Artificial Intelligence & Machine Learning
Semester-VII (A.Y 2021-22)

Harsh Dhiman, Ph.D.

Adani Institute of Infrastructure
Department of Electrical Engineering

Evaluation

Component	Quiz-I	Quiz-II	MSE	Quiz-III
Date	19 Jul	16 Aug	3 Sep	1 Oct

Introduction to Artificial Intelligence

- Age of AI

Wikipedia says

Artificial intelligence (AI) is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality.

- When a computer tries to mimic human actions/decisions
- Understanding how decisions can be made

Brief history of AI

- The Greek myths contain stories of mechanical men designed to mimic our own behaviour.
- European computers were conceived as "logical machines" and by reproducing capabilities such as basic arithmetic and memory.
- Artificial Intelligence – devices designed to act intelligently

Contd...

- 1950 : The Alan Turing who was an English mathematician and pioneered Machine learning in 1950. Alan Turing publishes “Computing Machinery and Intelligence" in which he proposed a test. The test can check the machine’s ability to exhibit intelligent behavior equivalent to human intelligence, called a Turing test.
- 1980 : After AI winter duration, AI came back with "Expert System". Expert systems were programmed that emulate the decision-making ability of a human expert.
- 2006 : AI came in the Business world till the year 2006. Companies like Facebook, Twitter, and Netflix also started using AI.

Life with AI

- Search engines like Google and Bing
- Digital assistants like Siri and Google assistant
- Shopping recommendation : Amazon, Flipkart

What is common in all of the above ?

DATA and lots of DATA !

- Data brings rich experience

Ai at Amazon.com

- Amazon is using AI for improving business efficiency and optimize delivery time
- Anticipating customer demands and realizing it to a mathematical model
- Location specific anticipation
- Dynamic pricing via AI tools. Increase in price of a product when demand is high

Case Study : Shipment Forecasting Solution to Enable Accurate Trade & Inventory Planning [1]

Objective

Our client is a global healthcare company, headquartered in the US. Their Trade S&OP team needed to accurately forecast shipments that would help them in effective inventory management.
This was required to be done for their diabetes care unit, specifically for high-selling glucose monitoring devices and test strips.

1. Inventory planning is the process of determining the optimal quantity and timing of inventory for the purpose of aligning it with sales and production capacity. Inventory planning affects a company's cash flow and profits while contributing to an efficient supply chain.

Challenges to the current problem

- Identifying and treating outliers in historical purchases
- Accounting for retailer level differences
- Aligning model based forecasting results to quarterly financial company targets

Solution Methodology

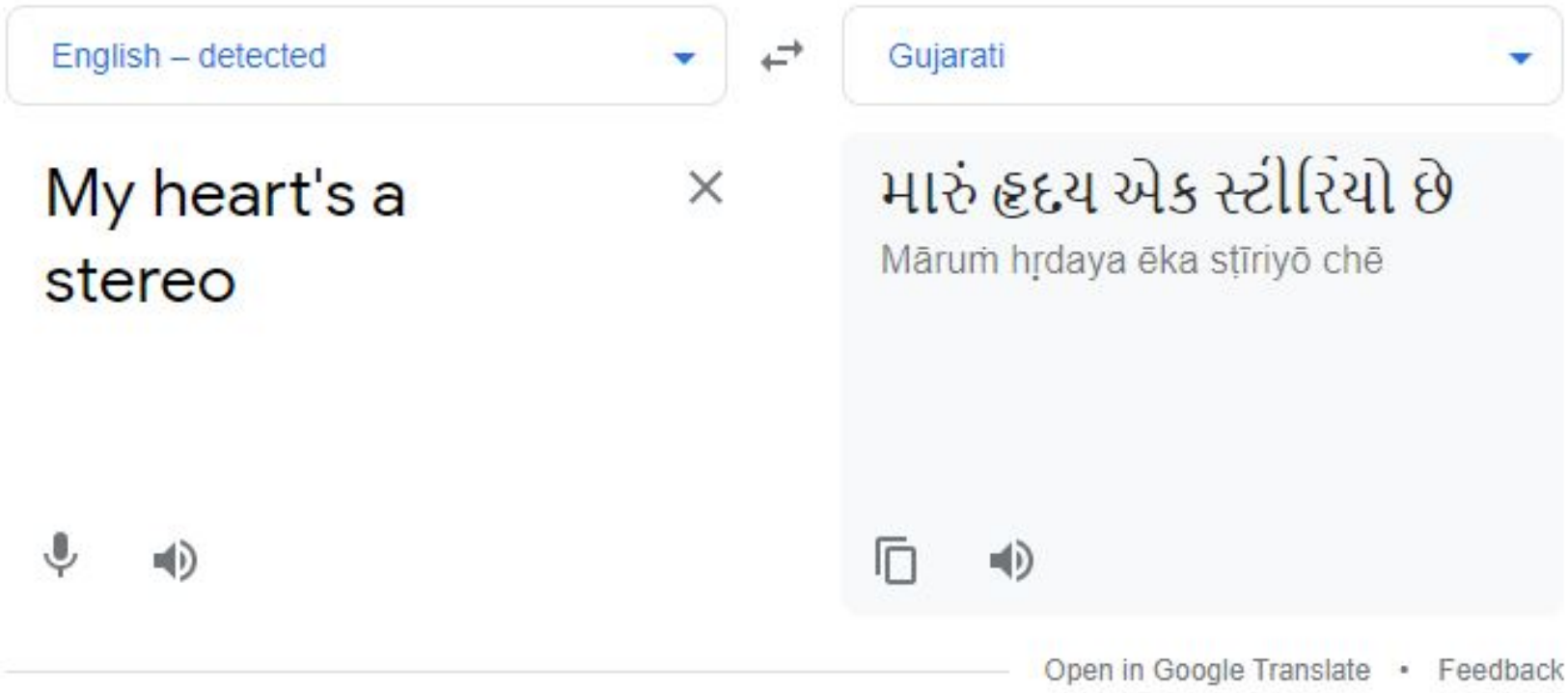
- Trade forecasting tool at a product SKU [2] , wholesaler/retailer customer level
- Seamless updated forecasts based on latest trends and data
- Customized BI/ Visualization reports

2. Stock keeping unit

Everyday AI : Shopping Recommendations on Amazon.in

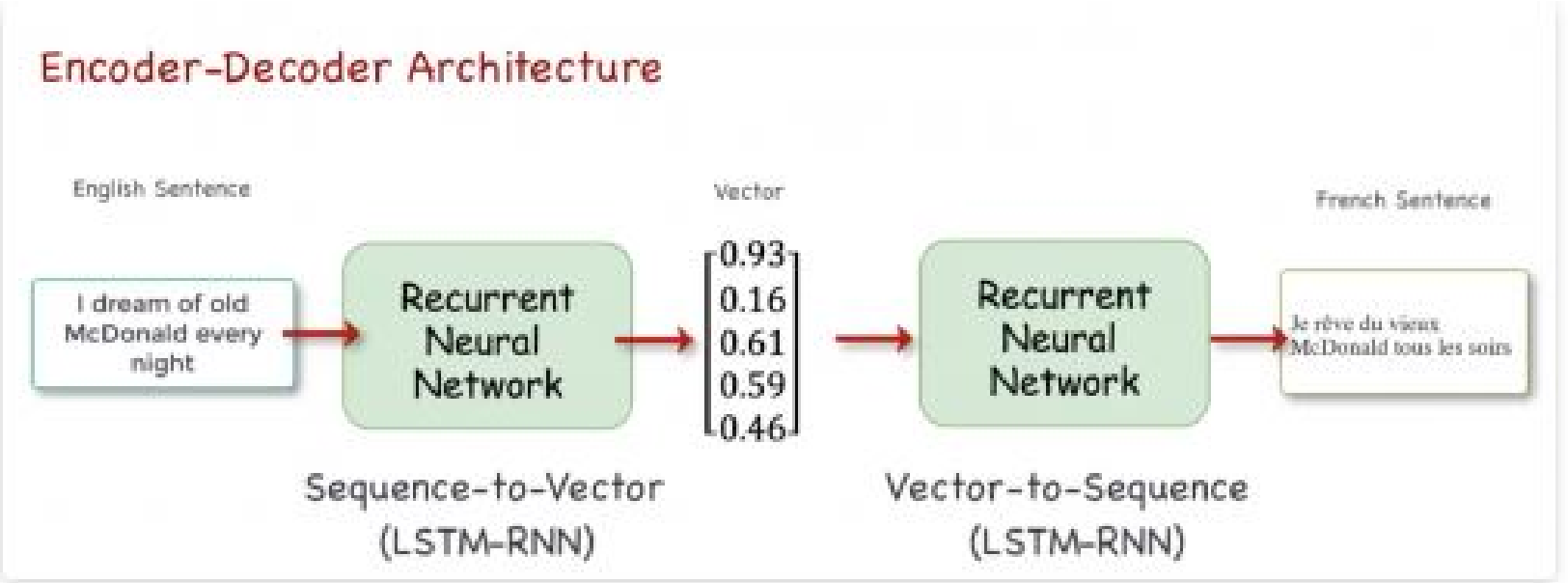
- Based on your history of purchases
- A complete recommendation system based on category of purchase

Google Translate : Natural Language Processing



- Google translate is based on Natural language processing
- Uses neural network architecture

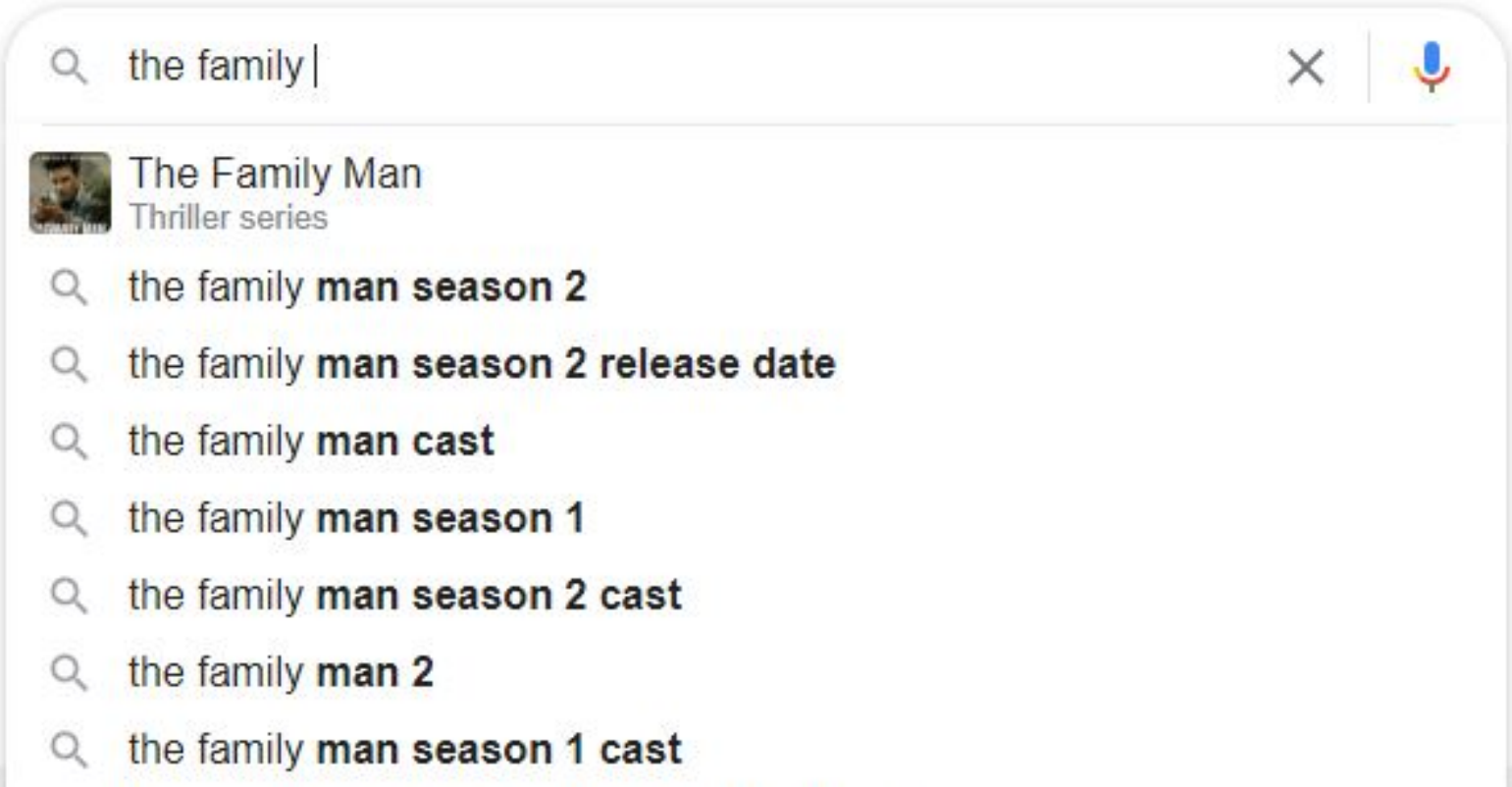
How does this architecture looks like ?



Everyday AI : Apple's Siri

- Siri is an integrated, voice-controlled personal assistant available for users of the Apple computing and telecommunications platform.
- Siri consists of three main components : a conversational interface, personal context awareness, and service delegation.
- Language data statistical analysis and machine learning power the personal context awareness portion of Siri. This enables the system to decipher the meaning of what you actually say to it.

Google Auto-complete



Contd...

- You'll notice we call these auto-complete "predictions" rather than "suggestions," and there's a good reason for that.
- Auto-complete is designed to help people complete a search they were intending to do, not to suggest new types of searches to be performed.
- The predictions change in response to new characters being entered into the search box.

AI on celluloid : Ultron from Avengers : Age of Ultron

AI on celluloid : R2D2 & C-3PO from Star Wars

Pros of AI

- Highly accurate
- Fast in decision making
- Highly reliable
- Aids in public utility like face recognition and speed to text using natural language processing

Why AI might be a bad idea ?

- High cost of implementation
- Can't replace humans
- Lacks creativity [3]
- Risk of unemployment
- AI is not empathetic

3. AI cannot be creative, in that it cannot experience something like inspiration, but rather it can learn creative behaviours and mimic the human creative process, via which it can produce an output

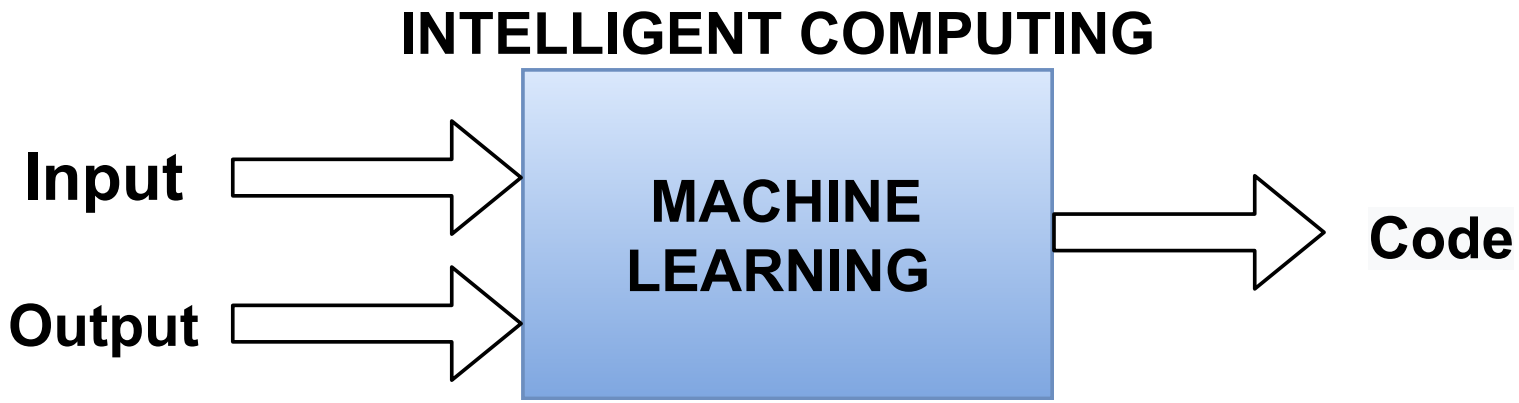
Chapter-2
Introduction to Machine Learning

Introduction

- Computers are powerful machines
- When a machine learns, it understands a behavior or pattern in the data
- Without data, learning is pointless

Data is the King!

Traditional v/s Intelligent Computing



Statistical learning

- Plays a key role in many areas of science, finance and industry. Here are some examples of learning problems
- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.

Contd...

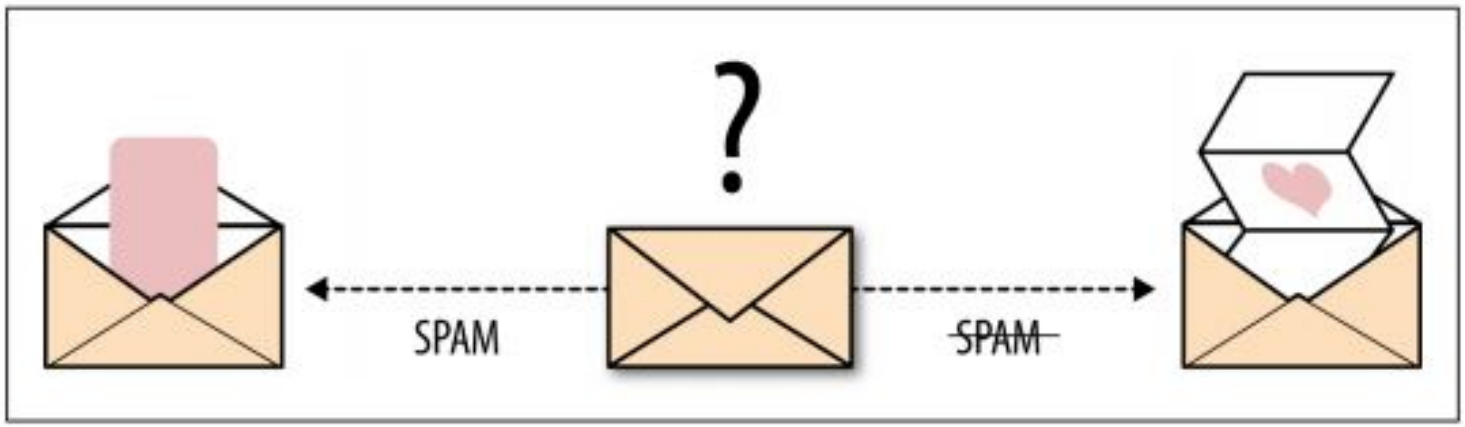
- In a typical scenario, we have an outcome measurement, usually quantitative (such as a stock price) or categorical (such as heart attack/no heart attack), that we wish to predict based on a set of features (such as diet and clinical measurements).
- We have a training set of data, in which we observe the outcome and feature measurements for a set of objects (such as people).
- Using this data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects.
- A good learner is one that accurately predicts such an outcome

Supervised Learning

- Learning from experience or under supervision
- Each input has a specified output/label

Case 1 : The Email that goes to Spam !

- The objective was to design an automatic spam detector that could filter out spam before clogging the users' mailboxes.



Contd...

- For all collected email messages, the true outcome (email type) email or spam is available.
- 57 of the most commonly occurring words and punctuation marks in the email message.
- This is a supervised learning problem, with the outcome the class variable email/spam.

Learning rule

- Our learning method has to decide which features to use and how
- For example,
  if (%cash $>$ 2) && (%donate $>$2.4), then spam ; else email

Case 2 : Computer Vision

- Is it a cat or a dog ?
- Image classification is a popular problem in the computer vision field.
- Here, the goal is to predict what class an image belongs to.
- In this set of problems, we are interested in finding the class label of an image.
- More precisely : is the image of a car or a plane ? A cat or a dog ?

Case 3 : House prices

- First, we need data about the houses : square footage, number of rooms, features, whether a house has a garden or not, and so on.
- We then need to know the prices of these houses, i.e. the corresponding labels.
- By leveraging data coming from thousands of houses, their features and prices, we can now train a supervised machine learning model to predict a new house's price based on the examples observed by the model.

Bayesian Classifier for Supervised Learning

- According to Bayes theorem, we can estimate conditional probability [4]
- The conditional probability of an event Y is the probability that the event will occur given the knowledge that an event X has already occurred.

4. Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome.

Contd...

- For example, the probability that any given person has a cough on any given day may be only 5%. But if we know or assume that the person is sick, then they are much more likely to be coughing.
- For example, the conditional probability that someone unwell is coughing might be 75%, in which case we would have that P(Cough) = 5% and P(Cough|Sick) = 75%.

Contd...

- This probability is written $P(Y|X)$, notation for the probability of Y given X.
- In the case where events X and Y are independent (where event X has no effect on the probability of event Y), the conditional probability of event Y given event X is simply the probability of event X, that is $P(Y)$.
- However, according to Bayes theorem

$$P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)} \tag{1}$$

Let us consider an example

- What is the probability that a patient has diseases meningitis with a stiff neck ?

Given data

A doctor is aware that disease meningitis causes a patient to have a stiff neck, and it occurs 80% of the time. He is also aware of some more facts, which are given as follows :

- The known probability that a patient has meningitis disease is 1/30000.
- The known probability that a patient has a stiff neck is 2%.

Let's apply Bayesian Stats

- Consider event X that a patient has a stiff neck and Y that a person has meningitis disease.
- Therefore, $P(X|Y) = 0.8,\ P(Y) = 1/30000,\ P(X) = 0.002$
- Hence,

$$P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)} \tag{2}$$

$$= \frac{0.8 \times (1/30000)}{0.002} \tag{3}$$

$$= 0.001333$$

Spam Mail Detection via Bayesian Classifier

- Consider a set of 12 emails out which 8 are normal and 4 are spam
- Consider the following distribution of words/phrases for normal and spam emails

Word	Normal	Spam
Dear	8	2
Friend	5	1
Lunch	3	0
Money	1	4

Contd...

- Probability of occurrence of the word dear in a normal email is $p(dear|normal) = 8/17 = 0.47$
- Similarly, the likelihood for other words are

Word	Normal	Spam
Dear	8/17	2/7
Friend	5/17	1/7
Lunch	3/17	0/7
Money	1/17	4/7

- The task is to estimate the probability is normal or spam given a set of words/phrases

Contd...

- What is the probability that the given email is normal given the word is Money ?
- Thus, we need to evaluate $P(Normal|Money)$
- As per Bayes theorem,

$$\begin{aligned} P(Normal|Money) &= \frac{P(Money|Normal).P(Normal)}{P(Money)} \\ &= \frac{1/17 \times 8/12}{5/24} \\ &= 0.1882 \end{aligned}$$

Example : Decide whether to play or not based on weather

- Following is the training data

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Contd...

- Consider prior class probabilities $p(yes)$ and $p(no)$

$$p(yes) = \frac{n(yes)}{N} = \frac{9}{14} \tag{4}$$

$$p(no) = \frac{n(no)}{N} = \frac{5}{14} \tag{5}$$

- Given the distribution, we now will calculate the likelihood $p(sunny|yes)$

$$p(sunny|yes) = \frac{3}{9} = \frac{1}{3} \tag{6}$$

Contd...

- Prior probability of predictor, $p(sunny)$ is given as

$$p(sunny) = \frac{5}{14} \tag{7}$$

Hence, the posterior probability

$$\begin{aligned} p(yes|sunny) &= \frac{p(sunny|yes) \times p(yes)}{p(sunny)} \\ &= \frac{1/3 \times 9/14}{5/14} \\ &= 0.6 \end{aligned}$$

Unsupervised Learning

- In supervised learning, the main idea is to learn under supervision, where the supervision signal is named as target value or label.
- In unsupervised learning, we lack this kind of signal. Therefore, we need to find our way without any supervision or guidance.
- This simply means that we are alone and need to figure out what is what by ourselves.

Examples

- However, we are not totally in the dark. We do this kind of learning every day.
- In unsupervised learning, even though we do not have any labels for data points, we do have the actual data points.
- This means we can draw references from observations in the input data.

Analogy for Unsupervised Learning

- Imagine you are in a foreign country and you are visiting a food market, for example. You see a stall selling a fruit that you cannot identify. You don't know the name of this fruit.
- However, you have your observations to rely on, and you can use these as a reference. In this case, you can easily the fruit apart from nearby vegetables or other food by identifying its various features like its shape, color, or size.

Contd...

- This is roughly how unsupervised learning happens. We use the data points as references to find meaningful structure and patterns in the observations.
- Unsupervised learning is commonly used for finding meaningful patterns and groupings inherent in data, extracting generative features, and exploratory purposes.

Common examples of unsupervised learning

Customer segments

- Clustering is an unsupervised technique where the goal is to find natural groups or clusters in a feature space and interpret the input data.
- Clustering is commonly used for determining customer segments in marketing data. Being able to determine different segments of customers helps marketing teams approach these customer segments in unique ways. (Think of features like gender, location, age, education, income bracket, and so on.)

Case 2 : Identifying Fake News

- Fake news is not a new phenomenon, but it is one that is becoming prolific.
- What the problem is : Fake news is being created and spread at a rapid rate due to technology innovations such as social media. The issue gained attention recently during the 2016 US presidential campaign. During this campaign, the term Fake News was referenced an unprecedented number of times.
- The way that the algorithm works is by taking in the content of the fake news article, the corpus, examining the words used and then clustering them. These clusters are what helps the algorithm determine which pieces are genuine and which are fake news

Sentiment Analysis

- Sentiment analysis is an automated process capable of understanding the feelings or opinions that underlie a text.
- It is one of the most interesting subfields of NLP [5], a branch of Artificial Intelligence (AI) that focuses on how machines process human language.
- Sentiment analysis studies the subjective information in an expression, that is, the opinions, appraisals, emotions, or attitudes towards a topic, person or entity.
- Expressions can be classified as positive, negative, or neutral.

5. Natural language processing

Examples for sentiment analysis

- " I really like your shirt, its Cool !" - Positive
- " I'm not sure if that's a good idea" - Neutral
- " The cappuccino at Starbucks is pathetic" -Negative

Chapter-3
Linear Regression in Machine Learning

Introduction

- Relationship between variables
- Independent and dependent variables
- One variable may or may not heavily depend on other variables

What are these variables ?

- Entities such as sales, income, stock price, and weather parameters
- Difference between prediction and classification.
- Prediction involves continuous variables or range. For example : stock price
- Classification deals with discrete variables. For example, classifying cancer stages

Linear Regression : Intro

- A statistical learning technique
- Often used for predictive modeling
- Benchmark model for R&D in AI & ML

Crux

Linear regression algorithms show a linear relationship between a dependent variable, y, and one or more independent variables,x i.e., how the value of the dependent variable, y changes according to the value of the independent variable.

Contd...

- Linear regression can be of univariate or multivariate nature
- Univariate linear regression deals with a single independent variable x
- Multivariate linear regression deals with more than one independent variable $x_1, x_2, \ldots, x_n$

Univariate Linear Regression

- A firm decides to check out the relationship between advertising and sales, which can be modeled by the use of linear regression.
- The estimation of the price of a house depending on the number of rooms it has increases or decreases.
- The estimation of the price of a house depending on the number of rooms it has increases or decreases.

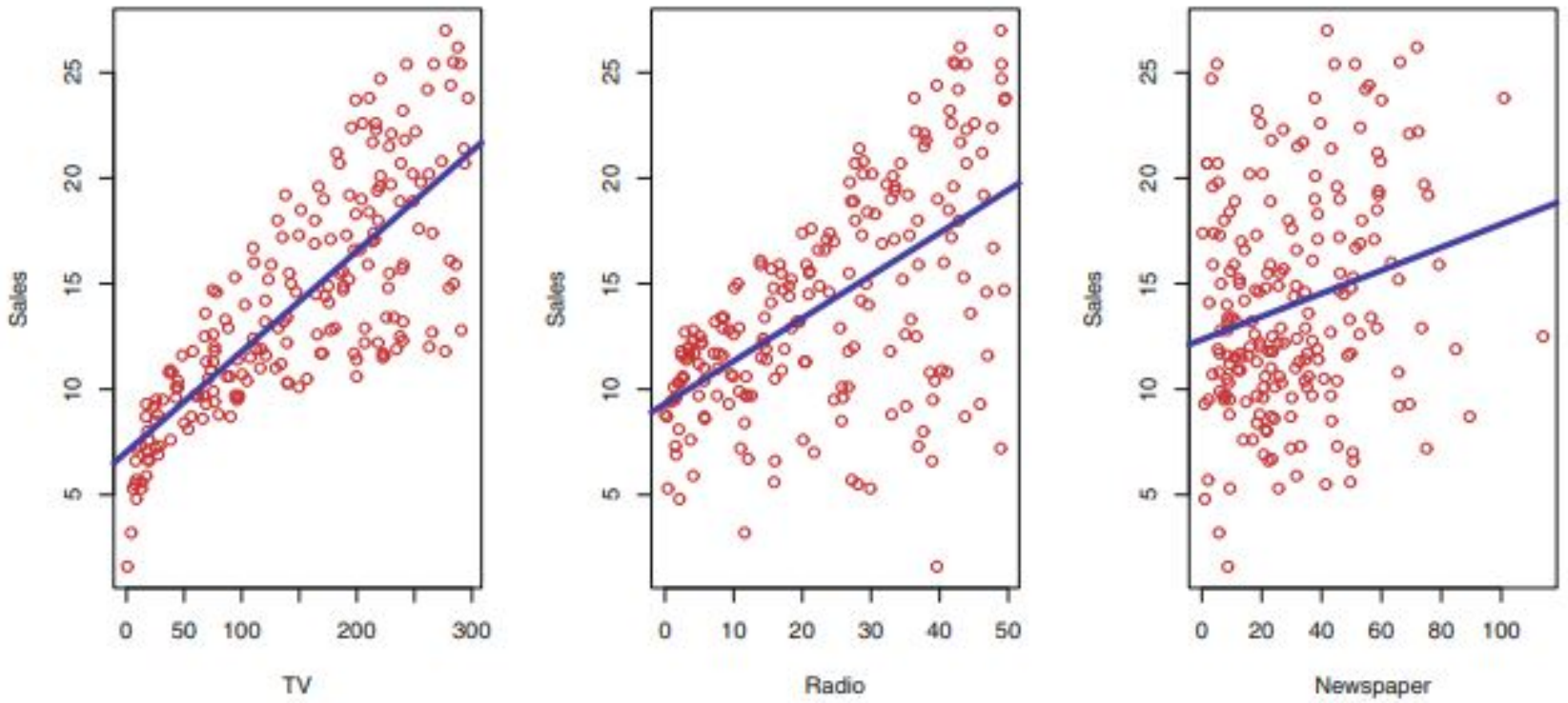
Case Study : Sales v/s Advertising Budget

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media : TV, radio, and newspaper.

Contd...

- In this setting, the advertising budgets are input variables while sales input is an output variable.
- The input variables are typically denoted using the symbol X, with a subscript to distinguish them. So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget.
- The inputs go by different names, such as predictors, independent variables, features or predictor
- In this case, sales—is variable often called the response or dependent variable, and is typically denoted using the symbol Y

Contd...


Sales
TV
Radio
Newspaper
0
50
100
200
300
10
20
30
40
60
80
5
15
25

Important Points

- Is there a relationship between advertising budget and sales ?
- How strong is the relationship between advertising budget and sales ?
- Which media contribute to sales ?

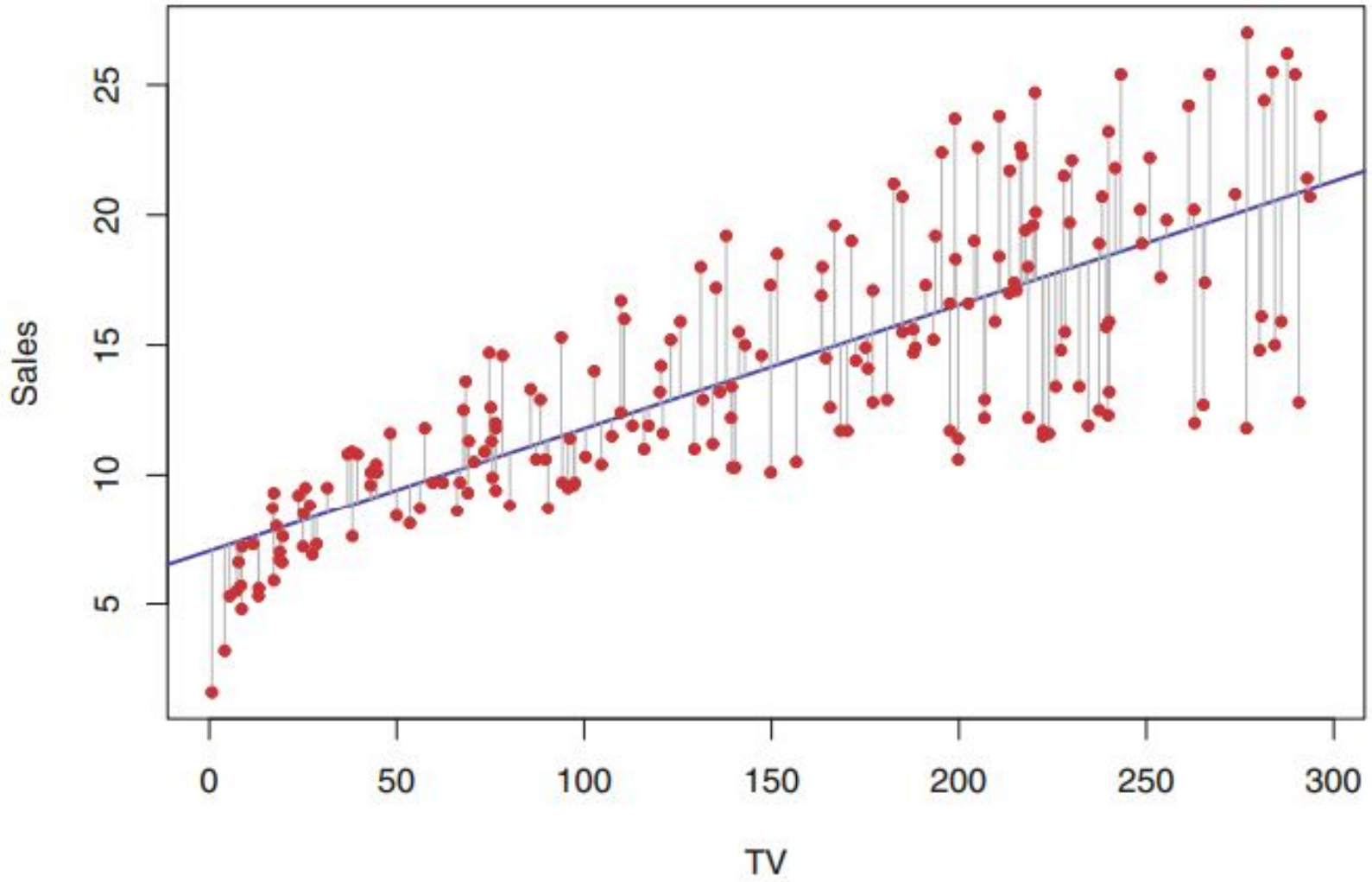
Univariate Regression for Sales v/s Advertising Budget

- Given predictor X (TV budget) and response Y (Sales), we have

$$\texttt{Sales} \approx \theta_1 \times \texttt{TV} + \theta_0 \tag{8}$$

- θ_1 and θ_0 are known as model coefficients or parameters
- Once these parameters are known from training data, we can predict future values of sales based on a given TV budget

Scatter plot for visualization



Univariate Linear Regression : Mathematical Perspective

- Consider a univariate variable x_i as input or independent variable where $i = 1, 2, \ldots, n$
- Given (x_i, y_i) as the pair of input and output variables
- A univariate linear regression model is expressed as

$$y = \theta_1 x + \theta_0 + \varepsilon \tag{9}$$

  where θ_1 is the slope parameter and θ_0 is the intercept
- θ_1 and θ_0 are known as regression parameters

Least-squares for coefficient estimation

- The linear regression model for each sample can be given as

$$y_i = \theta_1 x_i + \theta_0 + \varepsilon_i \tag{10}$$

- Consider the objective of linear regression to be $J(\theta_1, \theta_0)$ which is to be minimized. We can write

$$J(\theta_1, \theta_0) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \tag{11}$$

Contd...

- Partial derivative of J w.r.t. θ_0 and θ_1 gives

$$\frac{\partial J}{\partial \theta_0} = -2\sum_{i=1}^{n}\left(y_i - \theta_1 x_i - \theta_0\right) \tag{12}$$

$$\frac{\partial J}{\partial \theta_1} = -2\sum_{i=1}^{n}\left(y_i - \theta_1 x_i - \theta_0\right)x_i \tag{13}$$

- Equate (12) and (13) to zero and we get

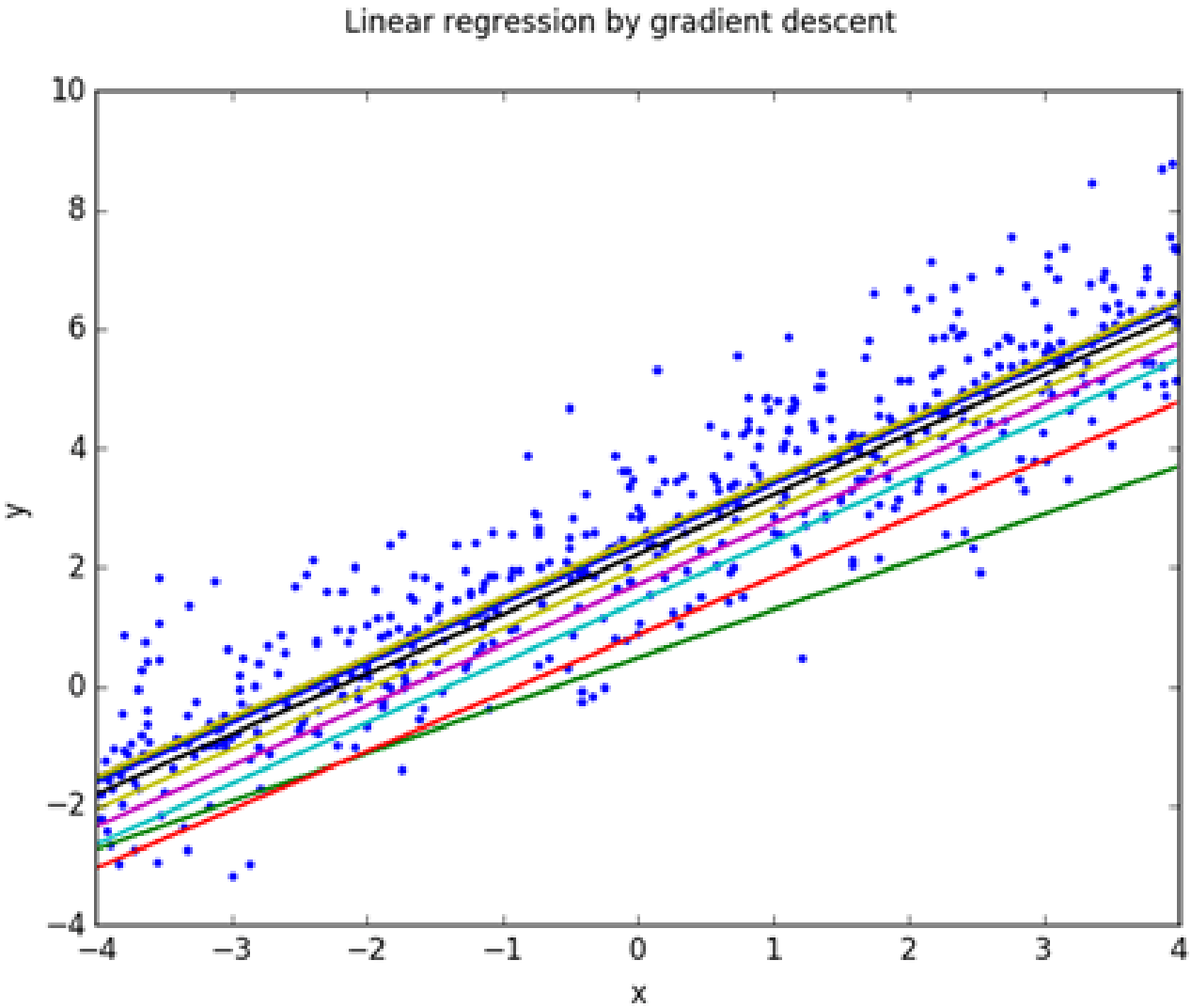
$$\theta_0 = \bar{y} - \theta_1 \bar{x} \tag{14}$$

$$\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{15}$$

How to optimize parameter estimates ?

- Different parameter estimates give different fit to the data
- The global aim is to minimize objective function $J(\theta_0, \theta_1)$
- We need the optimal parameters for our data
- We use gradient descent for this purpose

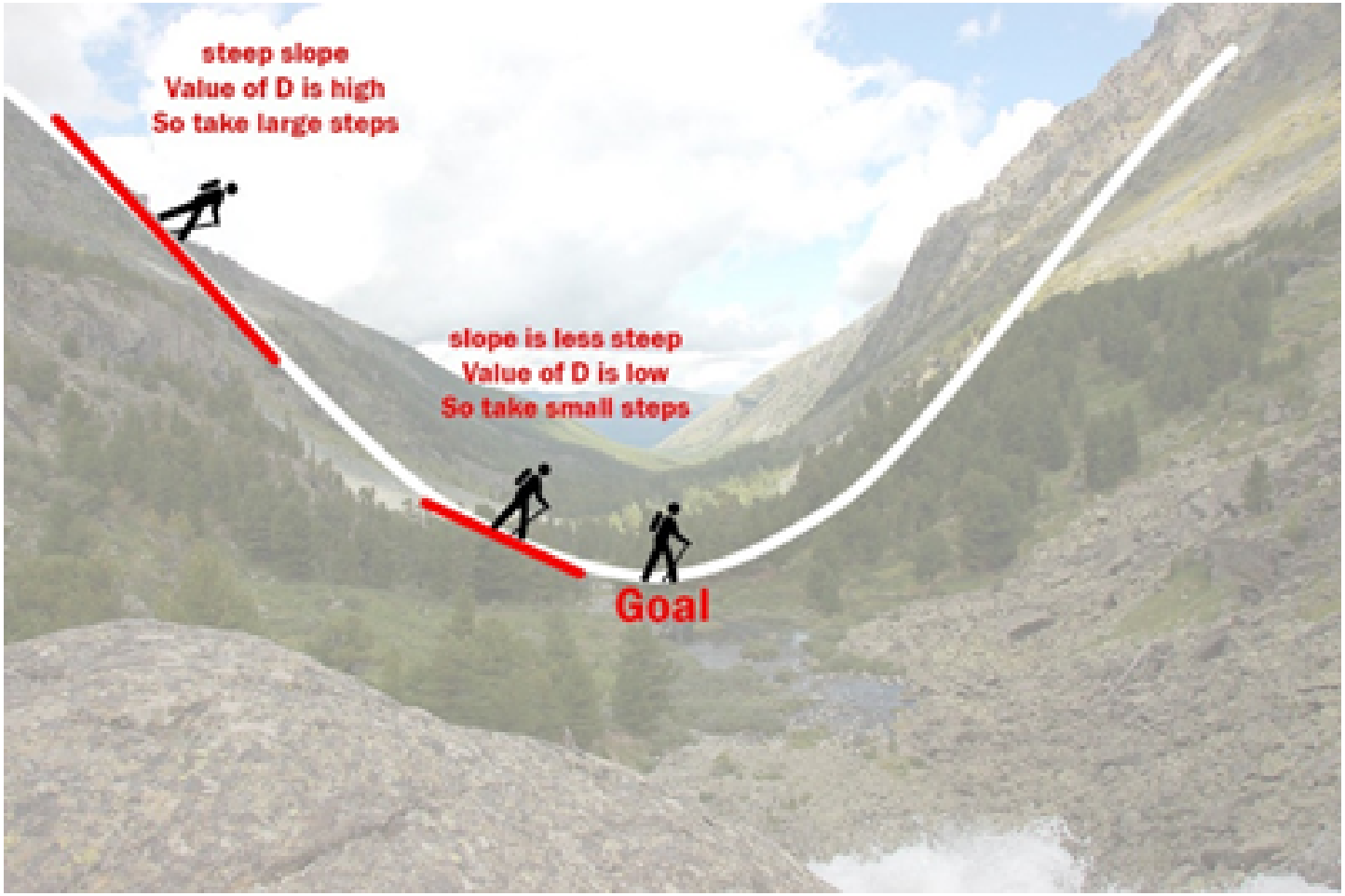
Why do we need gradient descent ?



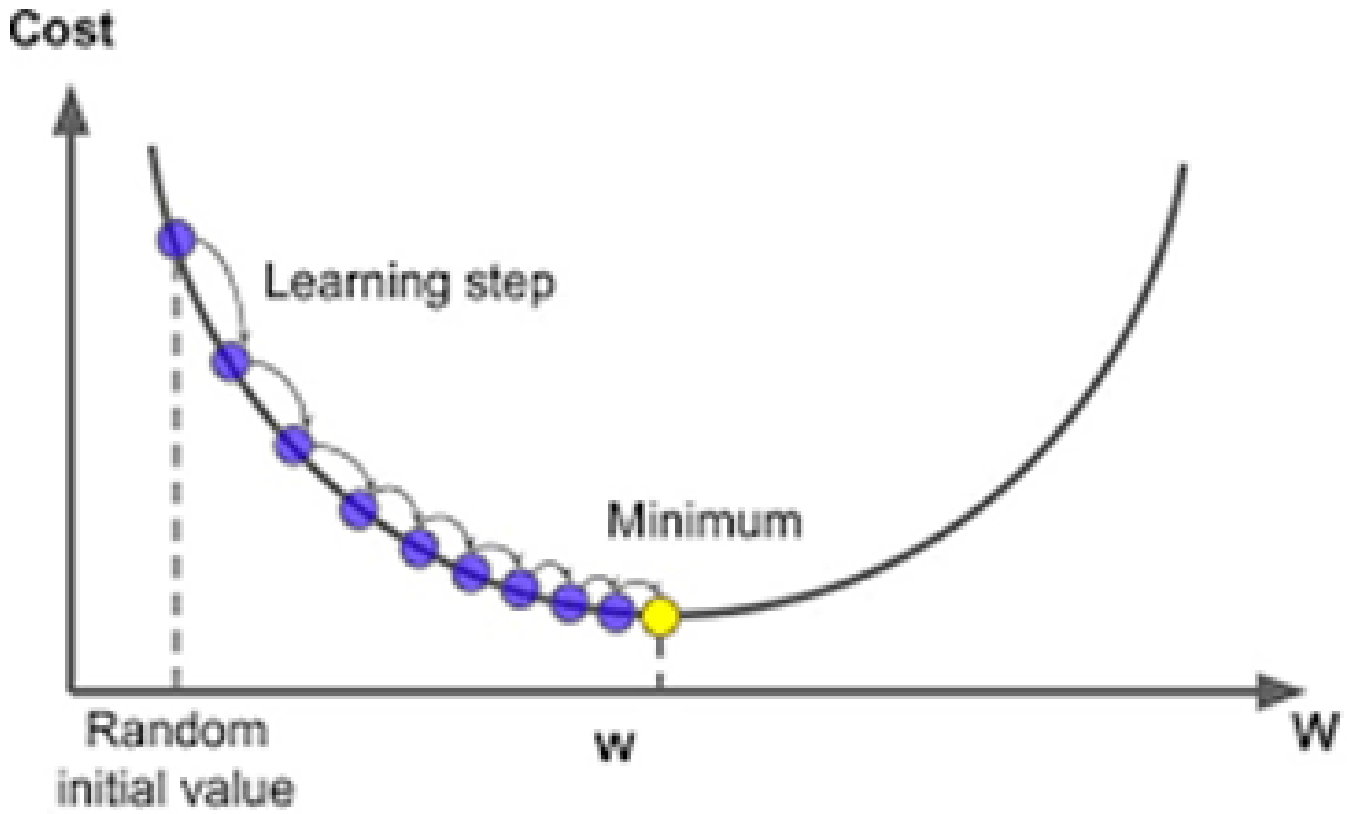
Gradient Descent for Linear Regression

- In machine learning, we use gradient descent to update the parameters of our model.
- Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.
- Initialize your parameters with random values
-

Visual representation of gradient descent



Contd...



Gradient Descent : Mathematical Perspective

- Consider objective function J

$$\min_{\theta} \; J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{16}$$

- In order to seek optimal parameters θ_0 and θ_1,

$$\frac{\partial J}{\partial \theta_0} = -2 \sum_{i=1}^{n} \left(y_i - \theta_1 x_i - \theta_0 \right) \tag{17}$$

$$\frac{\partial J}{\partial \theta_1} = -2 \sum_{i=1}^{n} \left(y_i - \theta_1 x_i - \theta_0 \right) x_i \tag{18}$$

Contd...

- Initialize $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_m]$ vector
- Update m weights (or parameters) vector using GD rule

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J}{\partial \theta_j} \tag{19}$$

- Repeat the above step until a convergence rule is reached

$$\| \theta_j - \theta_{j+1} \| < \varepsilon \tag{20}$$

Observations

- α plays an important role in fitting
- Small value of α leads to slow convergence
- Large values of α may fail to converge or might diverge instead

Github repo

Follow this link for python implementation of gradient descent
`https://github.com/harshdhiman-ai/Gradient-Descent-Linear-Regression`

Multivariate Linear Regression

- As the name implies, multivariate regression is a technique that estimates a single regression model with more than one outcome variable.
- When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

Examples

- A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students.
- She is interested in how the set of psychological variables is related to the academic variables and the type of program the student is in.

Contd...

- A doctor has collected data on cholesterol, blood pressure, and weight.
- She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week).
- She wants to investigate the relationship between the three measures of health and eating habits.

Multivariate regression : Mathematical perspective

- Consider data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Consider m features $(x_{i1}, x_{i2}, \ldots, x_{im})$
- We estimate coefficients $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$
- The objective is to minimize residual sum of squares (RSS)

$$RSS(\theta) = \sum_{i=1}^{n} \Big(y_i - h(\theta, x_i) \Big)^2 \tag{21}$$

Contd...

- In matrix form,

$$\text{RSS}(\theta) = (\boldsymbol{y} - \boldsymbol{X}\theta)^T(\boldsymbol{y} - \boldsymbol{X}\theta) \tag{22}$$

- Denote by $\boldsymbol{X}$ the $N \times (m+1)$ matrix with each row an input vector (with a 1 in the first position), and similarly let $\boldsymbol{y}$ be the N-vector of outputs in the training set.
- We have ,

$$\frac{\partial RSS}{\partial \theta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\theta) \tag{23}$$

Contd...

- Further,

$$\frac{\partial^2 RSS}{\partial \theta \partial \theta^T} = 2\boldsymbol{X}^T\boldsymbol{X} \tag{24}$$

- We set the first derivative to zero as $X^T X$ is positive definite matrix
- We get,

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\theta) = 0 \tag{25}$$

- Parameter estimates are

$$\hat{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{26}$$

Summary for Multivariate regression

- We have ,

$$Y = \theta_0 + \sum_{k=1}^{m} \theta_k X_k + \varepsilon \tag{27}$$

- ε is Gaussian random variable with zero mean and variance as σ^2
- We can express parameter estimates as

$$\hat{\theta} = \frac{\langle x, y \rangle}{\langle x, x \rangle} \tag{28}$$

Chapter 4
Logistic Regression

Introduction

- Linear regression is cool for linearly separable variables
- Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).
- Many problems require a probability estimate as output.
- Logistic regression is an extremely efficient mechanism for calculating probabilities.

Contd...

- In many cases, you'll map the logistic regression output into the solution to a binary classification problem, in which the goal is to correctly predict one of two possible labels (e.g., "spam" or "not spam")
- Classification can be followed using a discriminant based approach
- Task is to discriminate between multiple classes based on a given input

Contd...

- The linear model[6] is easy to understand : the final output is a weighted sum of the input attributes x_j.
- The magnitude of the weight θ_j shows the importance of x_j and its sign indicates if the effect is positive or negative.
- A generalized linear discriminant model can be written as

$$g(\boldsymbol{X}|\boldsymbol{\theta}) = \boldsymbol{X}^T\boldsymbol{\theta}, \tag{29}$$

where $\boldsymbol{\theta}$is a $(m+1) \times 1$ matrix and $\boldsymbol{X}$ is a $(m+1) \times n$ matrix

6. n indicates number of samples while m is number of predictor variables

Logistic Regression as Linear Discriminant

- For modeling binary classification, Bayes rule says that

$$P(C_1|x) = \frac{P(x|C_1) \times P(C_1)}{P(x)} \tag{30}$$

- Let $y = P(C_1|x)$ and $P(C_2|x) = 1 - y$, then as per classification task, we may say that

$$\text{choose } C_1 \text{ if } \begin{cases} y > 0.5 \\ \frac{y}{1-y} > 1 \\ \log \frac{y}{1-y} > 0 \end{cases} \text{ and } C_2 \text{ otherwise} \tag{31}$$

Contd...

- $\log\left(y/(1-y)\right)$ is known as logit transformation or log odds of y
- This can also be written in terms of prior probability of classes C_1 and C_2 as

$$\text{logit}\left(P\left(C_1 \mid x\right)\right) = \log \frac{P\left(C_1 \mid x\right)}{1 - P\left(C_1 \mid x\right)} = \log \frac{P\left(C_1 \mid x\right)}{P\left(C_2 \mid x\right)} \quad (32)$$

Contd...

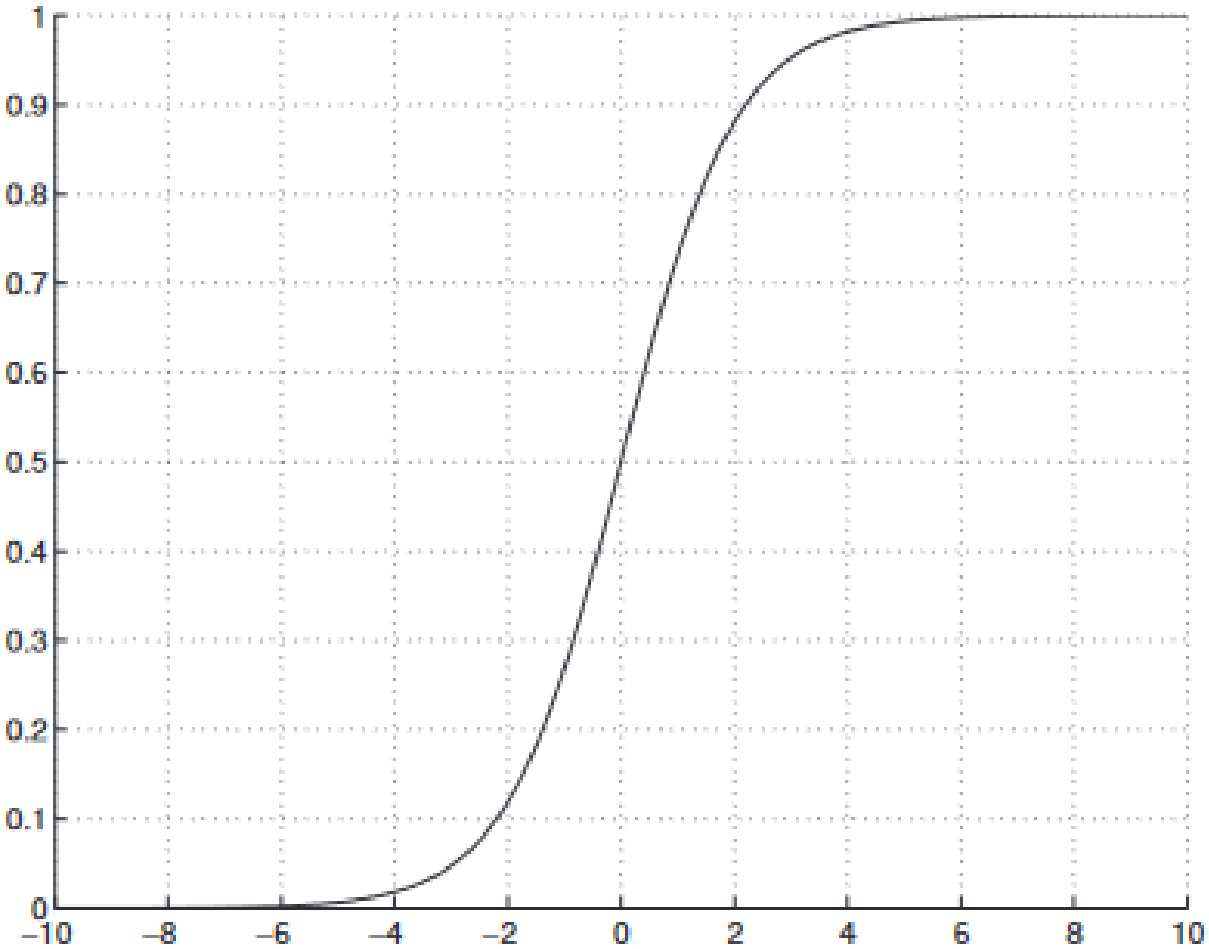
- Formally, the logistic regression model is given as

$$\log\left[\frac{p(C_1|x)}{1-p(C_1|x)}\right] = \theta_0 + \theta_1 x \tag{33}$$

- Solving for $p(x)$, we get,

$$p(x;\theta_1,\theta_0) = \frac{e^{\theta_0+x\cdot\theta_1}}{1+e^{\theta_0+x\cdot\theta_1}} = \frac{1}{1+e^{-(\theta_0+x\cdot\theta_1)}} \tag{34}$$

- Thus, when $\theta_0 + \theta_1 x$ is a positive value, $y = 1$ else $y = 0$
- Thus, a sigmoid function maps the input representation into a range [0,1]



- Hence, we can summarize for k th sample, the weighted sum $z_k = \sum_{i=1}^{m} \left(\theta_0 + \theta_i x_{ik}\right)$, then

$$sigmoid(z_k) = \frac{1}{1 + e^{-z_k}} \tag{35}$$

- Hence, given input x and model parameter θ, we can
  1. Calculate $g(\boldsymbol{x})$ and choose class C_1 if $g(\boldsymbol{x})>0$
  2. Calculate $y = sigmoid(\theta_0 + \boldsymbol{\theta^T x})$, choose class C_1 if $y > 0.5$

Maximum Likelihood Estimation for Logistic Regression

- Probabilistic framework for estimating the parameters of a model.
- In MLE, we choose parameters that maximize the conditional likelihood.
- The conditional likelihood $P(y|X, \theta)$ is the probability of observed values $y \in \mathrm{R}^n$.
- Given input $X = [x_1, x_2, \ldots, x_n] \in \mathrm{R}^{m \times n}$

Contd...

- Thus, we wish to maximize

$$P(\mathrm{y} \mid X, \theta) = \prod_{i=1}^{n} P\left(y_i \mid \mathrm{x}_i, \theta\right) \tag{36}$$

- Taking log both sides, we get,

$$\log\left(\prod_{i=1}^{n} P\left(y_i \mid \mathrm{x}_i, \theta\right)\right) = -\sum_{i=1}^{n} \log\left(1 + e^{-y_i \theta^T \mathrm{x}_i}\right)$$

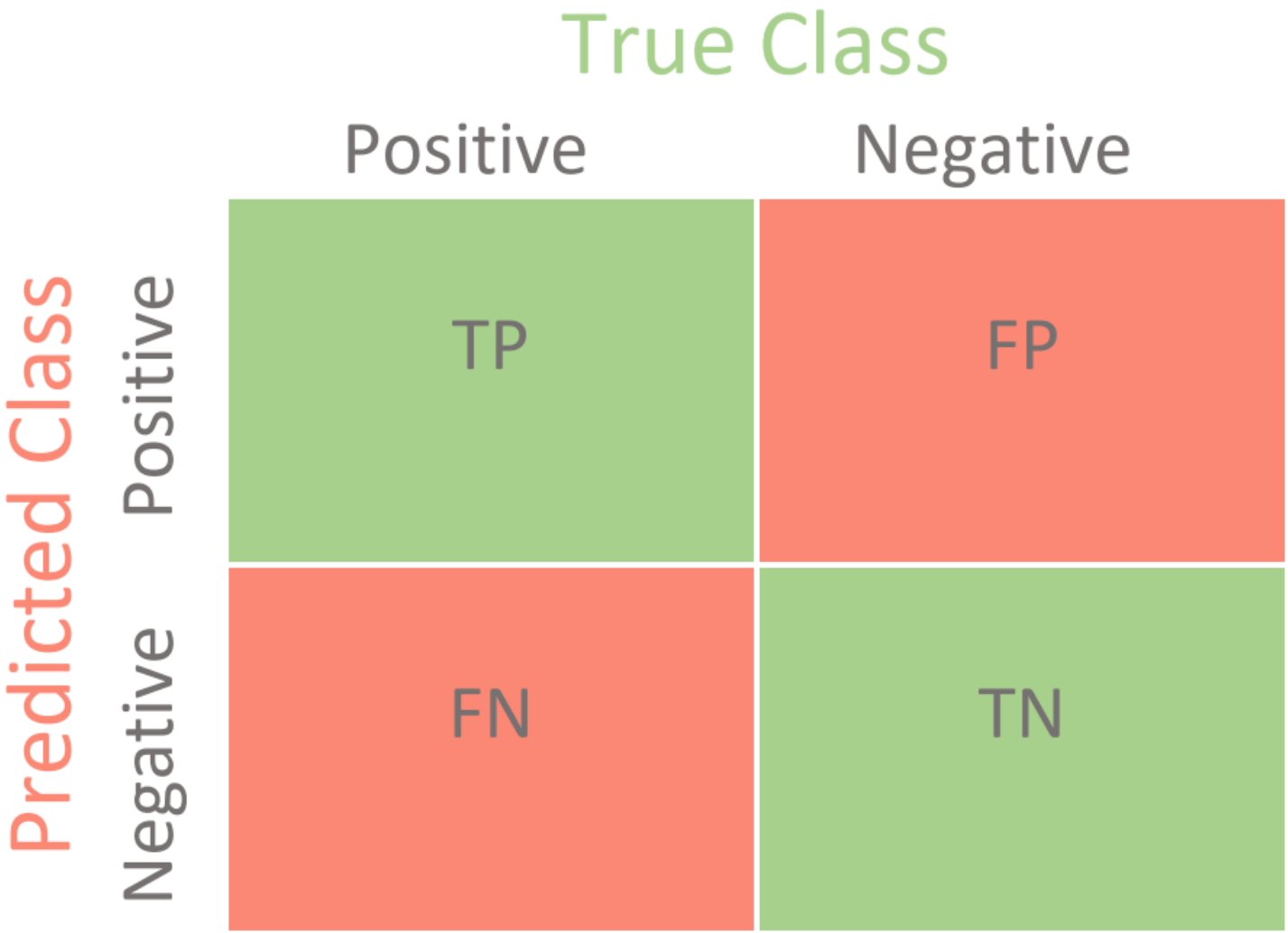
$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} -\sum_{i=1}^{n} \log\left(1 + e^{-y_i \theta^T \mathrm{x}_i}\right) \tag{37}$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \theta^T \mathrm{x}_i}\right)$$

Summary

- Logistic Regression is the discriminative counterpart to Bayes classifier.
- In Bayes classifier, we first model $P(x|y)$ for each label y, and then obtain the decision boundary that best discriminates between these two distributions.
- In Logistic Regression, we do not attempt to model the data distribution $P(x|y)$, instead, we model $P(y|x)$ directly.

Evaluating Classification Tasks



Confusion Matrix Terminologies

- True Positive (TP)
  (a)The predicted value matches the actual value
  (b)The actual value was positive and the model predicted a positive value
- True Negative (TN)
  (a)The predicted value matches the actual value (b)The actual value was negative and the model predicted a negative value

Contd...

- False Positive (FP) – Type 1 error
  (a) The predicted value was falsely predicted
  (b) The actual value was negative but the model predicted a positive value
  (c) Also known as the Type 1 error
- False Negative (FN) – Type 2 error
  (a) The predicted value was falsely predicted
  (b)The actual value was positive but the model predicted a negative value
  (c) Also known as the Type 2 error

Performance Metrics

- Accuracy

$$\frac{TP + TN}{TP + FP + FN + TN}$$

- Precision [7]

  tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not.

$$Precision = \frac{TP}{TP + FP} \tag{38}$$

7. What proportion of positive identifications was actually correct ?. **A model that produces no false positives has a precision of 1.0.**

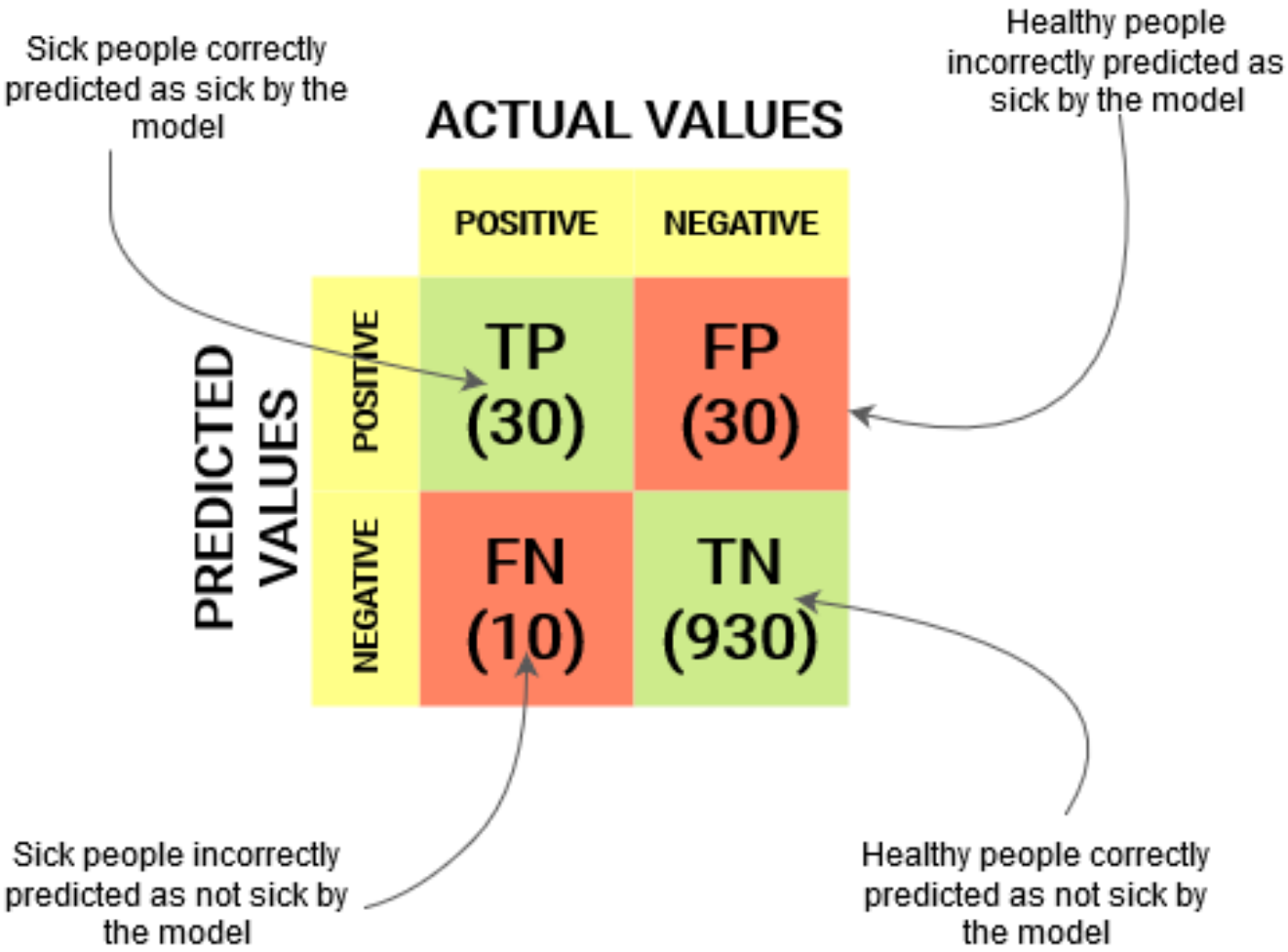
Recall

- Recall[8] tells us how many of the actual positive cases we were able to predict correctly with our model.
- Recall is also known as True Positive Rate (TPR) or hit rate

$$Recall = \frac{TP}{TP + FN} \tag{39}$$

- To fully evaluate the effectiveness of a model, you must examine both precision and recall. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

8. What proportion of actual positives was identified correctly?. A model that produces no false negatives has a recall of 1.0.


Sick people correctly predicted as sick by the model
Healthy people incorrectly predicted as sick by the model
ACTUAL VALUES
POSITIVE
NEGATIVE
PREDICTED VALUES
POSITIVE
NEGATIVE
TP (30)
FP (30)
FN (10)
TN (930)
Sick people incorrectly predicted as not sick by the model
Healthy people correctly predicted as not sick by the model

Contd...

- Classification metrics are as follows

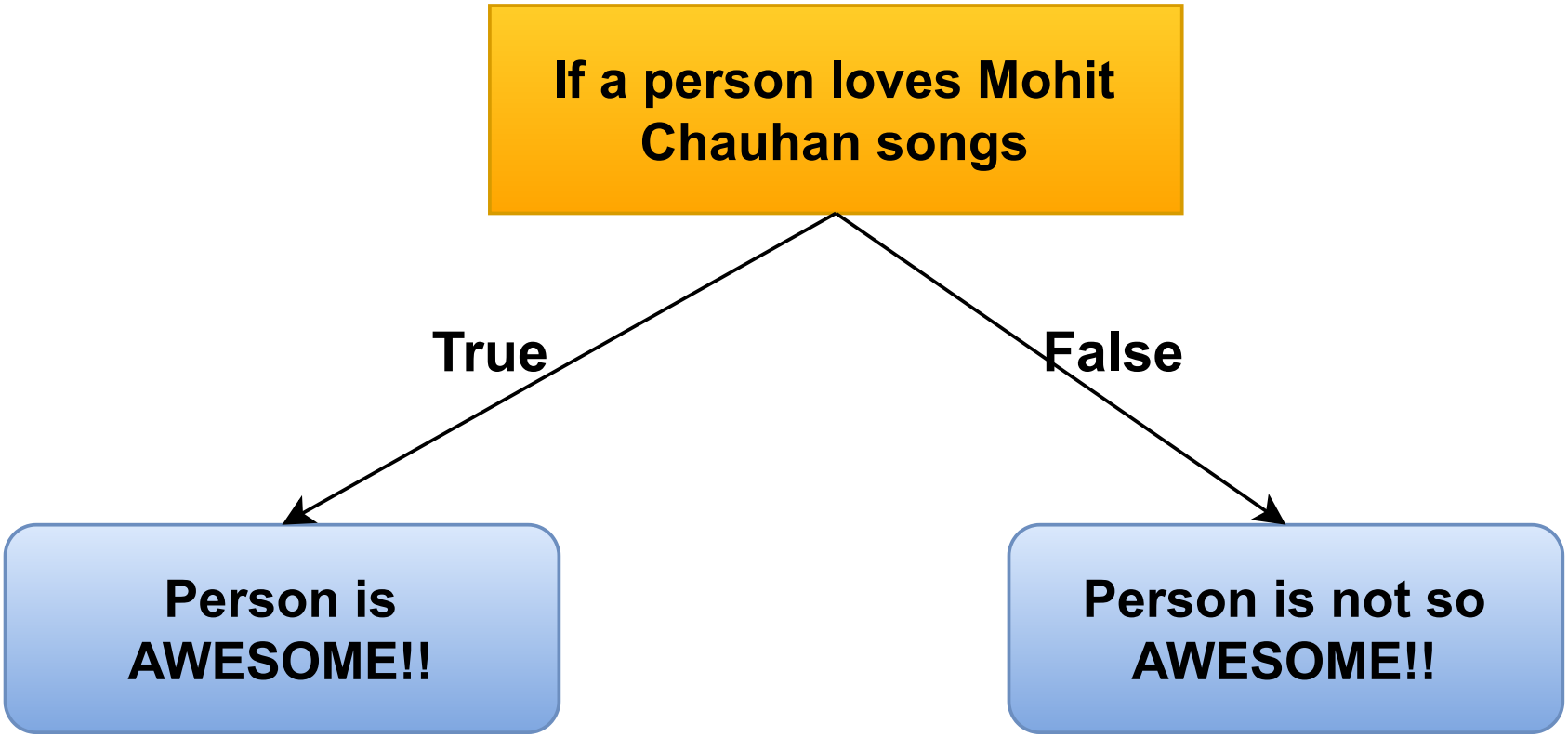
$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} = \frac{960}{1000} = 0.96 \\
Precision &= \frac{TP}{TP + FP} = \frac{30}{60} = 0.5 \\
Recall &= \frac{TP}{TP + FN} = \frac{30}{40} = 0.75
\end{aligned}
$$

Chapter-5
Decision Trees

Introduction

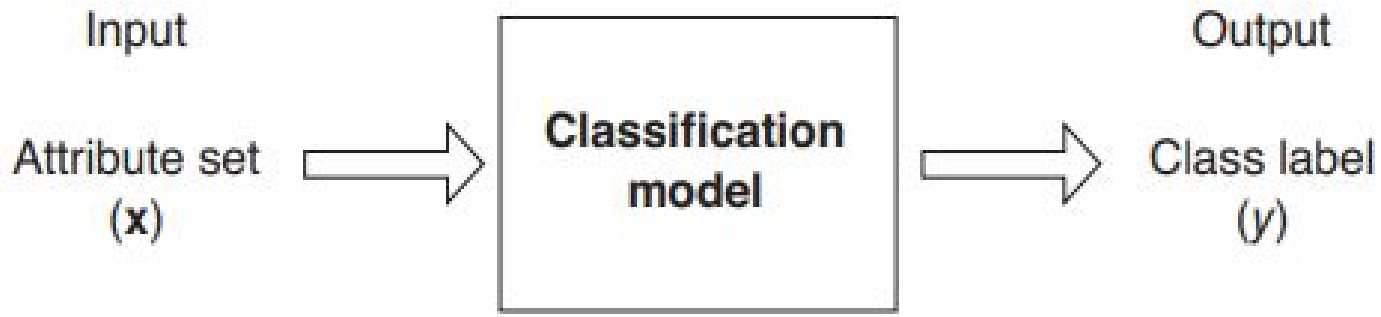
- Classification is a two-step process, learning step and prediction step, in machine learning.
- In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data.
- Decision Tree is one of the easiest and popular classification algorithms to understand and interpret.

Concept


If a person loves Mohit Chauhan songs
True
False
Person is AWESOME!!
Person is not so AWESOME!!

Contd...

- A method to approximate discrete-valued targets
- Majorly used for classification tasks
- A generic classification task looks like



Taxonomy

- Two types exist
  1. Categorical Variable Decision Tree : Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.
  2. Continuous Variable Decision Tree : Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

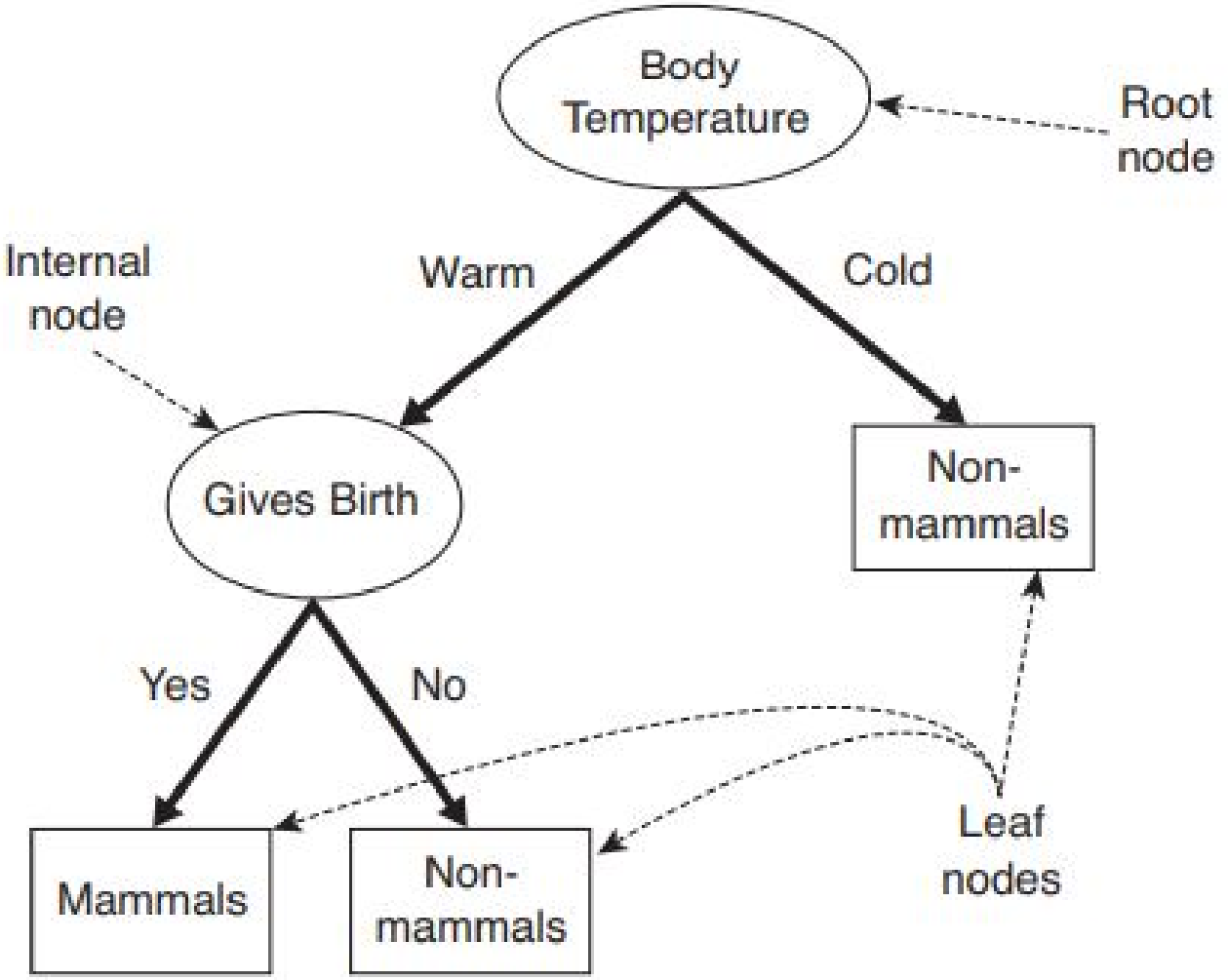
How a decision tree works ?

- Consider a task to classify a specie into mammal or a non-mammal
- One approach is to consider characteristics of the specie
- Either cold-blooded or warm-blooded ?
- It is possible to construct a hierarchical structure

Structure of a decision tree

- A root node has no incoming links and has zero or more outgoing links
- An internal node has one incoming link and has at least two or more outgoing links
- A Leaf/terminal node has exactly one incoming link and no outgoing link

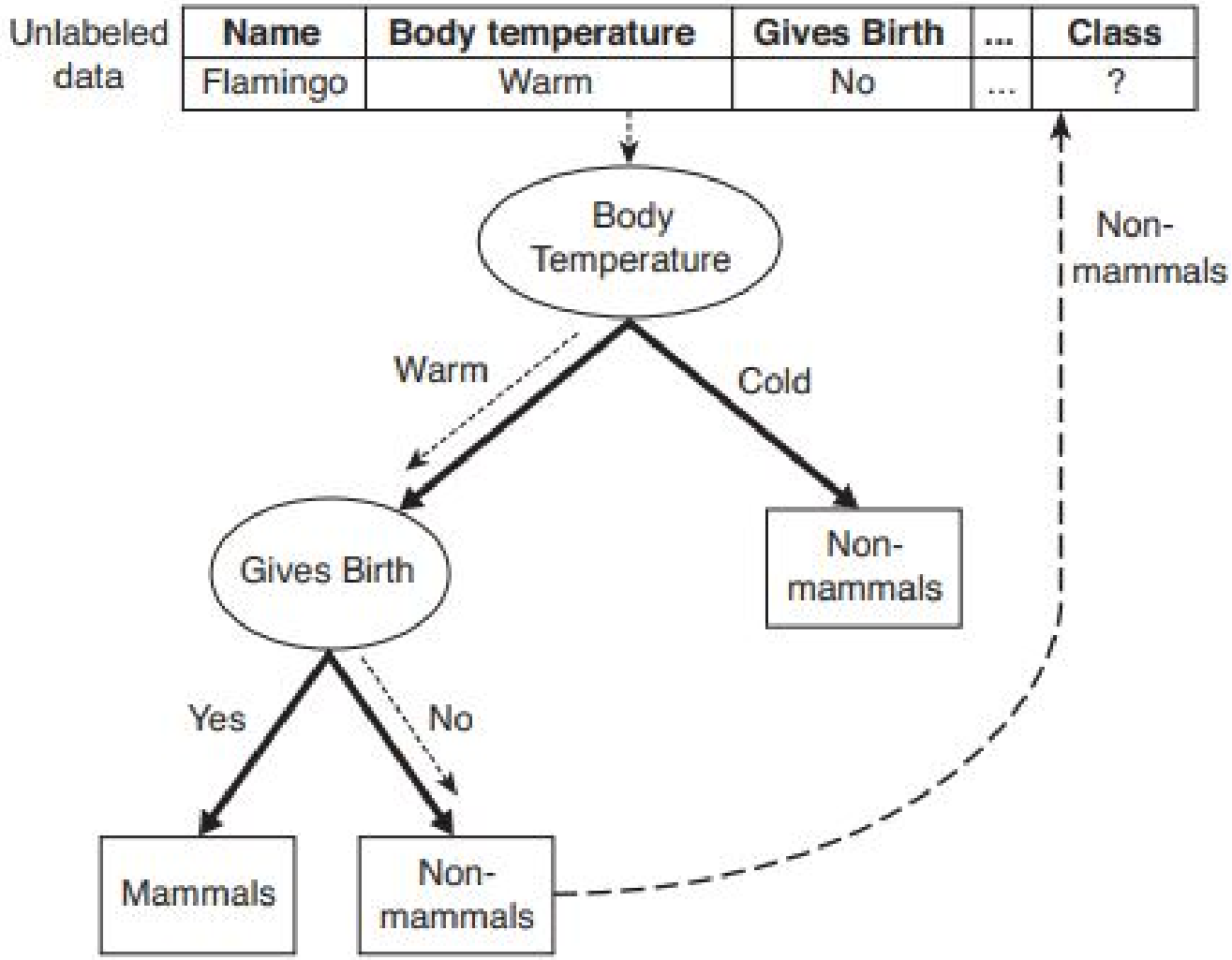
Example


Body Temperature
Root node
Internal node
Warm
Cold
Gives Birth
Non-mammals
Yes
No
Leaf nodes
Mammals
Non-mammals

Contd...

- Temperature as root node to distinguish between cold-blooded and warm-blooded
- All cold-blooded are non-mammals, hence it forms a leaf node
- If warm-blooded gives birth, then it is a mammal else not

Test example



Entropy & Information Gain for a Decision Tree

- Example : Consider the following data and classify the same using decision tree

Sr No	Temperature	Outlook	Windy	Humidity	Play Soccer ?
1	Hot	Sunny	false	High	No
2	Cool	Rain	false	Normal	Yes
3	Hot	Overcast	false	Normal	Yes
4	Cool	Sunny	false	Normal	Yes
5	Cool	Rain	true	Normal	Yes

How to select a good attribute ?

- A good attribute prefers attributes that split the data so that each successor node is as pure as possible
- i.e., the distribution of examples in each node is so that it mostly contains examples of a single class
- In other words, we want an attribute with
  (a) Maximum order : All examples are of the same class
  (b) Minimum order : All classes are equally likely
- Decision tree(s) use ID3 algo

Entropy for two-class classification problem

- Entropy (S) is defined as measure of randomness in a training set.
- Consider p_+ to be proportion of positive samples and p_- to be the proportion of negative samples.
- Entropy (S) is then given as

$$E(S) = -p_+ \log_2 p_+ - p_- \log_2 p_- \tag{40}$$

- In general, entropy for n classes is given as

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 \cdots - p_n \log p_n = \sum_{i=1}^{n} -p_i \log p_i \tag{41}$$

Let us evaluate Entropy for the given example

- Consider the attribute **Outlook= Sunny** : 1 Yes and 1 No
- E(Outlook=Sunny)= $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$
- Consider **Outlook=Rain** : 1 Yes and 1 No
- E(Outlook=Rain)= $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$
- Consider **Outlook=Overcast** : 1 Yes and 0 No
- E(Outlook=Overcast)= $-1 \log_2 1 - 0 \log_2 0 = 0$

Contd...

- Consider attribute **Humidity= High/Normal**
- Entropy (Humidity|High)= 1 No and 0 Yes,

$$E = -1 \log_2 1 - 0 \log_2 0 = 0 \tag{42}$$

- Entropy (Humidity|Normal)= 0 No and 4 Yes,

$$E = -1 \log_2 1 - 0 \log_2 0 = 0 \tag{43}$$

Information Gain in Decision Trees

- Entropy only computes the quality of a single (sub-)set of examples
- When an attribute A splits the set S into subsets S_i
- Information Gain (IG) is given as

$$\text{Gain}(S, A) = E(S) - I(S, A) = E(S) - \sum_i \frac{|S_i|}{|S|} \cdot E(S_i) \quad (44)$$

Let us evaluate IG of each attribute A_i

- Consider attribute A_1, that is, Outlook
- We have a total of 5 samples (4 Yes, 1 No)
- We represent Yes with '+' class and No with '-' class
- Overall entropy is given as

$$E(S) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} \quad (45)$$

$$= 0.72192 \quad (46)$$

Contd...

- IG for outlook is given as Sunny(1+,1-), Rain(1+,1-) and Overcast(1+,0-)

$$= E(S) - \frac{S_{A_1}}{S} E(A_1) - \frac{S_{A_2}}{S} E(A_2) - \frac{S_{A_3}}{S} E(A_3) \quad (47)$$

$$= 0.72192 - \frac{2}{5}(1) - \frac{2}{5}(1) - \frac{1}{5}(0) \quad (48)$$

$$= -0.07808 \quad (49)$$

Contd...

- Similarly, we can evaluate the IG for attributes Wind, Temperature and Humidity

Chapter-6
Evaluating Machine Learning Models

Introduction

- A prediction accuracy metric to report on the overall success of the model
- Visualizations to help explore the accuracy of your model beyond the prediction accuracy metric
- The ability to review the impact of setting a score threshold (only for binary classification)
- Alerts on criteria to check the validity of the evaluation

Evaluating Binary Classification Model

- ▸ Confusion Matrix
- The false positive rate (FPR) measures the false alarm rate or the fraction of actual negatives that are predicted as positive.
- The range is 0 to 1.
- A smaller value indicates better predictive accuracy
- This is given as

$$FPR = \frac{FP}{FP + TN} \tag{50}$$

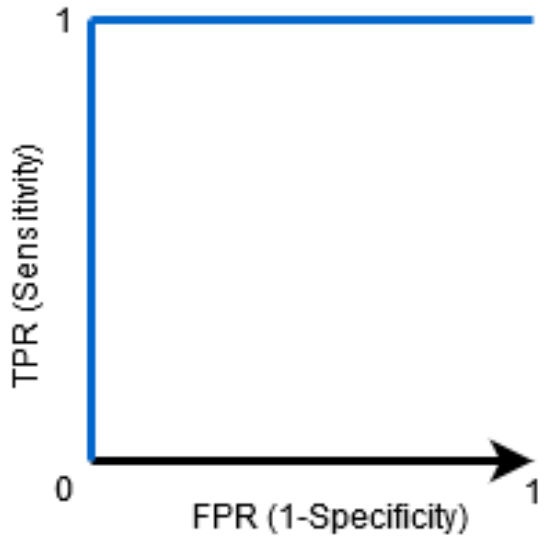
AUC-ROC Curve for Binary Classification

- The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems.
- It is a probability curve that plots the TPR [9] against FPR [10] at various threshold values and essentially separates the 'signal' from the 'noise'.
- The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

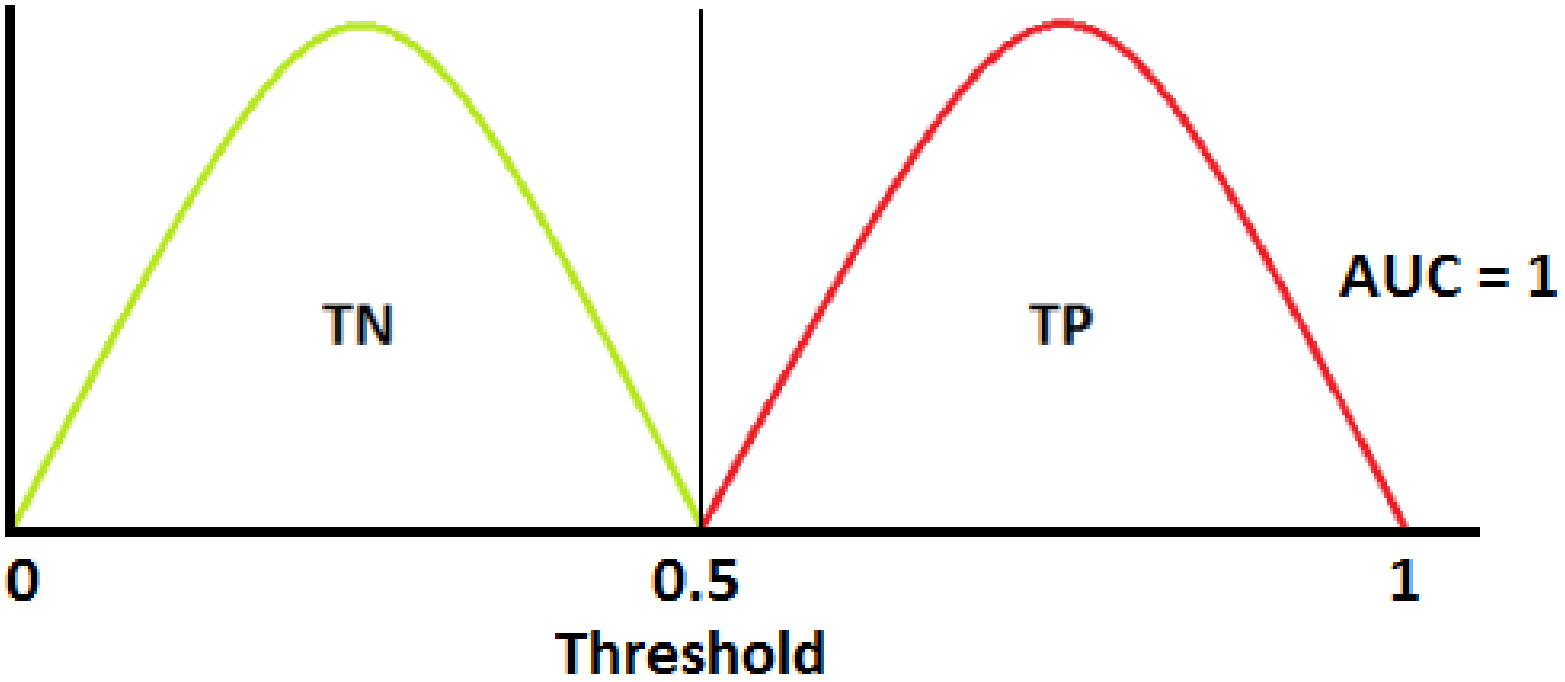
9. True positive rate
10. False positive rate

Contd...

- When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly.
- If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.

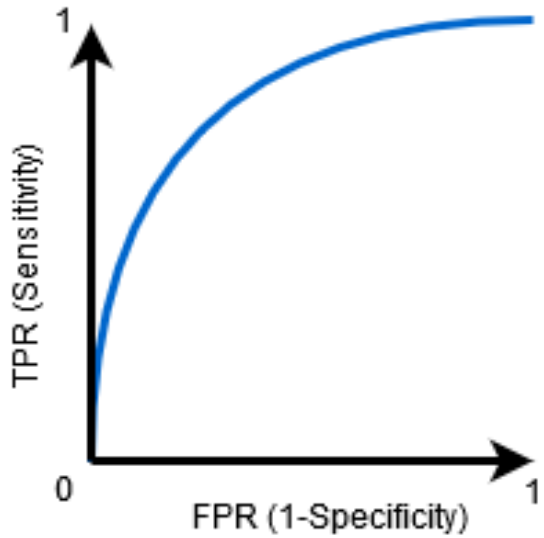


Classification with a threshold



Contd...

- When 0.5 <AUC< 1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.
- This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.



Crux for AUC-ROC curve

So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

Evaluating Regression Models

- The output of a regression ML model is a numeric value for the model's prediction of the target.
- For example, if you are predicting housing prices, the prediction of the model could be a value such as 254,013.
- Consider actual response to be y_i and predicted response as $\hat{y}_i$, the mean absolute error is given as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{51}$$

Contd...

- The mean squared error (MSE) is given as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{52}$$

- A popular metric root mean squared error (RMSE) is given as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{53}$$

Contd...

- Mean absolute percentage error (MAPE) is given as

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{54}$$

- Coefficient of regression (R^2) is given as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{55}$$

Importance of R^2

- R^2 is a statistic that will give some information about the goodness of fit of a model.
- In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.
- An R^2 of 1 indicates that the regression predictions perfectly fit the data.

Improving regression quality via Quantiles

- RMSE may be the most common metric, but it has some problems. Most crucially, because it is an average, it is sensitive to large outliers.
- If the regressor performs really badly on a single data point, the average error could be very big. In statistical terms, we say that the mean is not robust (to large outliers).

What are Quantiles?

- Quantiles (or percentiles), on the other hand, are much more robust.
- To see why this is, let's take a look at the median (the 50th percentile), which is the element of a set that is larger than half of the set, and smaller than the other half.
- If the largest element of a set changes from 1 to 100, the mean should shift, but the median would not be affected at all.

Decision Tree : Regression

- Decision tree builds regression or classification 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.
- The final result is a tree with decision nodes and leaf nodes.

- The topmost decision node in a tree which corresponds to the best predictor called root node.
- Decision trees can handle both categorical and numerical data.
- Let us discuss the example of regression problem under decision tree.

Example : Predict the number of hours a player will play given the covariates

Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

Contd...

- Let us calculate the standard deviation in the response variable (y)
- Standard deviation is given as

$$SD = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} \tag{56}$$

- We get $SD = 9.32$ and coefficient of variation (CV) as

$$CV = \frac{SD}{\bar{y}} = 23\% \tag{57}$$

The root node evaluation

- Given attribute X_i, the standard deviation for this attribute is

$$SD(Y|X_i) = \sum_{x \in X} P(x)S(x), \tag{58}$$

  where $P(x)$ is the probability of sub-class $x_i \in X$ and $S(x)$ is the standard deviation for the sub-class x_i

- Compute reduction in standard deviation as

$$SDR(X_i) = S(Y) - SD(Y|X_i) \tag{59}$$

Chapter
K-Nearest Neighbors

Introduction

1. K-Nearest Neighbour (KNN) is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
2. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
3. KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

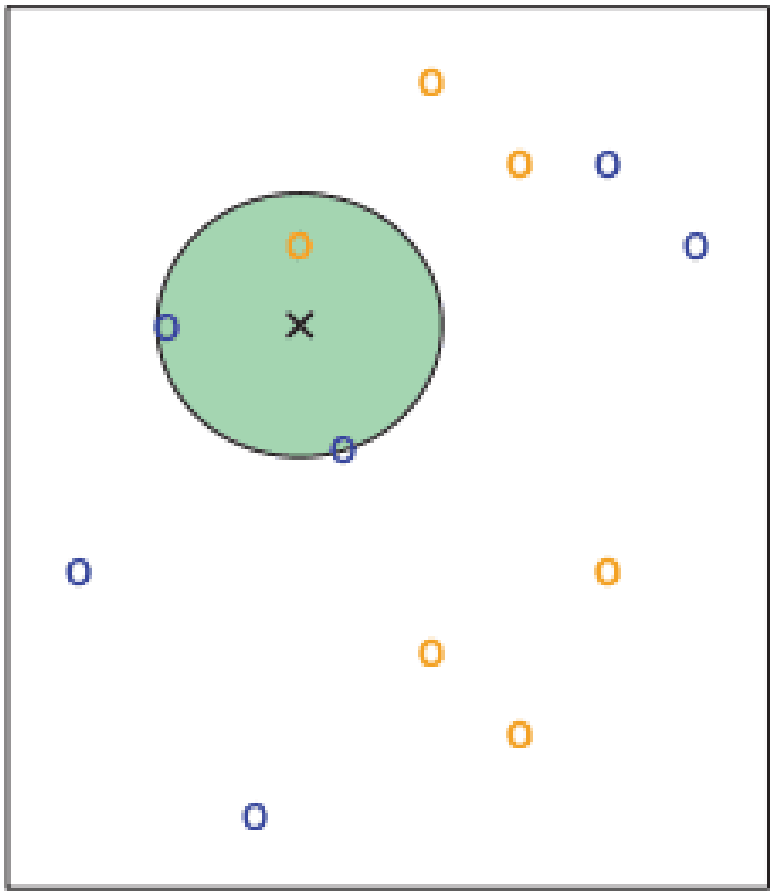
KNN for Classification Setup

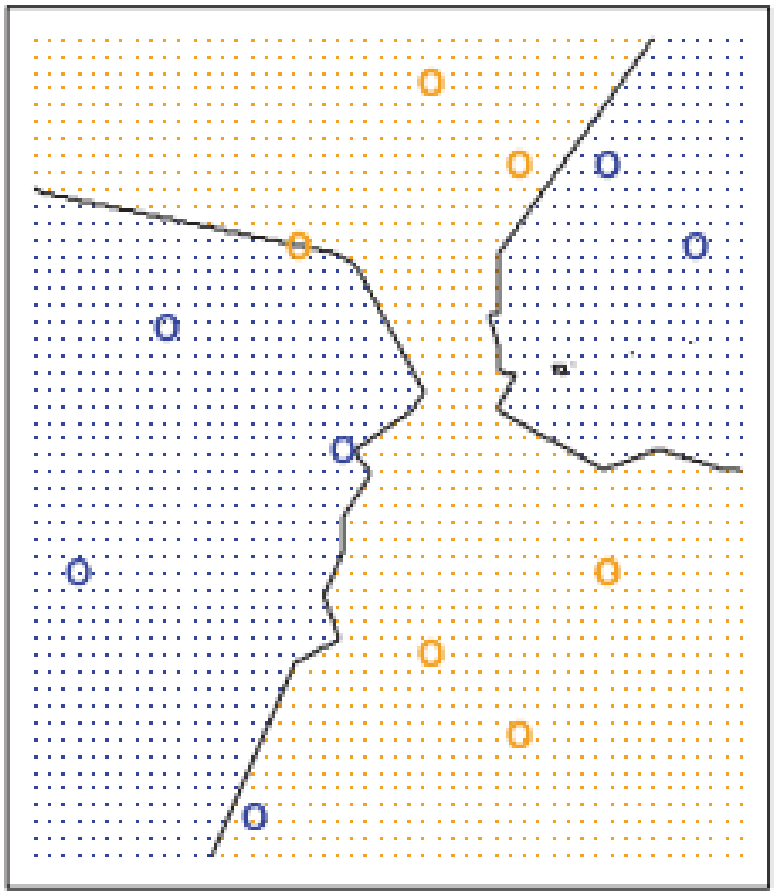
1. A classification problem has a discrete value as its output.
2. For example, "likes pineapple on pizza" and "does not like pineapple on pizza" are discrete. There is no middle ground.
3. The analogy above of teaching a child to identify a pig is another example of a classification problem.

KNN in practice

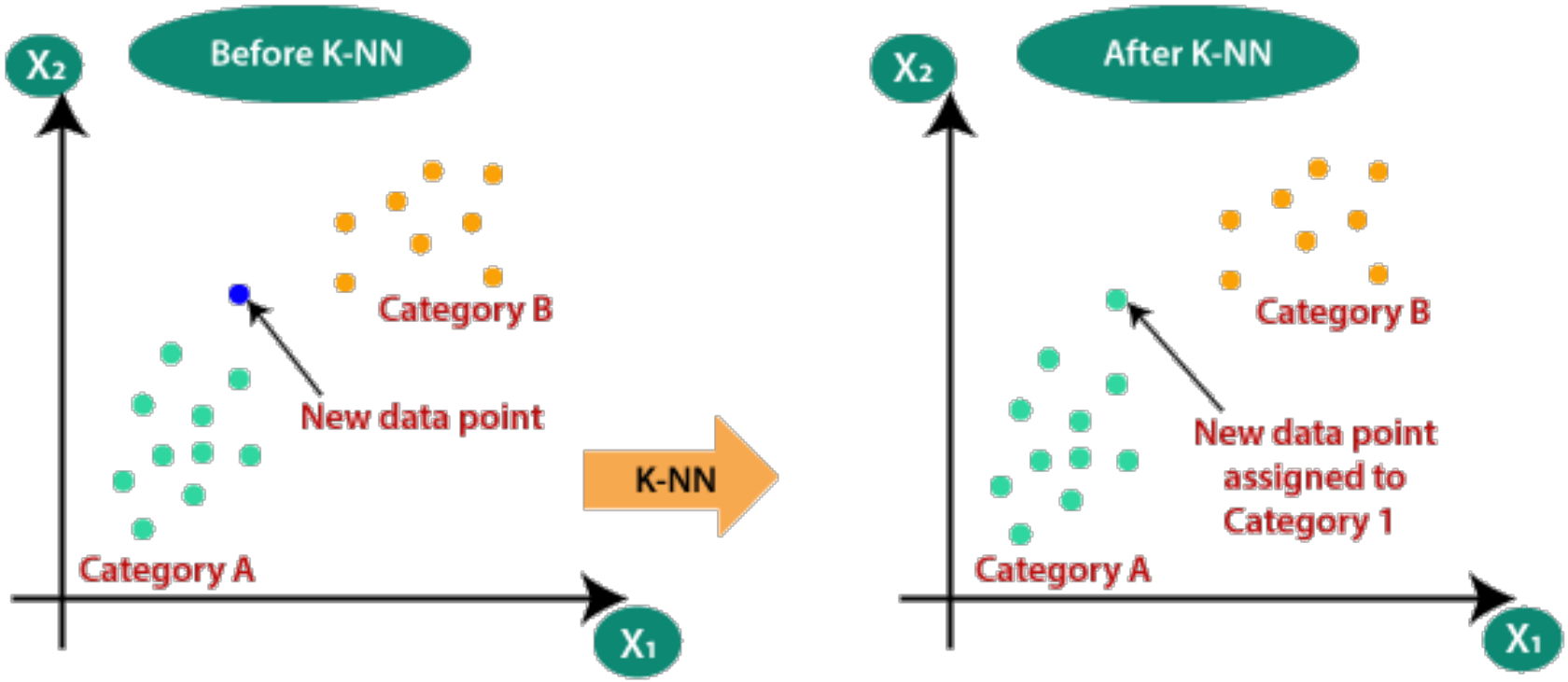
1. At scale, this would look like recommending products on Amazon, articles on Medium, movies on Netflix, or videos on YouTube.
2. Although, we can be certain they all use more efficient means of making recommendations due to the enormous volume of data they process.

KNN Visualized





Contd...


Before K-NN
X2
Category B
New data point
Category A
X1
K-NN
After K-NN
X2
Category B
New data point
assigned to
Category 1
Category A
X1

How KNN works ?

1. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
2. It first identifies the k points in the training data that are closest to the test value and calculates the distance between all those categories.
3. The test value will belong to the category whose distance is the least.

How do we find similarity in training data ?

1. Minkowski Distance – It is a metric intended for real-valued vector spaces.
2. We can calculate Minkowski distance only in a normed vector space, which means in a space where distances can be represented as a vector that has a length and the lengths cannot be negative.
3. Mathematically,

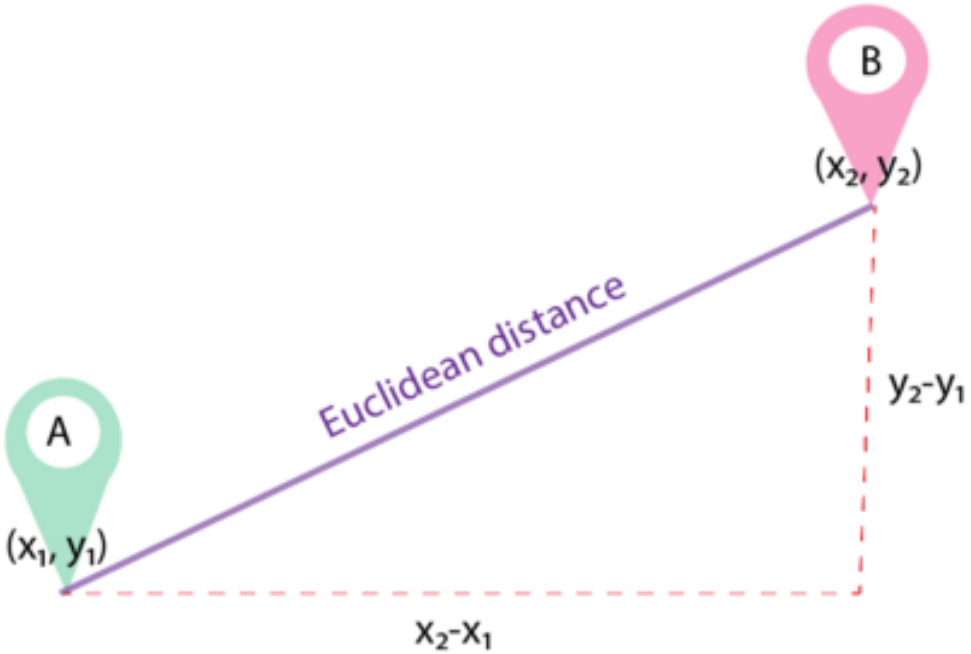
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{60}$$

Manhattan Distance

1. We use Manhattan Distance if we need to calculate the distance between two data points in a grid like path.
2. As mentioned above, we use Minkowski distance formula to find Manhattan distance by setting p's value as 1.
3. The distance between two points measured along axes at right angles. In a plane with $p1$ at (x_1, y_1) and $p2$ at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$
4. Thus,

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{61}$$

Euclidean distance



Contd...

1. For each dimension, we subtract one point's value from the other's to get the length of that "side" of the triangle in that dimension, square it, and add it to our running total.
2. The square root of that running total is our Euclidean distance.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{62}$$

Case Study : KNN for Prediction

Funny but true

New Chapter whose title is ? ? ? ? ?

Support Vector Machines

Introduction

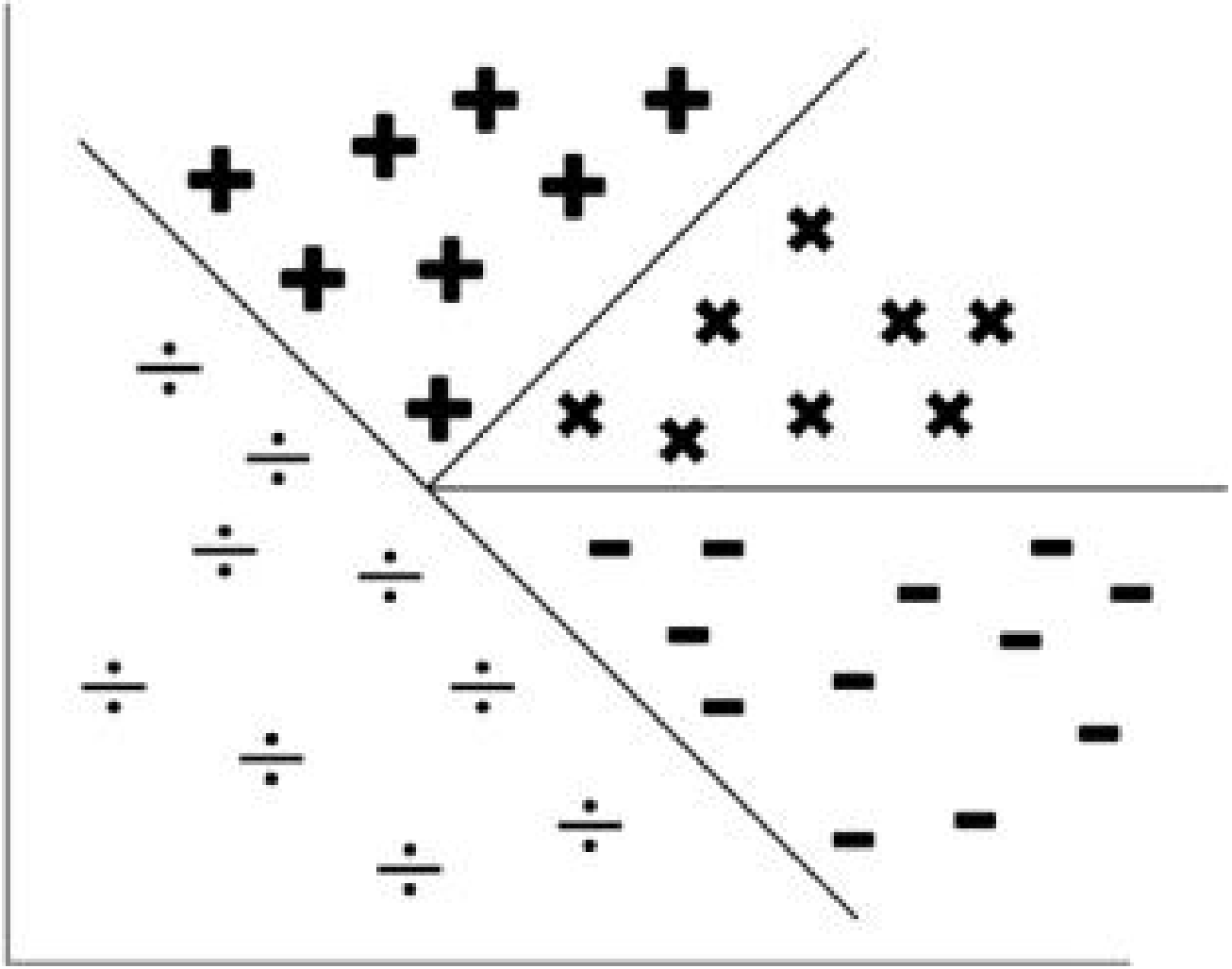
1. Here we approach the two-class classification problem in a direct way
2. We try and find a plane that separates the classes in feature space
3. Thus a hyperplane that separates “n" classes is what we need

What is a hyperplane?

1. A line that separates different classes in a multi-dimensional space
2. A generic hyperplane can be given as

$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n \tag{63}$$

Contd...

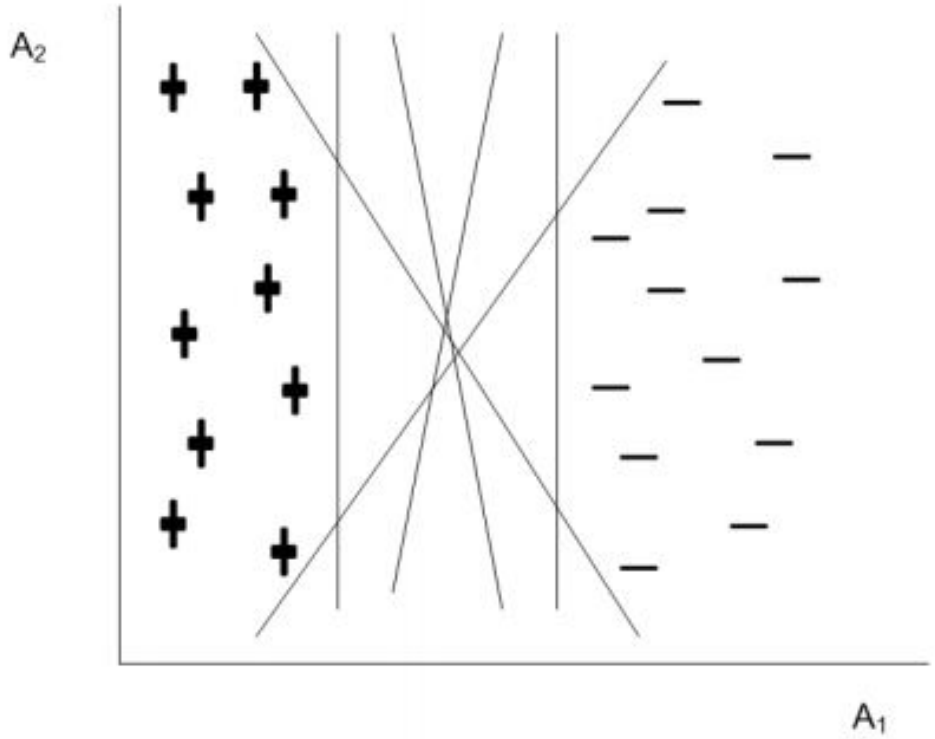


Let's define SVM formally

1. Ideology by Vladimir Vapnik
2. As a task of classification, it searches for optimal hyperplane(i.e., decision boundary) separating the tuples of one class from another.
3. SVM works well with higher dimensional data and thus avoids dimensionality problem.

Margin in SVM

1. Consider training data $y = (y_1, y_2, \ldots, y_n)$ belonging to either + or - class
2. Consider two attributes/features/inputs/predictors as A_1 and



A_2

Contd...

1. Figure shows a plot of data in 2-D.
2. Another simplistic assumption here is that the data is linearly separable, that is, we can find a hyperplane (in this case, it is a straight line) such that all +'s reside on one side whereas all -'s reside on other side of the hyperplane
3. However, we want to address the following ?
   - Whether all hyperplanes are equivalent so far the classification of data is concerned ?
   - If not, which hyperplane is the best ?

Contd...

1. We may note that so far the classification error is concerned (with training data), all of them are with zero error.
2. However, there is no guarantee that all hyperplanes perform equally well on unseen (i.e., test) data.
3. Thus, for a good classifier it must choose one of the infinite number of hyperplanes, so that it performs better not only on training data but as well as test data.

Contd...

1. Two hyperplanes H_1 and H_2 have their own boundaries called decision boundaries(denoted as b_{11} and b_{12} for H_1 and b_{21} and b_{22} for H_2).
2. A decision boundary is a boundary which is parallel to hyperplane and touches the closest class in one side of the hyperplane.
3. The distance between the two decision boundaries of a hyperplane is called the margin. So, if data is classified using Hyperplane H_1, then it is with larger margin then using Hyperplane H_2.
4. The margin of hyperplane implies the error in classifier. In other words, the larger the margin, lower is the classification error.

SVM for Linear Data

1. A SVM which is used to classify data which are linearly separable is called linear SVM.
2. In other words, a linear SVM searches for a hyperplane with the maximum margin.
3. This is why a linear SVM is often termed as a maximal margin classifier (MMC)

Chapter-2
Neural Networks

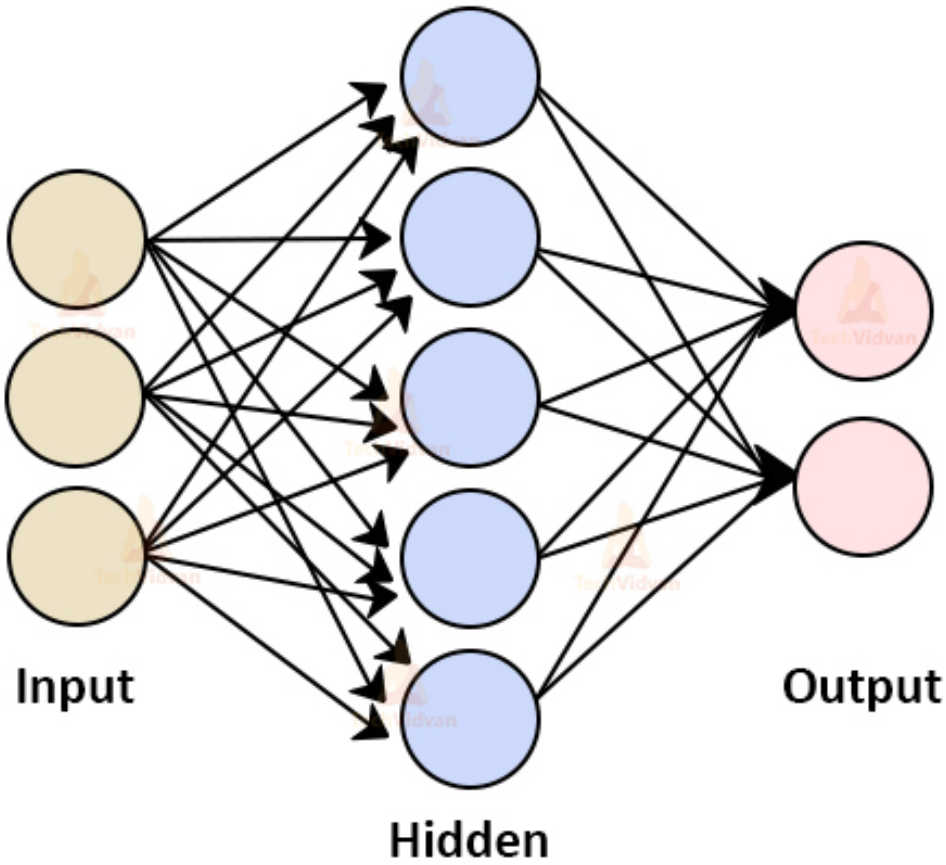
Introduction

1. Machine behaves like human brain
2. Brain is responsible for information flow from one point to another
3. Mimic behavior of human brain for modeling scientific problems
4. Human brain has the ability to perform tasks such as pattern recognition, perception and motor control much faster than any computer

The science behind human brain

- Brain has neurons that carry information.
- Neurons work together to form an incredible processing unit.
-

Architecture of Artificial Neural Network



Activation functions for NNs

Perceptron Learning

- The perceptron is the simplest form of a neural network used for the classification of patterns said to be linearly separable
- Basically, it consists of a single neuron with adjustable synaptic weights and bias
- Consider input vector $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ and output $\boldsymbol{y}$

Contd..

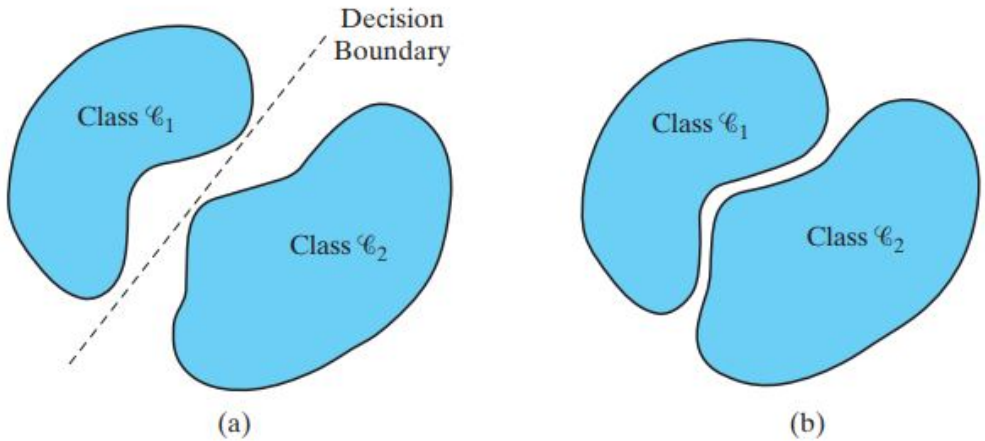
- Consider n inputs with weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ with external bias b
- The goal of the perceptron is to classify externally applied stimuli $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ into classes C_1 or C_2
- The decision rule for the classification is to assign the point represented by the inputs $(x_1, x_2, \ldots, x_n)$ to C_1 if perceptron output y is $+1$
- Similarly, the input is assigned to class C_2 if perceptron output is -1

Perceptron learning as hyperplane

- Input to the activation function is given as

$$v = \sum_{i=1}^{n} w_i x_i + b \tag{64}$$

- Thus, the output is given as $\hat{y} = sgn(v)$
- Decision boundary must be linearly separable



Contd...

- The activation function for linearly separable problems is given as

$$\text{sgn}(v) = \begin{cases} +1 & \text{if } v > 0 \\ -1 & \text{if } v < 0 \end{cases} \tag{65}$$

Multi-layer Perceptron Neural Networks

1. Perceptron learning is often linear
2. One approach to doing this is to chain together a collection of perceptrons to build more complex neural networks.