **Model**

Distance	Time
4.9m	1s
19.6m	2s
44.1m	3s
78.5m	4s

$$g = 9.8m/s^2$$

3. Generalisation

   - We may find out lot of patterns during the abstraction stage
   - In this stage, we limit these patterns to the relevant problem only

4. Evaluation

   - Here we evaluate the model using a test data set to see if it works correctly for new data

Machine Learning in Practice

1. Data Collection
   - Normally, collected data is combined into a single text file, spreadsheet or a database
2. Data Exploration and Preparation
   - Data exploration is the detailed analysis of data to gain better understanding of it
   - For this, statistical analysis and visualisation tools are used
   - Data Preparation is the stage in which we prepare the whole data set needed for the learning process
   - This involves data cleaning, removing duplicate data etc.

Machine Learning in Practice

3. Model Training
   - Here we will be choosing an appropriate machine learning algorithm that will represent data in the form of a model
4. Model Evaluation
   - Here we will evaluate the accuracy of the model using a test data set
5. Model Improvement
   - In this stage, we try to improve the performance of the model using certain advanced strategies
   - Sometimes, we switch to a different type of model, if really needed
   - If the model is performing well, it can be deployed for its intended task

Types of Machine Learning Algorithms

- It can be classified into 3 types

1. Predictive Model
2. Descriptive Model
3. Meta Learning Model

Types of Machine Learning Algorithms

1. Predictive Model

- Here we predict the value of a target variable based on a set of input variables
- The value predicted can be a category or a numeric value
- Examples
- Predicting whether a person has cancer
- Predicting whether a football team will win or lose
- Predicting the date of a baby's conception using the mother's present-day hormone levels
- supervised learning model

Types of Machine Learning Algorithms

2. Descriptive Model

- Here we identify useful patterns / associations within data
- Examples
- Identify the items which are bought together from a shop
- Recommend items to users based on their shopping behaviour
- Detect fraudulent credit card transactions
- Identify hot spots for criminal activities
- unsupervised learning model

3. Meta Learning Model

- This model deals with learning how to learn more effectively
- This can be used for improving the performance of machine learning algorithms

Lazy Learning

- This is a learning approach in which classification is done only after preparing the test data
- This approach spends less time on training, and more time on predicting
- This method is also known as instance based learning or rote learning
- Example - k-Nearest Neighbour Algorithm (k-NN Algorithm)

Classification Using k-NN Algorithm

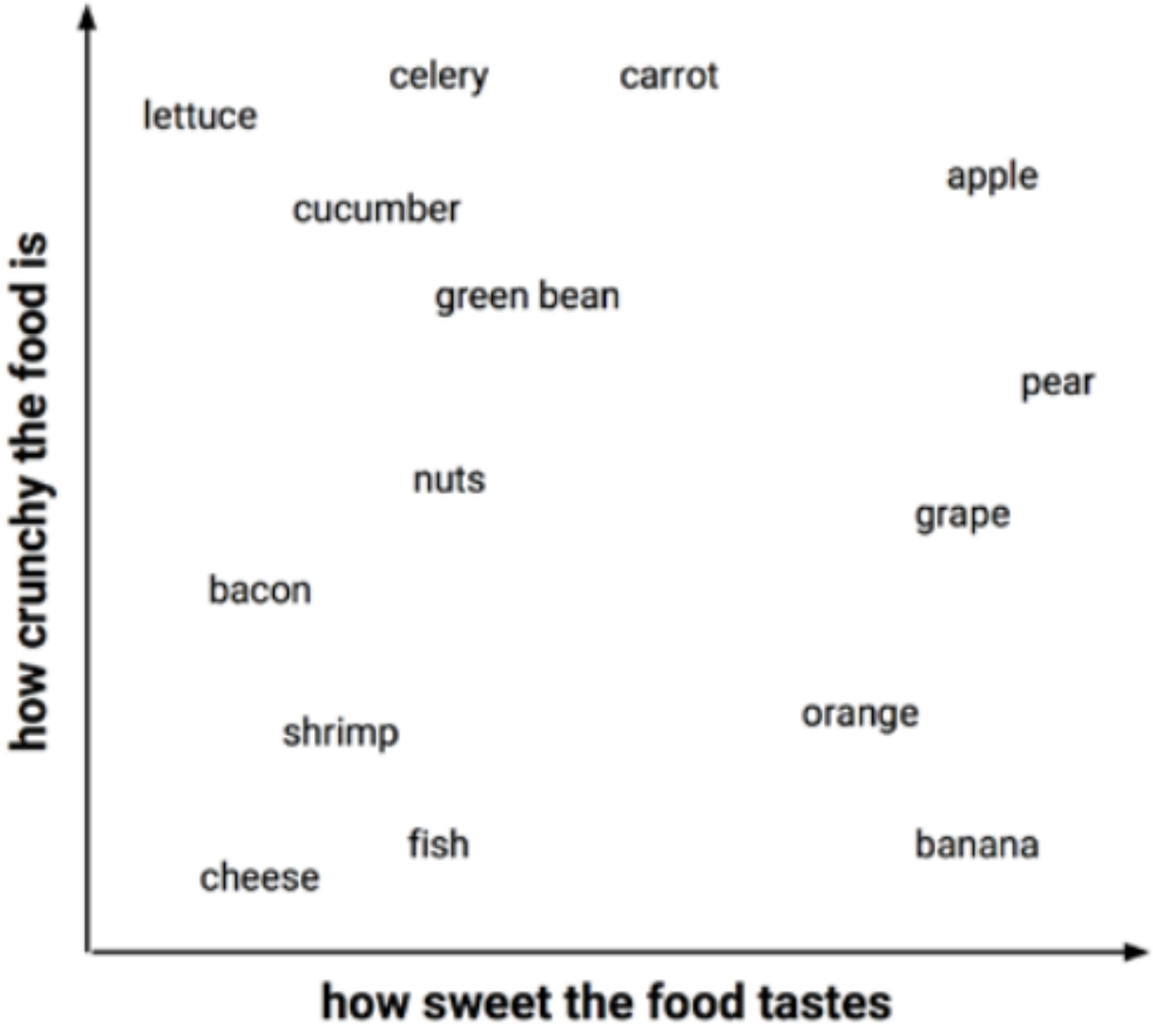
- Example - Food Data Set

Ingredient	Sweetness	Crunchiness	Food type
apple	10	9	fruit
bacon	1	4	protein
banana	10	1	fruit
carrot	7	10	vegetable
celery	3	10	vegetable
cheese	1	1	protein

- Here we consider 2 features of food items - sweetness and crunchiness
- For each food item, they are measured in a scale from 1 to 10

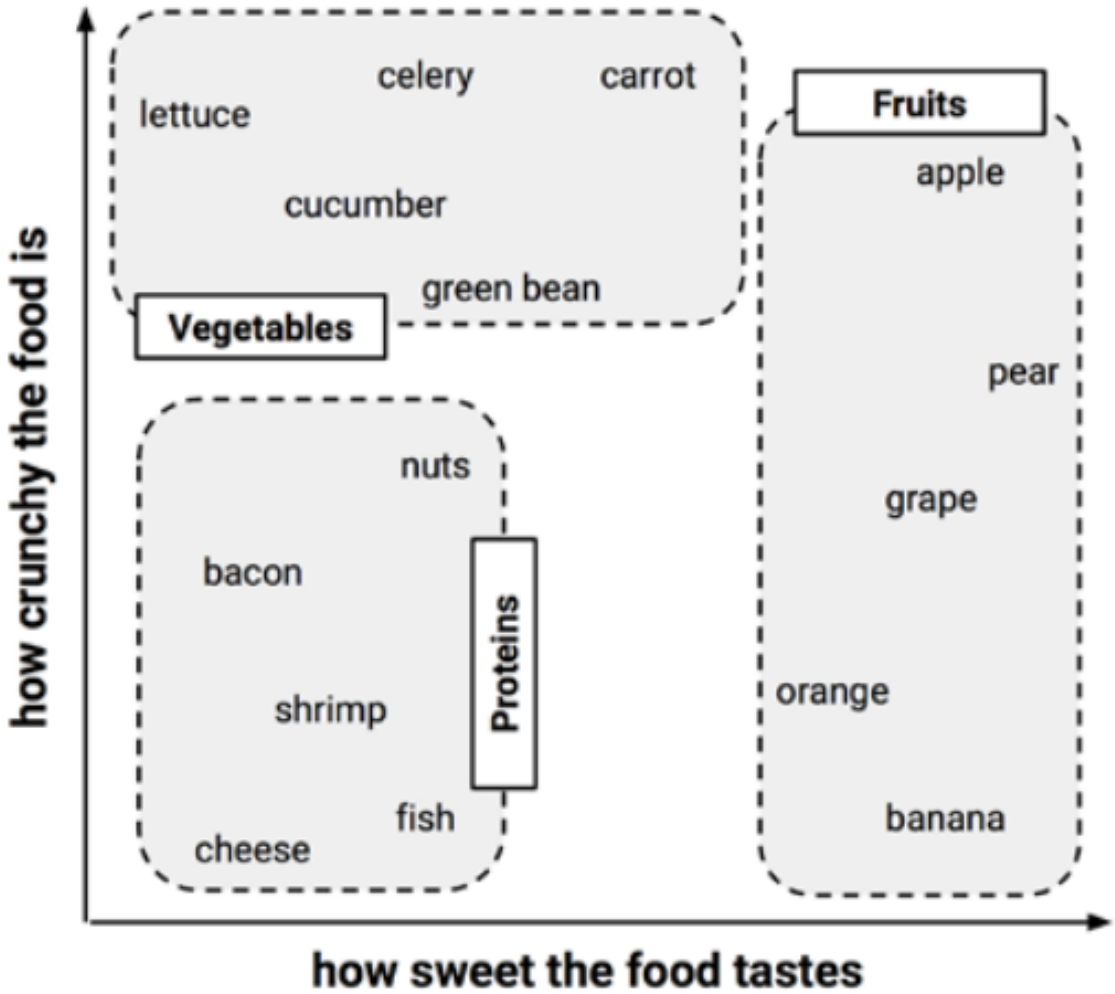
Classification Using k-NN Algorithm

- Scatter Plot of Features of Food Items (More items are added)



Classification Using k-NN Algorithm

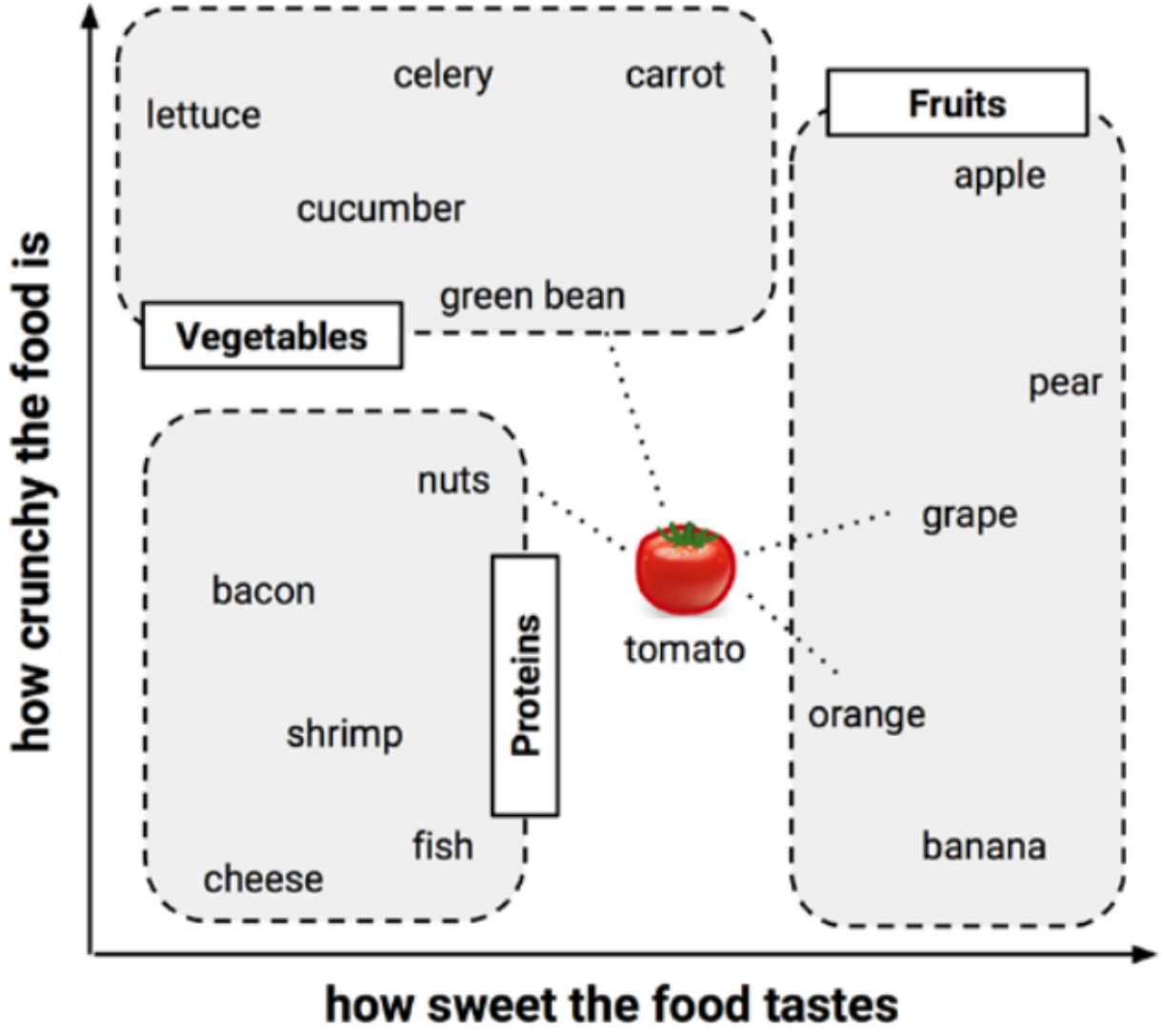
- Scatter Plot of Features of Food Items



- Similar food items are grouped closely together

Classification Using k-NN Algorithm

- Classification of Tomato using k-NN Algorithm



Classification Using k-NN Algorithm

- Classification of Tomato using k-NN Algorithm
- Measuring similarity with distance
- Here we identify tomato's nearest neighbours using a distance function
- Traditionally, the K-NN algorithm uses Euclidean Distance
- Formula

$$\text{dist}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

- Here p and q are items to be compared, having n features
- p_1 refers to the value of first feature of p
- q_1 refers to the value of first feature of q

Classification Using k-NN Algorithm

- Classification of Tomato using k-NN Algorithm
- distance between the tomato (sweetness = 6, crunchiness = 4), and the green bean (sweetness = 3, crunchiness = 7)

$$\text{dist}(\text{tomato, green bean}) = \sqrt{(6-3)^2 + (4-7)^2} = 4.2$$

- distance between the tomato and its closest neighbors

Ingredient	Sweetness	Crunchiness	Food type	Distance to the tomato
grape	8	5	fruit	*sqrt((6 - 8)^2 + (4 - 5)^2) = 2.2*
green bean	3	7	vegetable	*sqrt((6 - 3)^2 + (4 - 7)^2) = 4.2*
nuts	3	6	protein	*sqrt((6 - 3)^2 + (4 - 6)^2) = 3.6*
orange	7	3	fruit	*sqrt((6 - 7)^2 + (4 - 3)^2) = 1.4*

Classification Using k-NN Algorithm

- Classification of Tomato using k-NN Algorithm
- To classify tomato, we will assign it the food type of its nearest neighbour(s)
- k indicates the number of nearest neighbours we consider
- If $k = 1$, orange is the nearest neighbour with a distance of 1.4
- Since orange is a fruit, tomato will also be a fruit
- If $k = 3$, orange, grape and nuts are its nearest neighbours
- Since majority of them are fruits, tomato will be a fruit

k-Nearest Neighbour Algorithm

1. Load the data
2. Divide the data into training data and test data
3. Initialise the value of k
4. Calculate the Euclidean distance between the test data and each of the training data
5. Find k number of nearest neighbours having minimum distance values
6. Get the most frequent class of these neighbours
7. This will be the predicted class

Choice of k

- If k is small, the result will be affected by noisy data
- If k is large, the algorithm will be computationally expensive
- The best k value is somewhere between these two extremes
- We can choose k as the square root of the number of training data
- An alternative approach is to test several k values on a variety of test data sets and choose the one that delivers the best performance

Preparing Data for Use With k-NN

- To get accurate result, ranges of features need to be standardised (Uniform Scaling of Ranges)

1. min-max normalisation

$$\mathrm{X}_{new} = \frac{\mathrm{X} - \min(\mathrm{X})}{\max(\mathrm{X}) - \min(\mathrm{X})}$$

2. z-score standardisation

$$\mathrm{X}_{new} = \frac{\mathrm{X} - \mu}{\sigma} = \frac{\mathrm{X} - \mathrm{Mean}(\mathrm{X})}{\mathrm{StdDev}(\mathrm{X})}$$

- For calculating Euclidean distance, all values need to be converted to numeric
- Example - male = 0, female = 1

Probabilistic Learning: Understanding Naive Bayes

- In probabilistic learning, the probabilities associated with various events are used for learning and prediction
- The probability is a number between 0 and 1, which indicates the chance that an event will occur
- A probability of 0 indicates that there is no chance for the event to occur
- A probability of 1 indicates that there is 100 percent chance for the event to occur
- This learning model is used in Naive Bayes Algorithm
- The technique used in this algorithm is based on the work by 18th Century Mathematician, Thomas Bayes

Conditional Probability and Bayes Theorem

- Naive Bayes algorithm is based on Bayes Theorem
- Bayes theorem gives the conditional probability of an event A given another event B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- P(A|B) = Conditional probability of A given B
- P(B|A) = Conditional probability of B given A
- P(A) = Probability of event A
- P(B) = Probability of event B

Conditional Probability and Bayes Theorem

- Example - Tossing 2 Coins
- Sample Space = {HH, HT, TH, TT}
- H: Head, T: Tail

```
P(second coin being head given first coin is tail) = P(A|B)
     = [P(B|A) * P(A)] / P(B)
     = [P(First coin is tail given second coin is head) *
        P(Second coin being Head)] / P(first coin being
        tail)
     = [(1/2) * (1/2)] / (1/2)
     = (1/2)
     = 0.5
```

Naive Bayes Algorithm for Classification

- Example - Given the following data on a certain set of patients seen by a doctor. Can the doctor conclude that a person having chills, fever, mild headache and without running nose has flu? (Use Naive Bayes classification).

Chills	Running nose	Headache	Fever	Has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

Naive Bayes Algorithm for Classification

- Here we calculate the conditional probability of Flu = Y using Bayes Theorem
- Let W1=Chills, W2=Running Nose, W3=Headache, W4=Fever
- Now the conditional probability of Flu=Y can be defined as
- P(Flu=Y | W1=Y $\bigcap$ W2=N $\bigcap$ W3=mild $\bigcap$ W4=Y)
  $= \frac{P(W1=Y \bigcap W2=N \bigcap W3=mild \bigcap W4=Y|Flu=Y)P(Flu=Y)}{P(W1=Y \bigcap W2=N \bigcap W3=mild \bigcap W4=Y)}$
  $\propto$ P(W1=Y | Flu=Y)P(W2=N | Flu=Y)P(W3=mild | Flu=Y)P(W4=Y | Flu=Y)P(Flu=Y)
  $\propto$ (3/5)*(1/5)*(2/5)*(4/5)*(5/8)
  $\propto$ 0.024
- Naive Bayes algorithm assumes independence among events of the same class
- For independent events, $P(A \bigcap B) = P(A)P(B)$

Naive Bayes Algorithm for Classification

- Here we calculate the conditional probability of Flu = N using Bayes Theorem
- Let W1=Chills, W2=Running Nose, W3=Headache, W4=Fever
- Now the conditional probability of Flu=N can be defined as
- P(Flu=N | W1=Y $\bigcap$ W2=N $\bigcap$ W3=mild $\bigcap$ W4=Y)
  $= \frac{P(W1=Y \bigcap W2=N \bigcap W3=mild \bigcap W4=Y|Flu=N)P(Flu=N)}{P(W1=Y \bigcap W2=N \bigcap W3=mild \bigcap W4=Y)}$
  $\propto$ P(W1=Y | Flu=N)P(W2=N | Flu=N)P(W3=mild | Flu=N)P(W4=Y | Flu=N)P(Flu=N)
  $\propto$ (1/3)*(2/3)*(1/3)*(1/3)*(3/8)
  $\propto$ 0.009
- Since the proportional conditional probability of Flu=Y is greater, our prediction will be that the person will have flu

Naive Bayes Algorithm for Classification

- R program

```
# Loads e1071 package containing naiveBayes
library(e1071)
# Read the csv file into a data frame
symptoms = read.csv("symptoms.csv")
#Training Data and Test Data
symptoms_train = symptoms[,1:4]
symptoms_test = data.frame(Chills="Y",RunningNose="N",
Headache="mild",Fever="Y")
#Naive Bayes Classification
classifier_cl <- naiveBayes(symptoms_train,symptoms$HasFlu)
# This will print classical and conditional probabilities
classifier_cl
# Predicting on test data
symptoms_test_pred <- predict(classifier_cl, symptoms_test)
cat("Prediction of Flu:\n")
symptoms_test_pred
```

Naive Bayes Algorithm for Classification

- Output

```
Naive Bayes Classifier for Discrete Predictors
Call:
naiveBayes.default(x = symptoms_train, y = symptoms$HasFlu)
A-priori probabilities:
symptoms$HasFlu
    N     Y
0.375 0.625
Conditional probabilities:
                Chills
symptoms$HasFlu         N         Y
              N 0.6666667 0.3333333
              Y 0.4000000 0.6000000
                RunningNose
symptoms$HasFlu         N         Y
              N 0.6666667 0.3333333
              Y 0.2000000 0.8000000
```

Naive Bayes Algorithm for Classification

- Output - continued

```
                Headache
symptoms$HasFlu      mild        no    strong
              N 0.3333333 0.3333333 0.3333333
              Y 0.4000000 0.2000000 0.4000000
                Fever
symptoms$HasFlu         N         Y
              N 0.6666667 0.3333333
              Y 0.2000000 0.8000000
Prediction of Flu:
[1] N
Levels: N Y
```

Naive Bayes Algorithm for Classification

1. Load the Data
2. Divide the data into training data and test data
3. Calculate the classical and conditional probabilities of each event in the training data
4. Assume independence among various events of the same class
5. Use mean, standard deviation and normal distribution formulas to calculate the conditional probabilities of events with numeric values
6. Calculate the proportional conditional probabilities of each class for the test data
7. The predicted class will be the one with greater proportional conditional probability

Module 3
Decision Tree Learning:
Concept of Decision Tree

- Decision tree is a powerful machine learning tool for classification and prediction
- It is a flowchart like tree structure
- Here every non leaf node indicate a test on an attribute
- Each branch represent an outcome on the test
- Every leaf node indicate a class label

Concept of Decision Tree

Outlook	Temp	Play?
Sunny	30	Yes
Overcast	15	No
Sunny	16	Yes
Cloudy	27	Yes
Overcast	25	Yes
Overcast	17	No
Cloudy	17	No
Cloudy	35	Yes

Classification Problem

Weather -> Play (Yes, No)

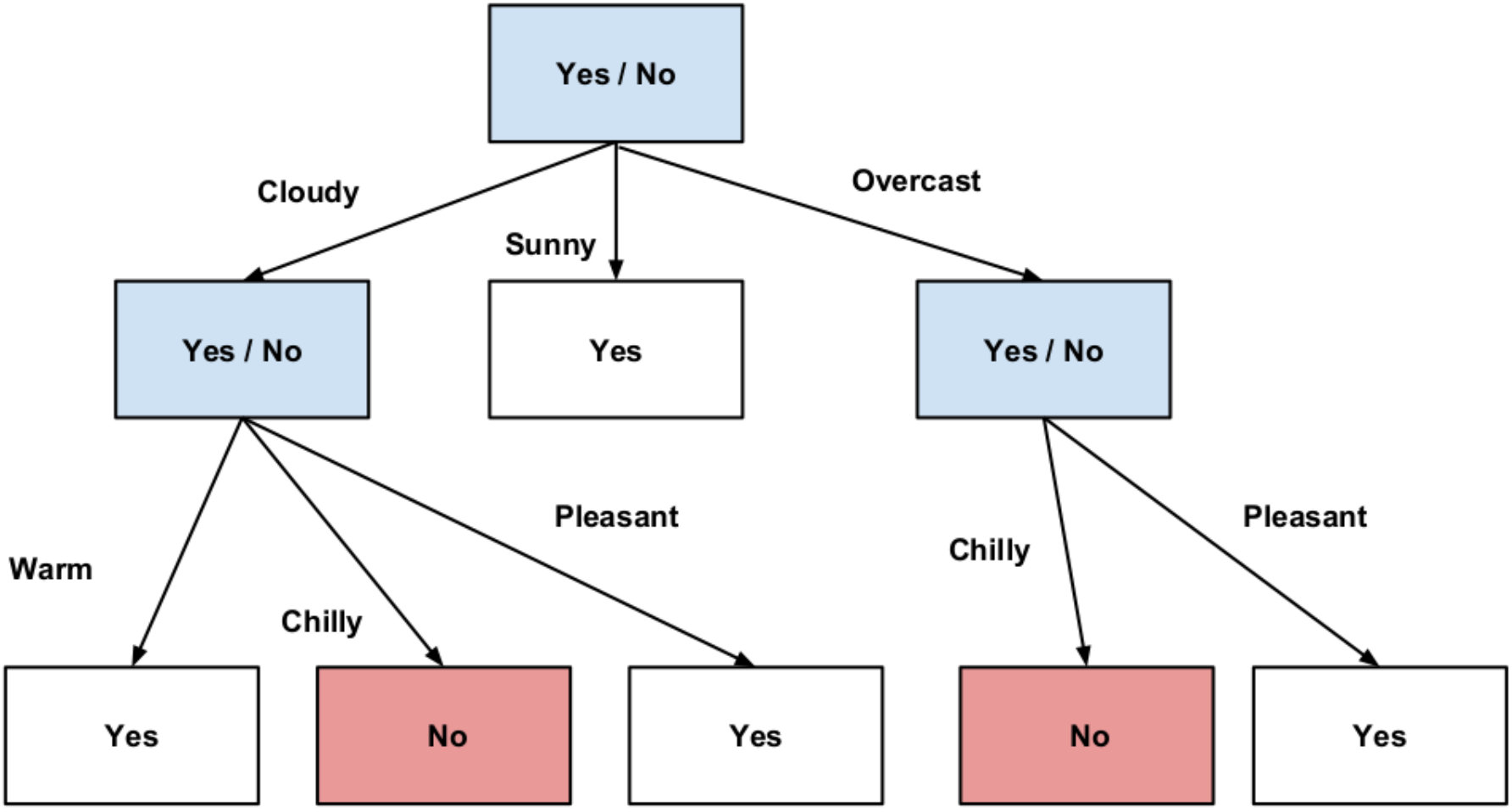
Concept of Decision Tree

- Here we convert temperature values from numeric to categorical
- Chilly: < 20
- Pleasant: 20 - 30
- Warm: > 30

Outlook	Temp	Play?
Sunny	Warm	Yes
Overcast	Chilly	No
Sunny	Chilly	Yes
Cloudy	Pleasant	Yes
Overcast	Pleasant	Yes
Overcast	Chilly	No
Cloudy	Chilly	No
Cloudy	Warm	Yes

Concept of Decision Tree

- Decision Tree for Classification



Divide and Conquer Approach

- For classification, decision trees use divide and conquer approach
- Initially the whole data set is divided into several subsets
- These subsets are again divided into even smaller subsets and so on
- This process is continued until we arrive at the solution
- At first, the root node represent the entire data set
- Next the decision tree algorithm chooses a feature or attribute to split upon
- Based on the distinct values of this feature, the data set is partitioned into groups, and the first set of tree branches are formed
- Next we choose another feature to split and this process is continued till we arrive at the final solution

C5.0 Decision Tree Algorithm

- We can implement decision trees in different ways
- It is an efficient implementation developed by computer scientist J. Ross Quinlan
- Choosing the best split
- When there are several features in the problem, we have to decide which feature is to be split initially and subsequently after each iteration
- This is done based on 2 metrics called entropy and information gain
- Entropy is a measure of disorder among a set of class values

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \; log_2(p_i)$$

C5.0 Decision Tree Algorithm

- Here S is the data segment
- c refers to the number of class levels
- p_i refers to the proportion of values falling into class level i
- suppose we have a partition of data with two classes: red (60 percent) and white (40 percent)
- We can calculate the entropy as follows
- Entropy(S) = -0.60 * log2(0.60) - 0.40 * log2(0.40) = 0.9709506
- information gain = entropy(parent) - entropy(children)
- Here entropy(children) is the average entropy of child nodes
- We have to make splitting in such a way that there is more information gain
- Pruning the decision trees
- Here the size of the decision tree is reduced by cutting of the unwanted branches

Classification Rules Learning: Concept of Classification Rules

- Classification Rules Learning is a machine learning approach based on classification rules
- A classification rule represent knowledge in the form of logical if..else statements
- This knowledge can be used for classification and prediction
- General Form
- If condition Then conclusion
- Example
- If Weather = Sunny Then GamePlay = Yes
- If part of the classification rule is called antecedent
- Then part of the classification rule is called consequent

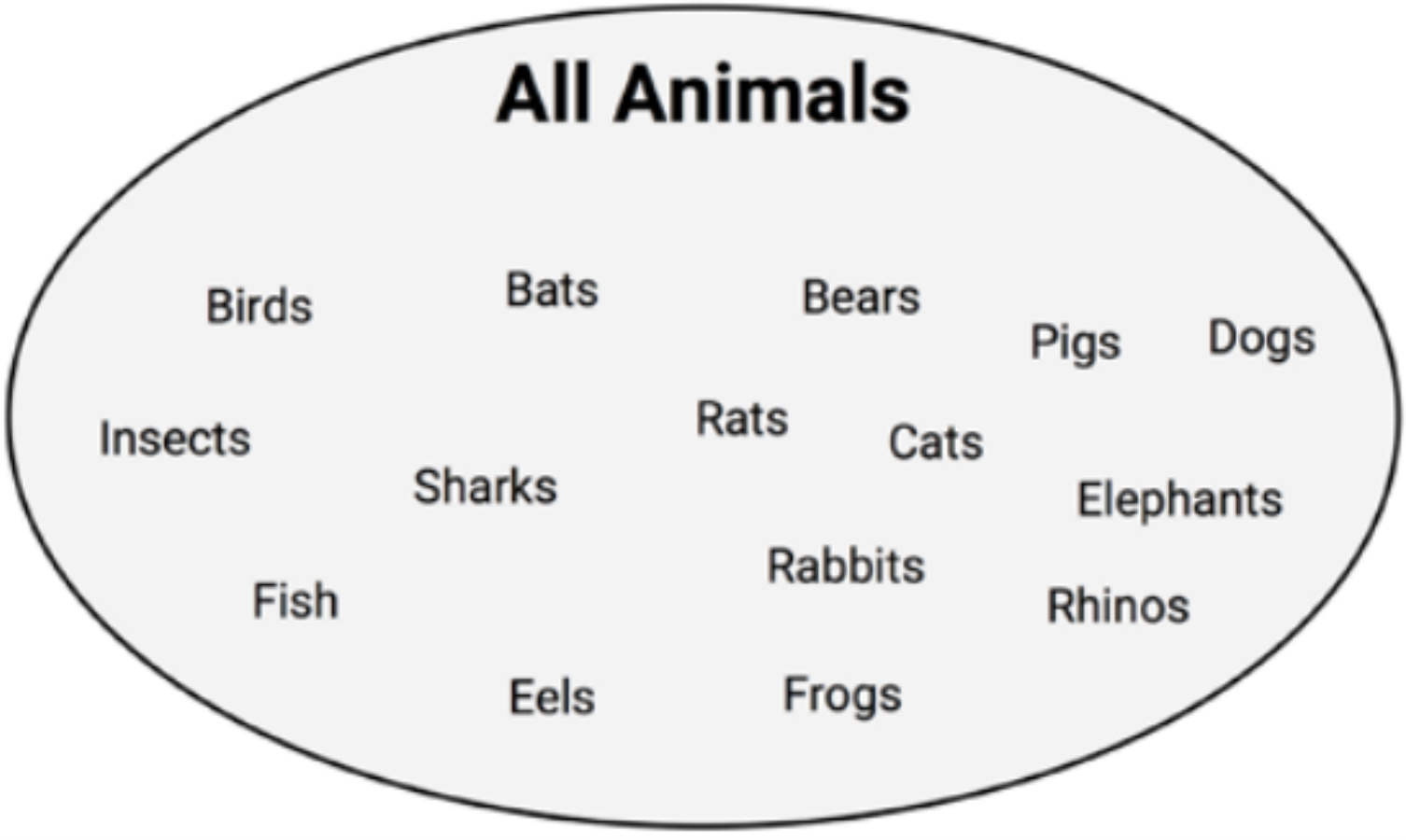
Separate and Conquer Approach

- Machine learning algorithms based on classification rules use the separate and conquer approach for solving problems

1. Identify a rule that covers a subset of the training data
2. Separate this subset from the training data
3. Try to identify more rules that separates more subsets
4. This process is continued till the whole training data is divided into several subsets

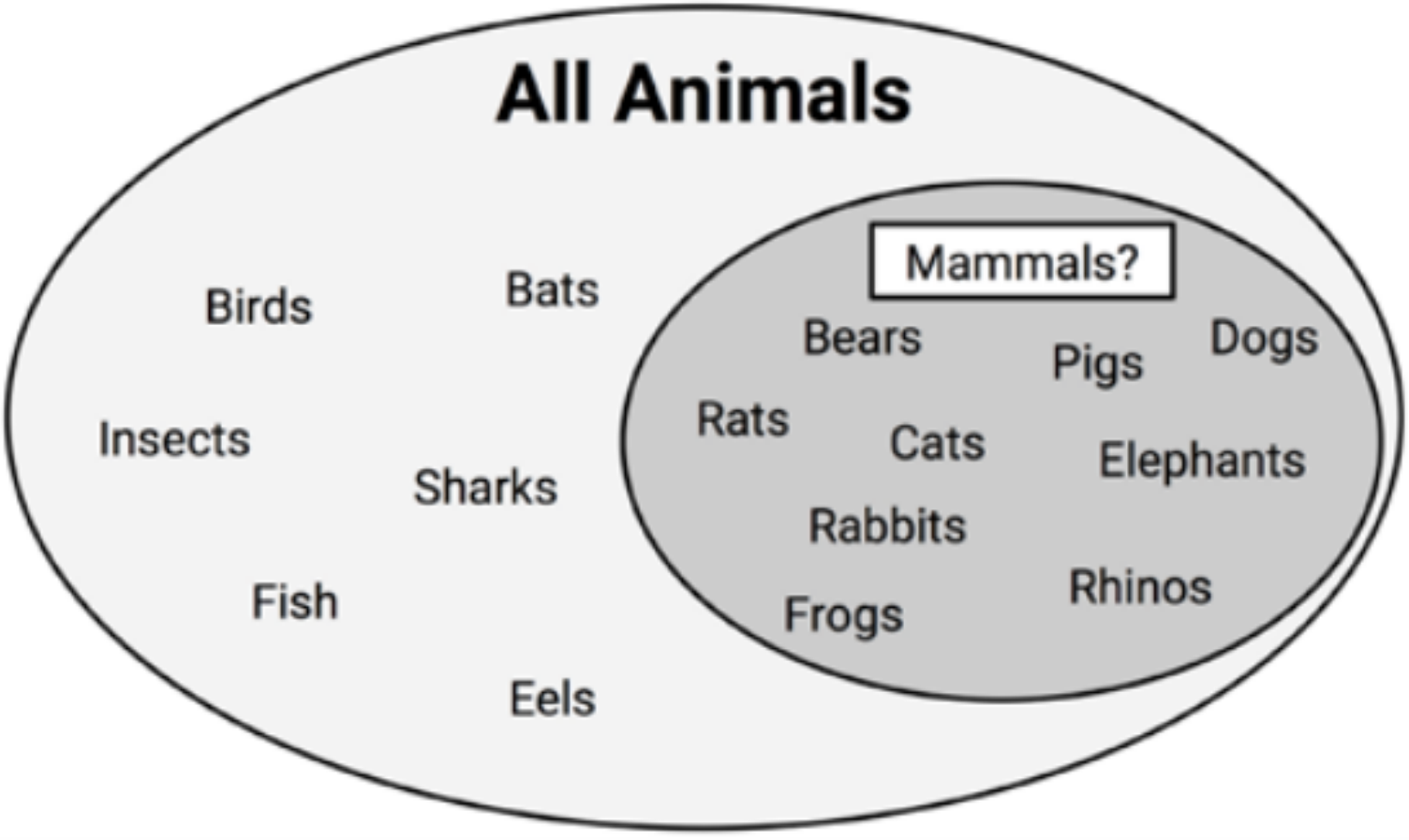
Separate and Conquer Approach

- Example - Create rules to identify whether or not an animal is a mammal



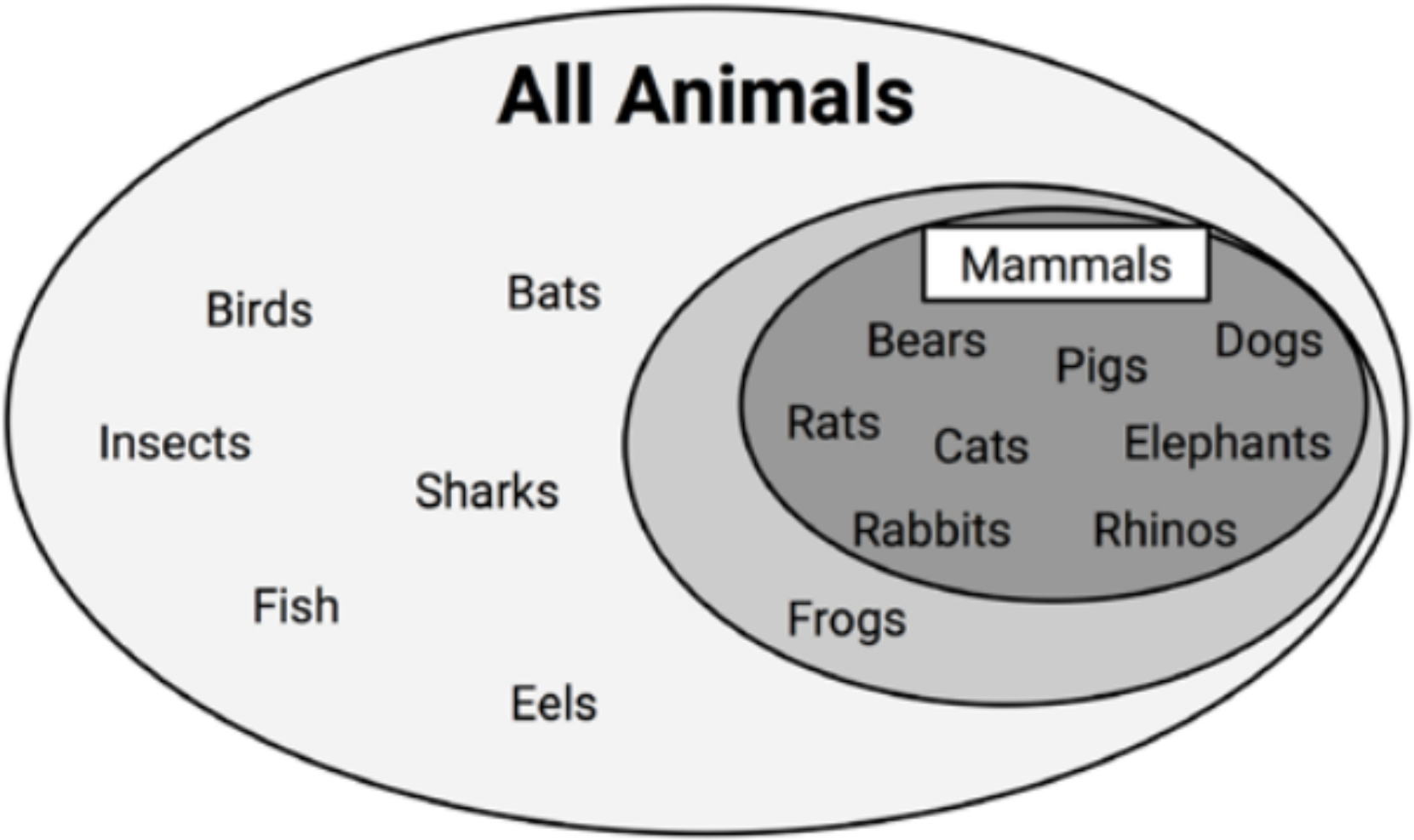
Separate and Conquer Approach

- Rule 1 - All animals that walk on land are mammals



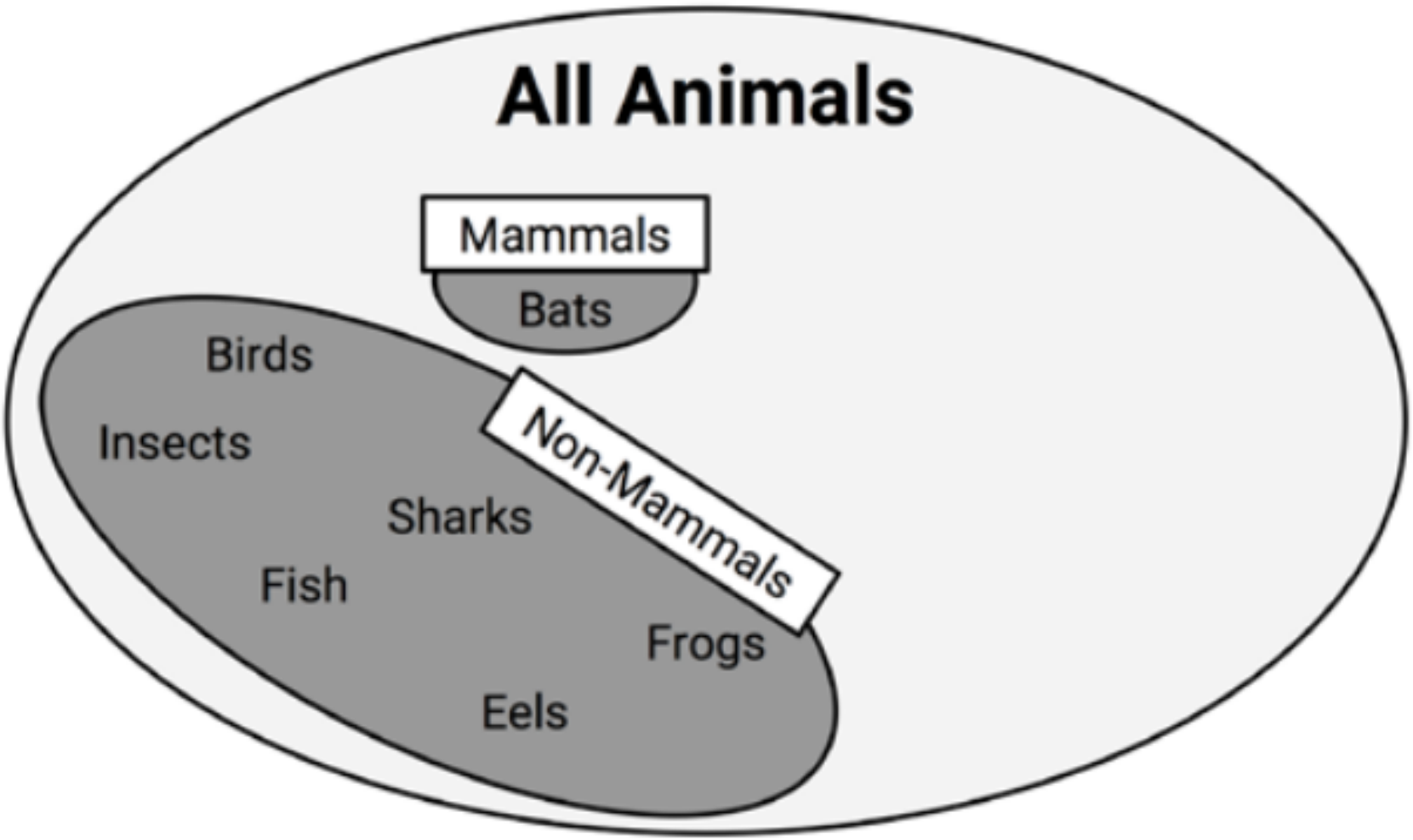
Separate and Conquer Approach

- Frogs are amphibians not mammals
- Rule 1 (modified) - All animals that walk on land and have tails are mammals



Separate and Conquer Approach

- Bats are mammals
- Rule 2 - If the animal does not have fur, it is not a mammal
- Rule 3 - Otherwise, the animal is a mammal



The 1R Algorithm

- It is a simple classification rules machine learning algorithm
- It is also called One Rule or OneR Algorithm

1. For each feature, divide the data into groups based on similar values of the feature
2. For each group, predict the class, based on the assumed rule
3. The error rate for the rule based on each feature is calculated
4. The rule with the fewest errors is chosen as the one rule.

The 1R Algorithm

- Working on Animal Data

Animal	Travels By	Has Fur	Mammal
Bats	Air	Yes	Yes
Bears	Land	Yes	Yes
Birds	Air	No	No
Cats	Land	Yes	Yes
Dogs	Land	Yes	Yes
Eels	Sea	No	No
Elephants	Land	No	Yes
Fish	Sea	No	No
Frogs	Land	No	No
Insects	Air	No	No
Pigs	Land	No	Yes
Rabbits	Land	Yes	Yes
Rats	Land	Yes	Yes
Rhinos	Land	No	Yes
Sharks	Sea	No	No

Full Dataset

Travels By	Predicted	Mammal	
Air	No	Yes	✖
Air	No	No	
Air	No	No	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	No	✖
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Land	Yes	Yes	
Sea	No	No	
Sea	No	No	
Sea	No	No	

Rule for "Travels By"
Error Rate = 2 / 15

Has Fur	Predicted	Mammal	
No	No	No	
No	No	No	
No	No	Yes	✖
No	No	No	
No	No	No	
No	No	No	
No	No	Yes	✖
No	No	Yes	✖
No	No	No	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	
Yes	Yes	Yes	

Rule for "Has Fur"
Error Rate = 3 / 15

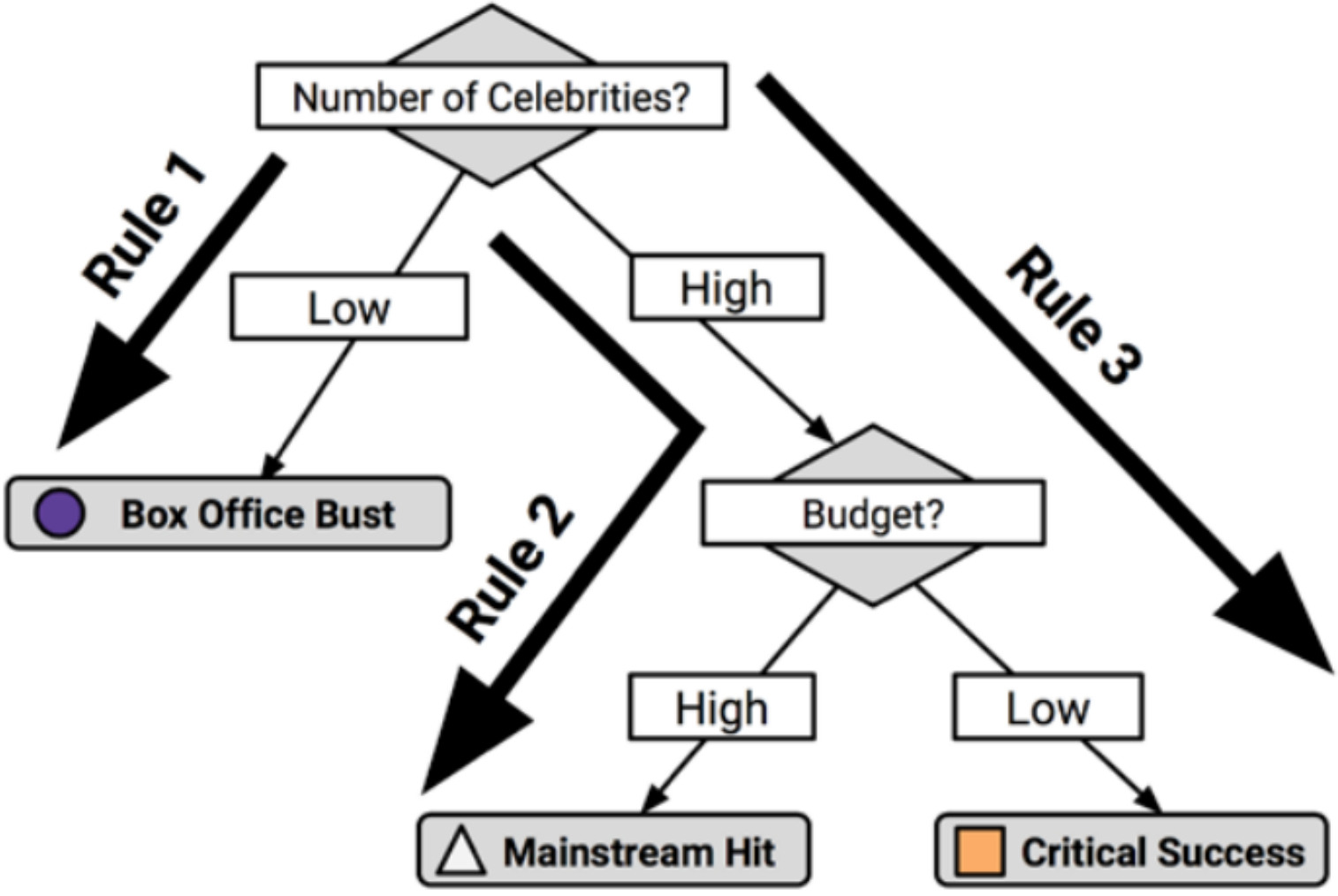
The 1R Algorithm

- For the 'Travels By' feature, animals who travel by land were predicted to be mammals
- For the 'Travels By' feature, bats and frogs were the wrong predictions
- For the 'Has Fur' feature, animals having fur were predicted to be mammals
- For the 'Has Fur' feature, pigs, elephants and rhinos were the wrong predictions
- 'Travels By' feature has fewer errors, hence the following rules are returned

1. If the animal travels by air, it is not a mammal
2. If the animal travels by land, it is a mammal
3. If the animal travels by sea, it is not a mammal

Rules from Decision Trees

- Classification rules can also be obtained directly from decision trees



Rules from Decision Trees

- We can deduce the following rules by following the paths from root node to the leaf nodes in the above tree

1. If the number of celebrities is low, then the movie will be a Box Office Bust
2. If the number of celebrities is high and the budget is high, then the movie will be a Mainstream Hit
3. If the number of celebrities is high and the budget is low, then the movie will be a Critical Success

Regression Methods:
Concept of Regression

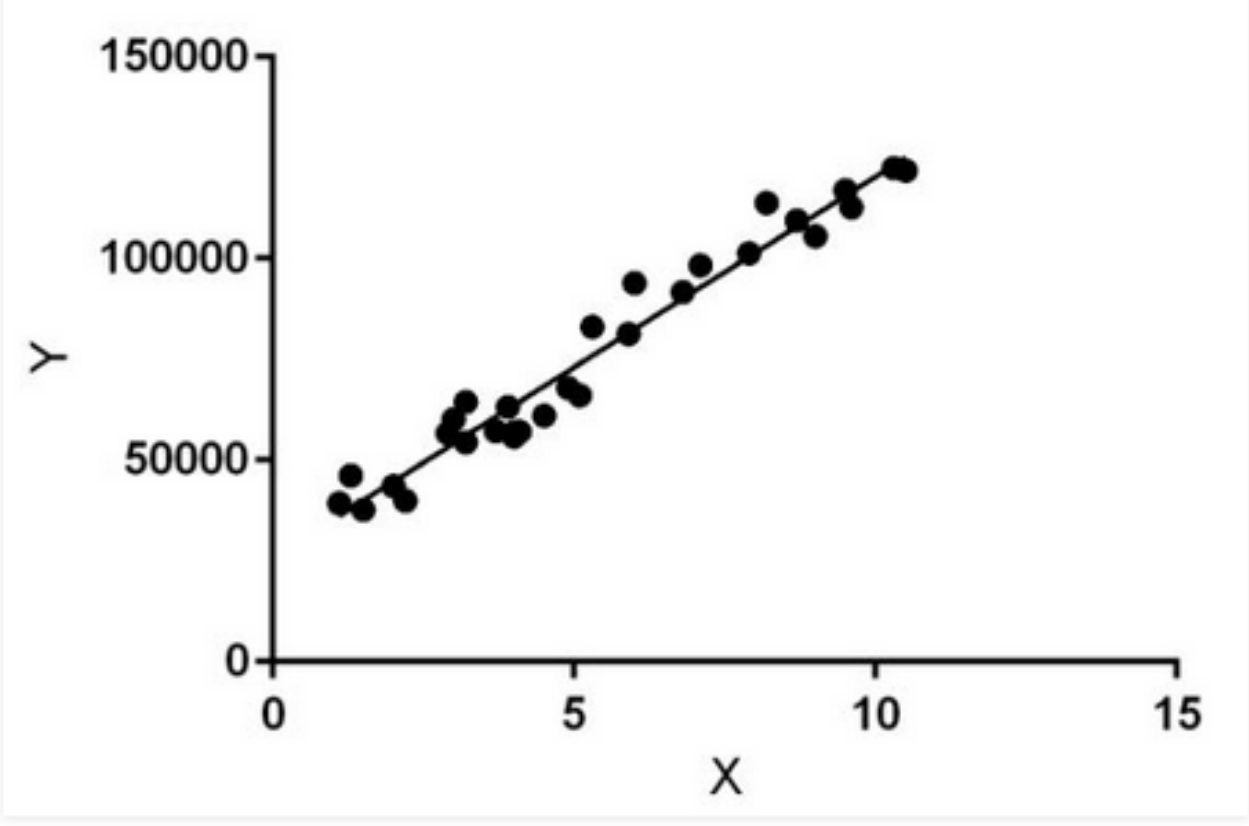
- Regression is a statistical technique used for estimating relationships between numerical data
- Regression specifies the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors)
- The dependent variable depends on the values of independent variable(s)

Simple Linear Regression

- Simple Linear Regression finds out the linear relationship between single numeric dependent variable and a single numeric independent variable
- Equation
- $y = a + bx$
- Here y is the dependent variable
- x is the independent variable
- a and b are constants

Simple Linear Regression

- Example
- Here X is work experience and Y is salary



Ordinary Least Squares Estimation

- Ordinary Least Squares (OLS) is an estimation method that minimise the vertical distance between the predicted y value and the actual y value
- In mathematical terms, it is the task of minimising the following equation

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

- e is the error, y_i is the actual y value, $\hat{y}_i$ is the predicted y value
- The error values are squared and summed across all the points in the data

Correlation

- Using correlation, we can quickly gauge relationship between dependent and independent variables
- Correlation between two attributes is commonly measured by the Pearson correlation coefficient

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

- It ranges from -1 to 1, where negative values indicate negative correlation, positive values indicate positive correlation and 0 indicates no correlation

Multiple Linear Regression

- Multiple Linear Regression finds out the linear relationship between single numeric dependent variable and multiple numeric independent variables
- Equation
- $y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$
- Here y is the dependent variable
- $x_1, x_2, ..., x_n$ are independent variables
- $a, b_1, b_2, ..., b_n$ are constants
- Most real world analyses have more than one independent variable
- multiple linear regression is an extension of simple linear regression

Module 4
Neural Network Learning:

- Neural Network Learning is a machine learning model inspired by the functioning of a biological brain
- A brain consists of a network of interconnected cells called neurons, which helps it to solve problems
- An Artificial Neural Network(ANN) consists of a network of artificial neurons or nodes to solve learning problems

Neural Network Learning:

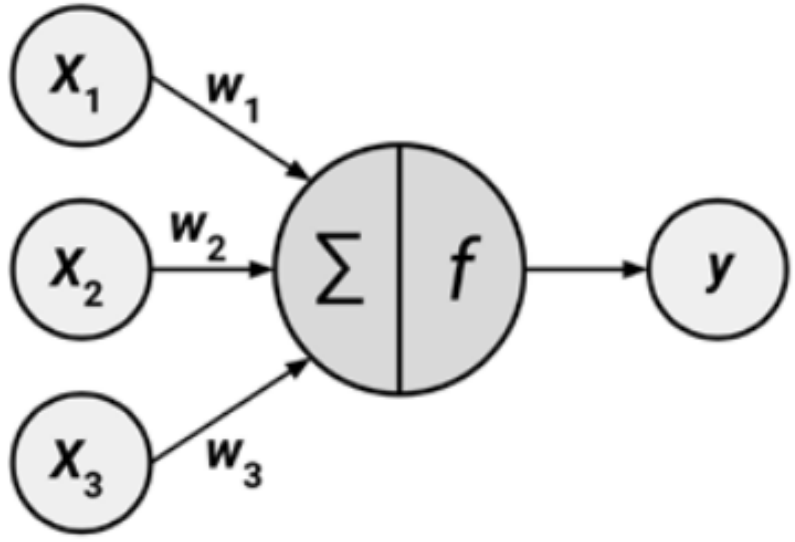
- The human brain is made up of about 85 billion neurons
- A cat has roughly a billion neurons
- A mouse has about 75 million neurons
- A cockroach has only about a million neurons
- A fruit fly brain has 100,000 neurons
- ANNs contain only several hundred neurons
- Still, ANNs can be applied to learning tasks like classification, numeric prediction and unsupervised pattern recognition

Artificial Neurons

- A network topology describes the number of neurons in the model as well as the number of layers and how they are connected
- A typical artificial neuron has n input signals $x_1, x_2, ..., x_n$ with weights $w_1, w_2, ..., w_n$
- The weighted sum of the input signals is used by an activation function to produce the output signal y

Artificial Neurons

- An artificial neuron with 3 input signals

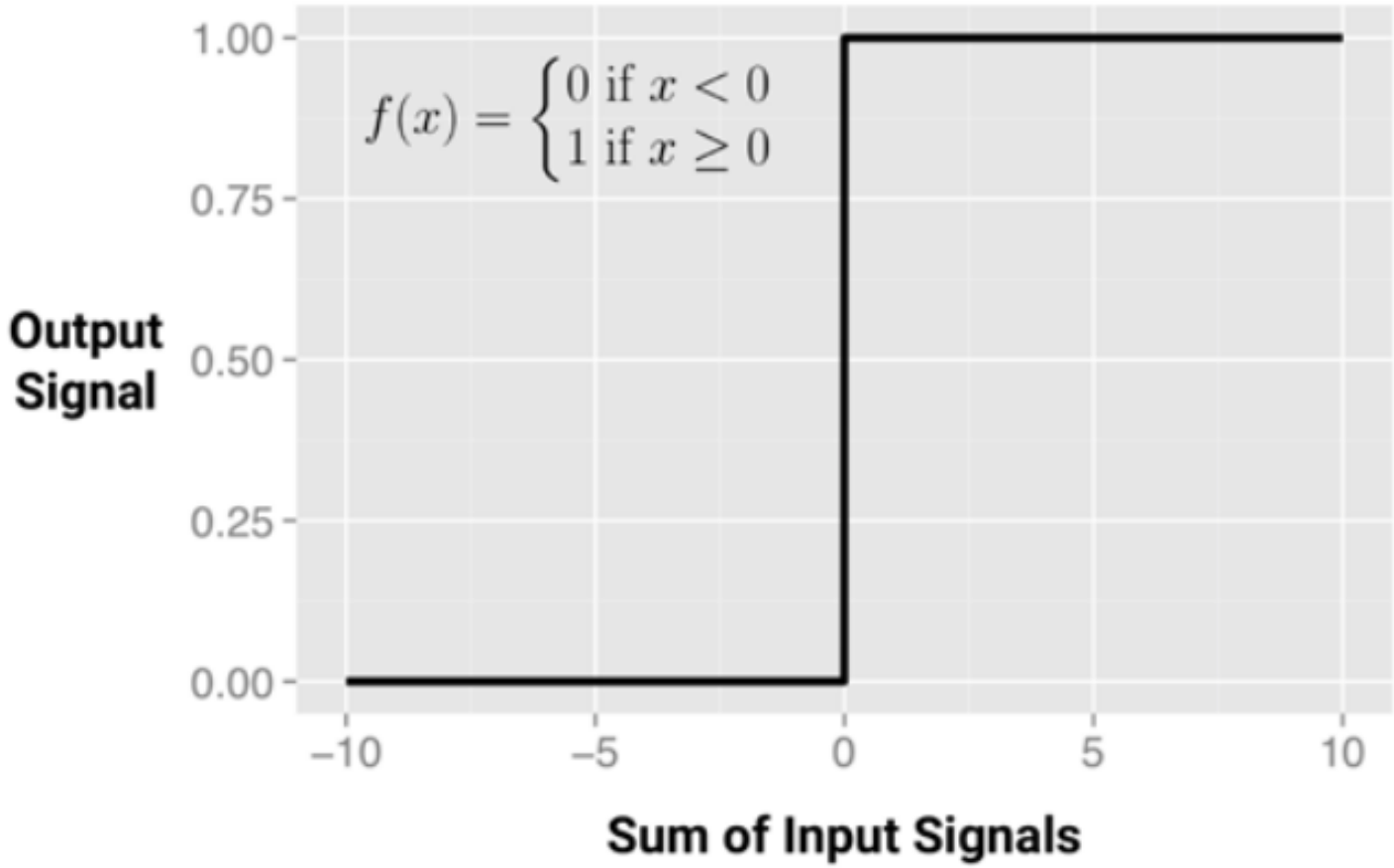


- General formula for the output signal from an artificial neuron

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

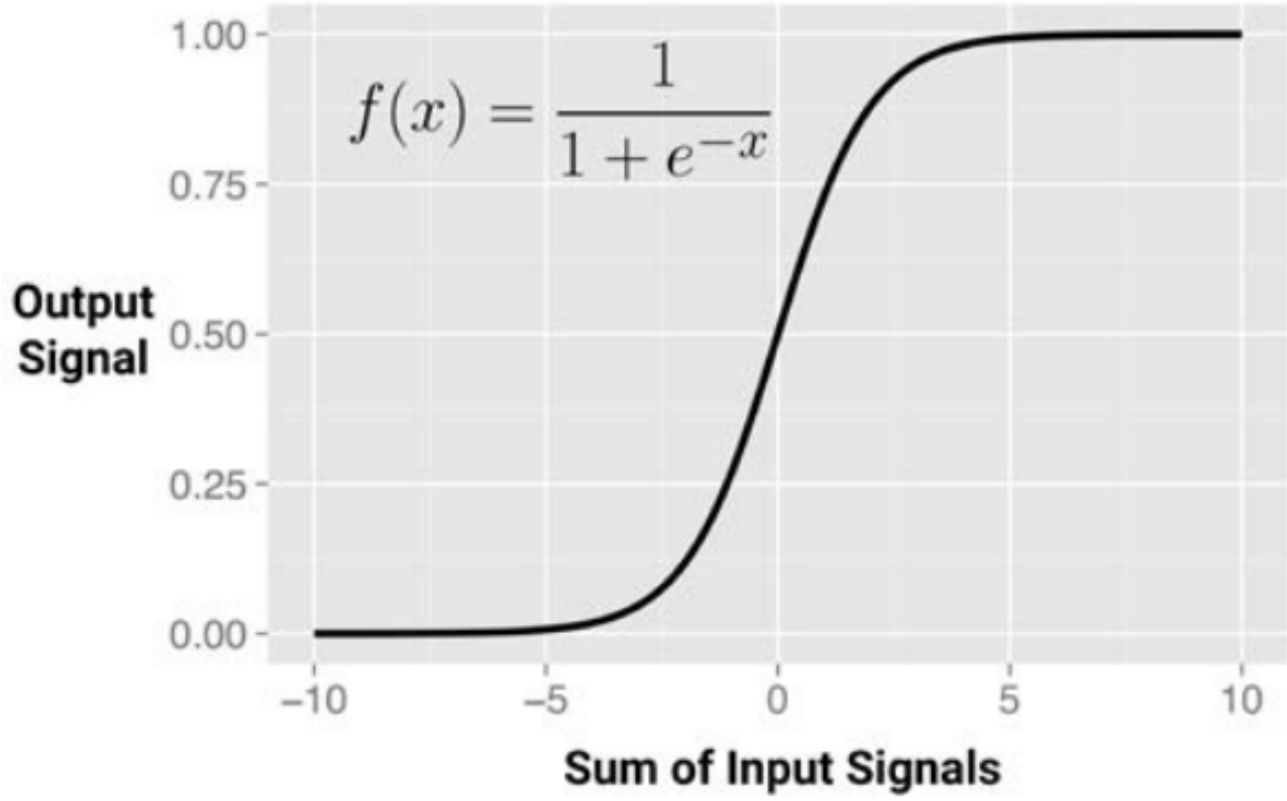
Activation Functions

- An activation function processes the input signals to produce the output signal
- Different neural network algorithms use different activation functions
- Unit Step Activation Function



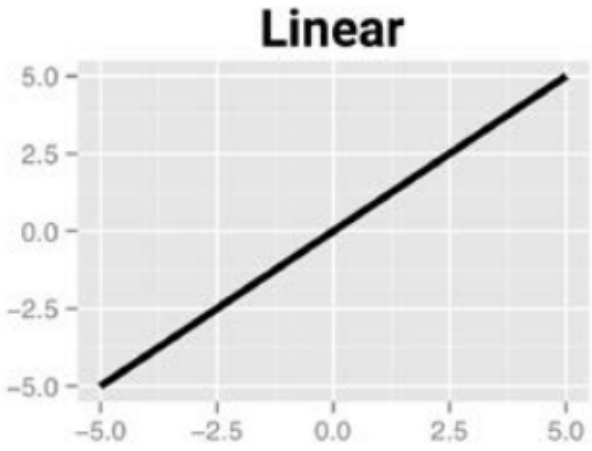
Activation Functions

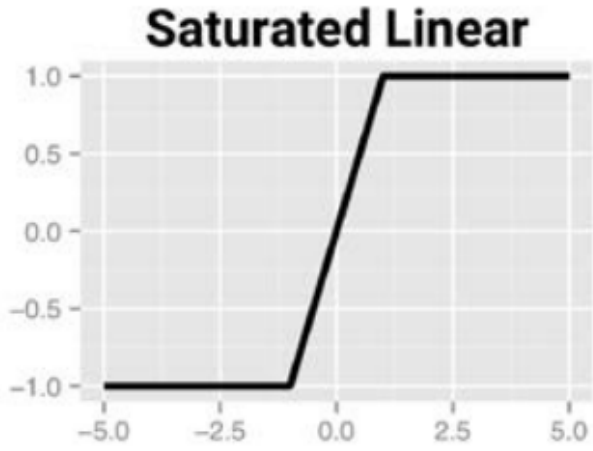
- Sigmoid Activation Function
- Here, output values can fall anywhere in the range from 0 to 1

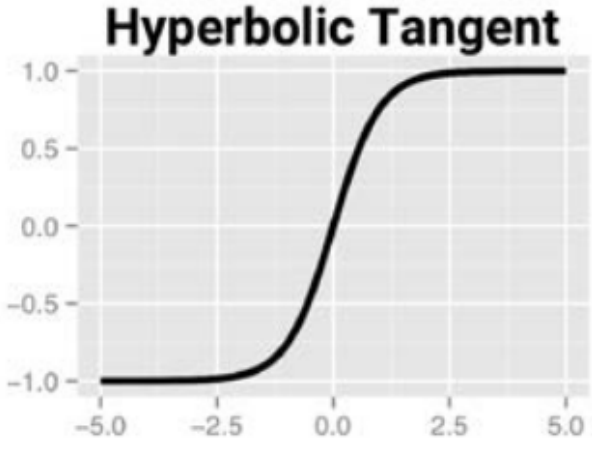


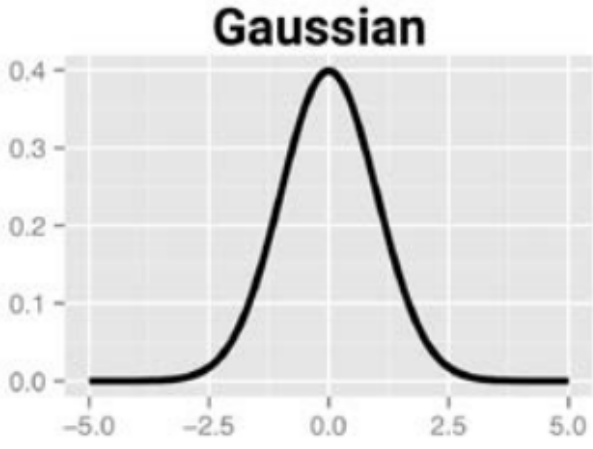
Activation Functions

- Other Activation Functions







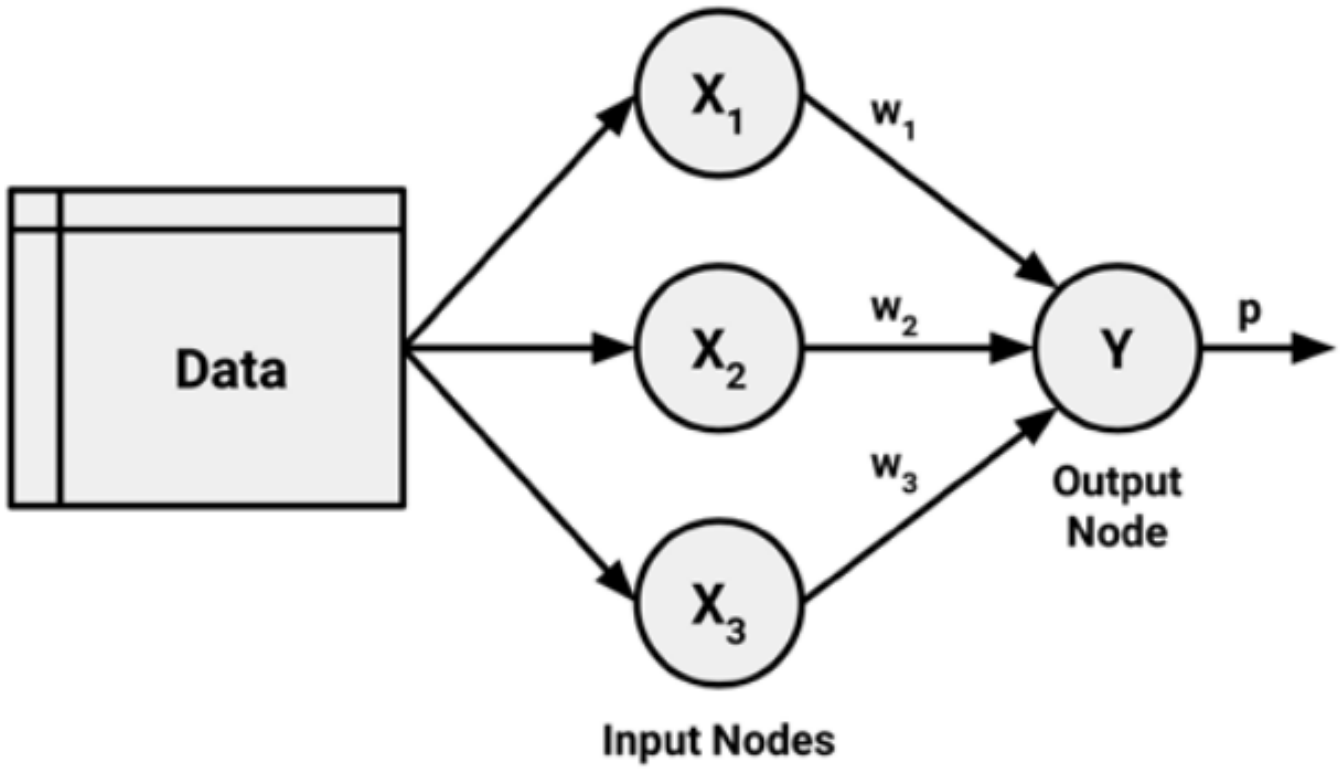


Network Topology

- Depending on the network topology, an artificial neural network can be classified into 2 types

1. Single Layer Network

- Here there is only one group of input nodes which are connected to the output node

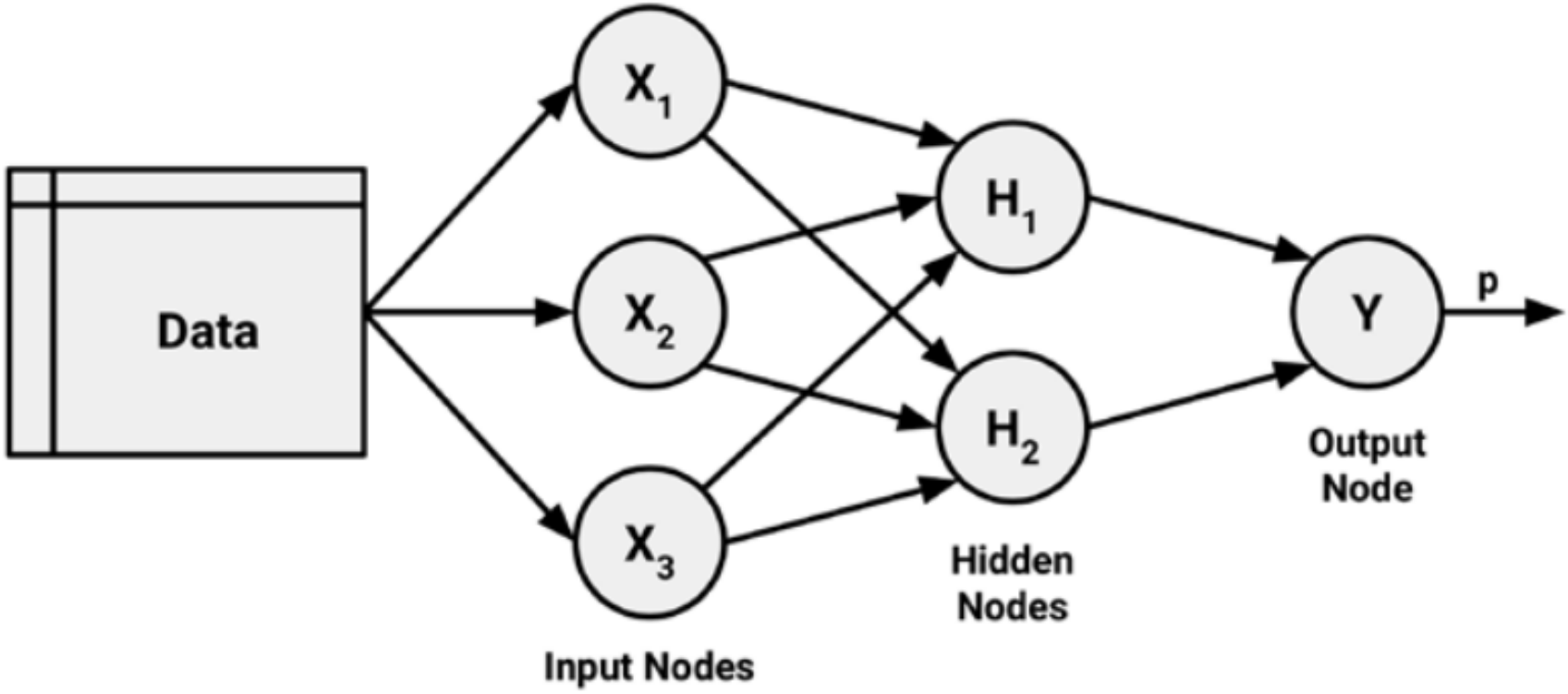


Network Topology

- Depending on the network topology, an artificial neural network can be classified into 2 types

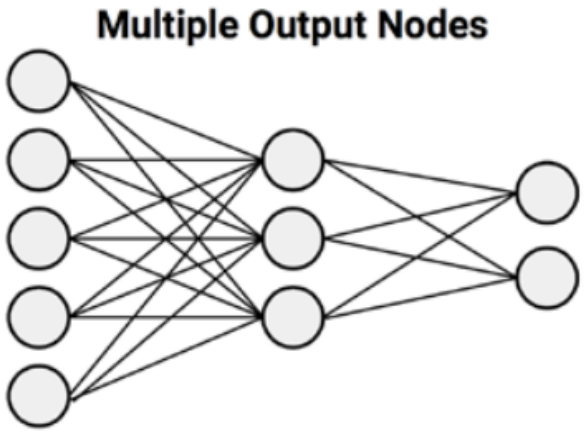
1. Multi Layer Network

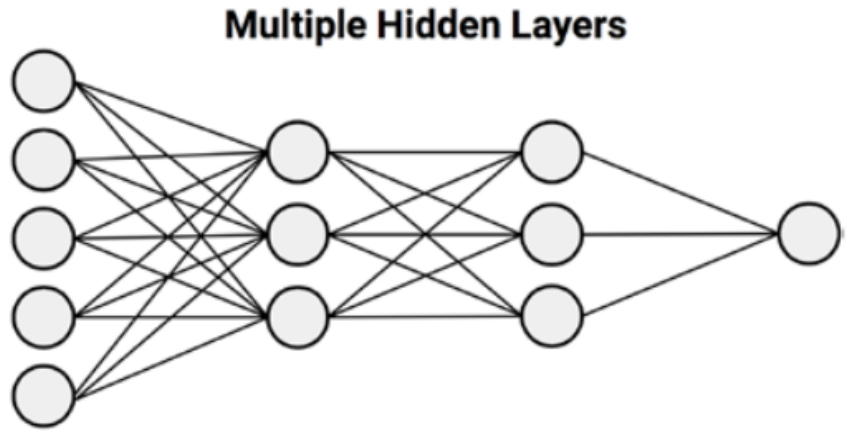
- Here there are one or more groups of hidden nodes between input nodes and the output node



Network Topology

- Depending on the neural network algorithm, there can be multiple output nodes or multiple hidden layers
- A neural network with multiple hidden layers is called a Deep Neural Network (DNN) and the learning performed there is called Deep Learning





Training neural networks with back propagation

- The back propagation algorithm is commonly used in neural networks to reduce its errors
- Initially weights are assigned to input signals at random
- This algorithm completes in multiple cycles, each one called an epoch
- Each cycle has 2 phases

1. Forward Phase

- The weighted input signals are processed across different layers and upon reaching the final layer, the output signal is produced

2. Backward Phase

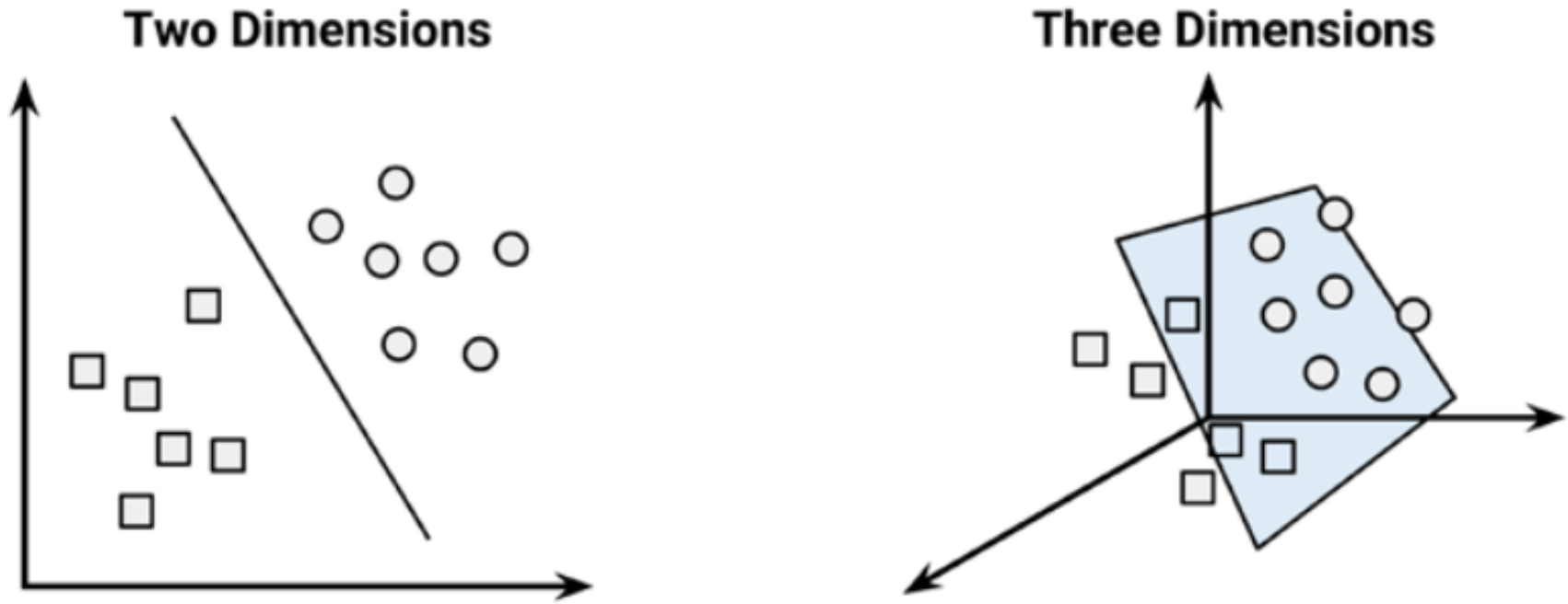
- The result produced by the output signal is compared with the output produced by the training data
- The corresponding error is propagated backwards through the network
- The weights of input signals are modified to reduce future errors

Support Vector machines: Hyperplanes

- A support vector machine(SVM) is a supervised machine learning algorithm used for classification and regression
- A support vector machine uses a boundary called hyperplane to partition data into similar class values
- Applications
- Classification of gene data to identify cancer or other genetic diseases
- Classification of documents by subject matter
- Detection of earthquakes

Classification Using Hyperplanes

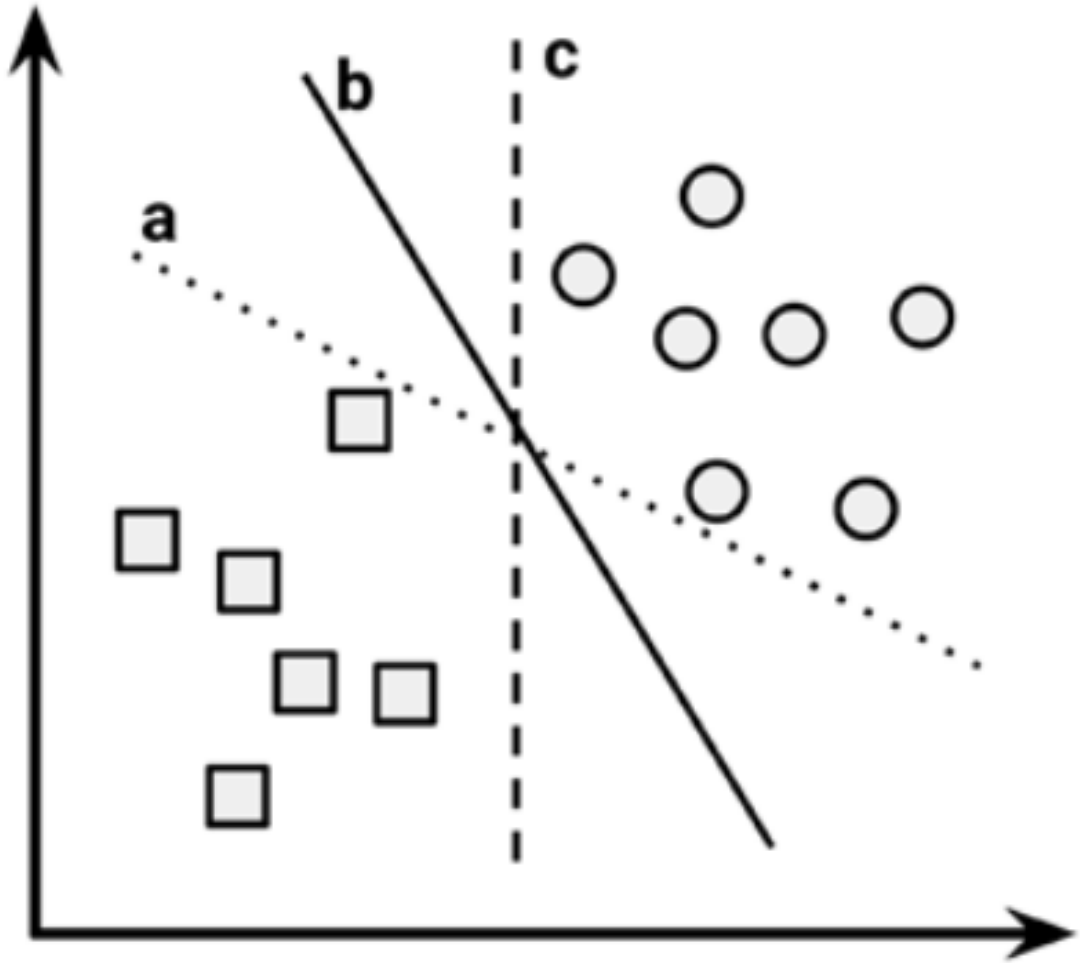
- A hyperplane that separates groups of circles and squares in two and three dimensions



- Since circles and squares can be perfectly separated using a straight line, they are said to be linearly separable

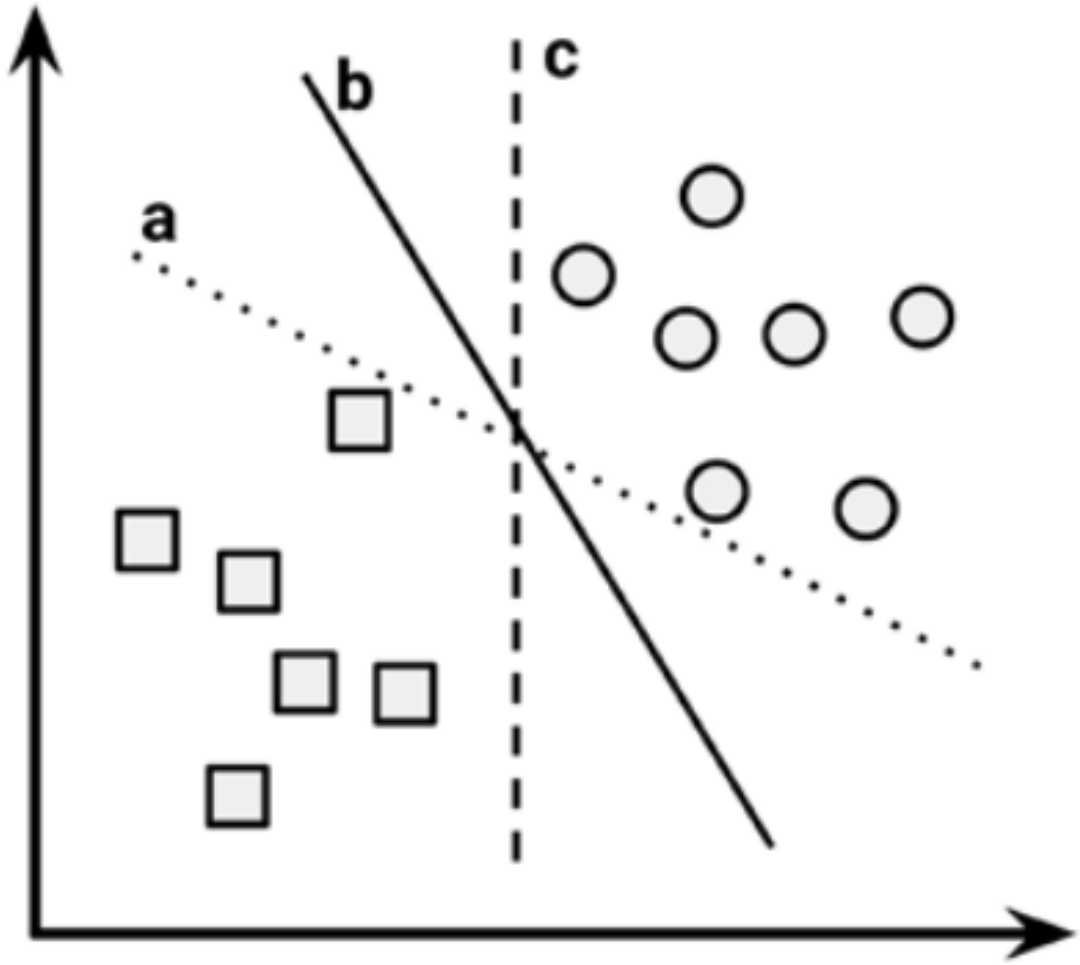
Maximum margin hyperplanes in linearly separable data

- There is more than one choice of dividing line between the groups of circles and squares
- The task of the SVM algorithm is to identify a line that separates the two classes



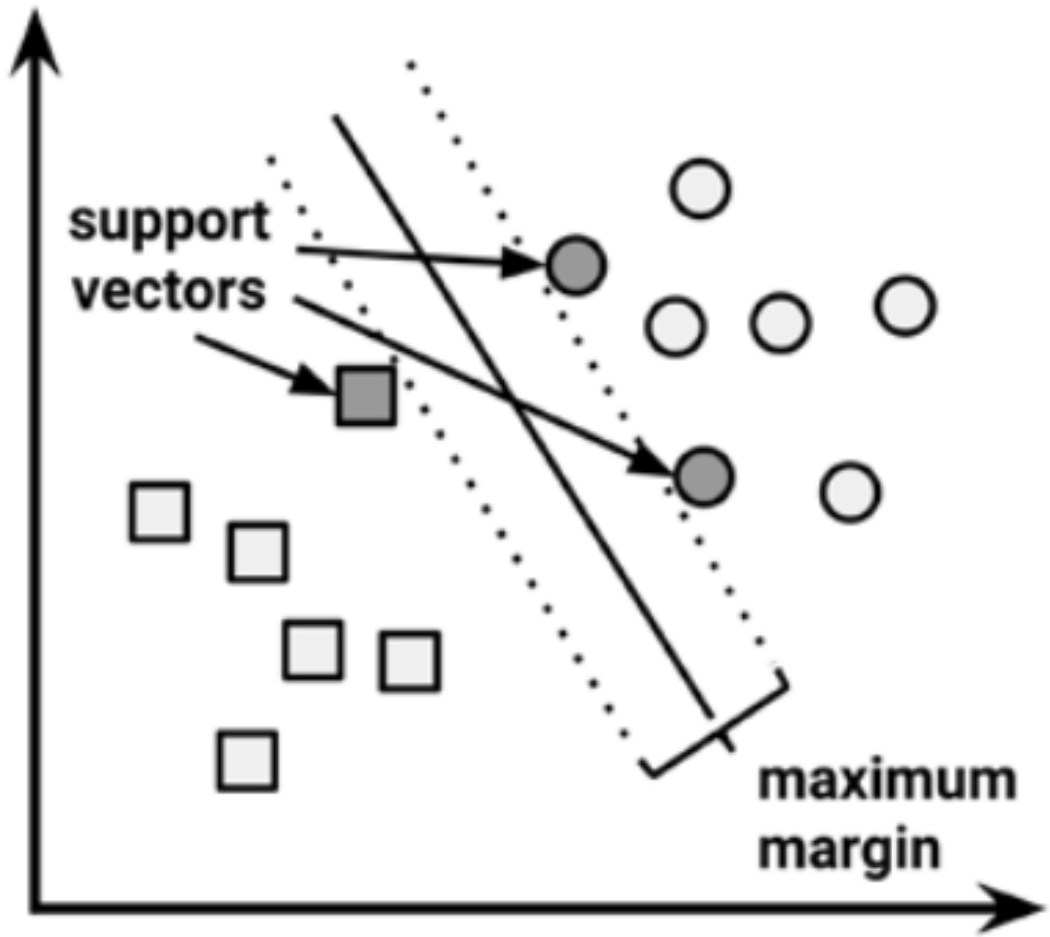
Maximum margin hyperplanes in linearly separable data

- The SVM algorithm searches for a line that causes greatest separation between the two classes
- This is called the Maximum Margin Hyperplane(MMH)



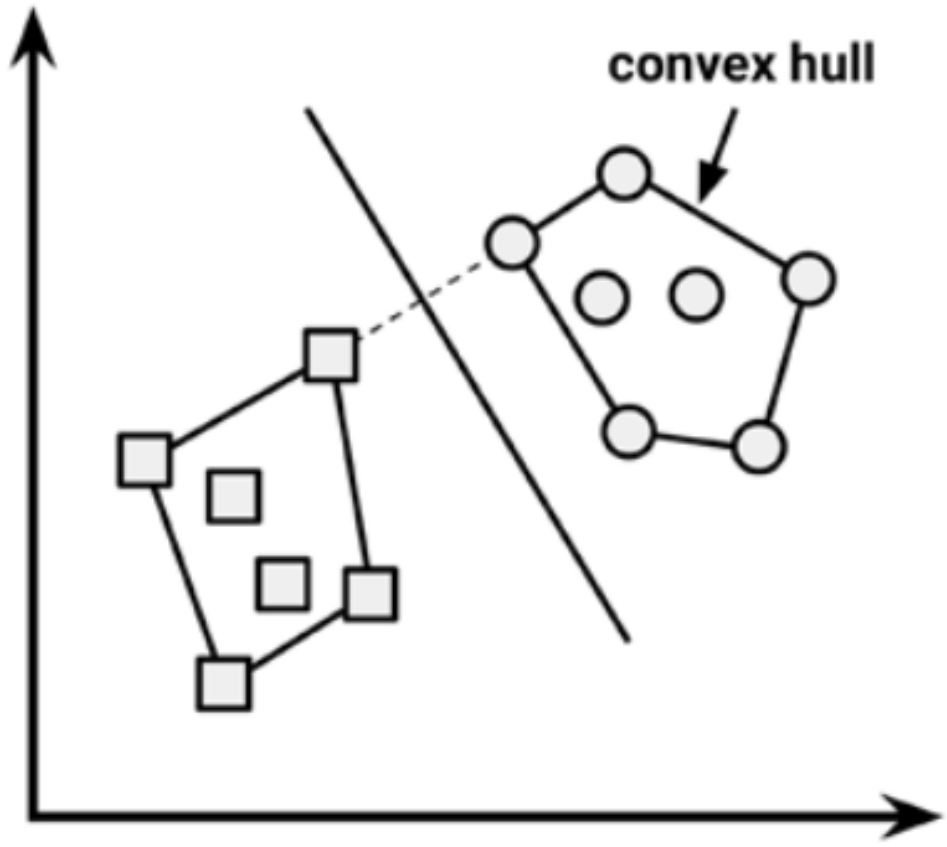
Maximum margin hyperplanes in linearly separable data

- The support vectors are the points from each class that are the closest to the MMH
- Each class must have one or more support vectors
- We can identify MMH using support vectors



Maximum margin hyperplanes in linearly separable data

- The MMH is as far away as possible from the outer boundaries of the two groups of data points
- These outer boundaries are known as the convex hull
- The MMH is then the perpendicular bisector of the shortest line between the two convex hulls

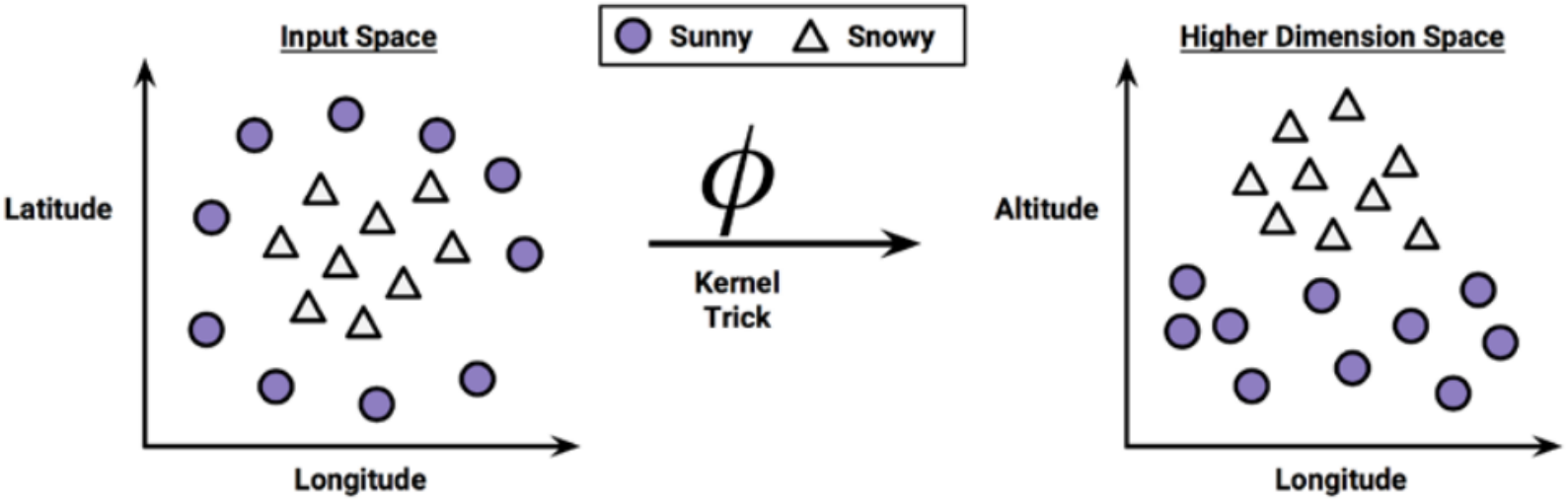


Using kernels for non-linear spaces

- In many real-world applications, the relationships between variables are nonlinear
- SVM can convert a non-linear relationship to a linear one using a process known as kernel trick
- In this process, the problem is mapped into a higher dimension space

Using kernels for non-linear spaces

- Weather reports from various stations



- While applying kernel trick, the dimension latitude is replaced by the dimension altitude
- The choice of the new dimension is done based on the knowledge that snowy weather is found at higher altitudes
- Now the classes are perfectly linearly separable

References

1. Vijay Kotu, Bala Deshpande, "Data Science Concepts and Practice", Morgan Kaufmann Publishers, 2018 (Module 1)
2. Brett Lantz, "Machine Learning with R", Second edition, PackT publishing, 2015 (Modules 2 to 5)