Machine Learning

Dr. Teena Sharma

Ph.D., University of Quebec at Chicoutimi, Canada

(tsharma@etu.uqac.ca)

Summer 2024

Week	Topic	Required Reading
2	Introduction	Chap. 1
3	Data understanding	Chap. 2 and 3
4	Visualization	
5	Data preparation	Chap. 7
6	Classification	Chap. 18, 19, 21 and 22
7	Regression	
8	Clustering	Chap. 13 and 17
9	Big Data (Processing and Storage)	
10	Deep Learning	
11	Deep Learning	
12		

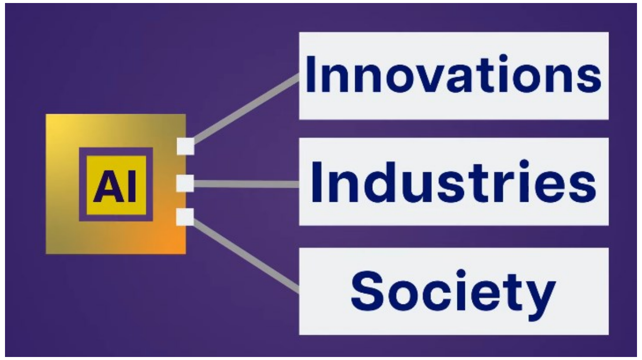
Expert Talk –foreign, India

1. Prof. Rajasen Gupta (Professor Mcgill University, Montreal Canada)
2. Prof Abdellah Chehri (Royal Military College, Kingston, Canada)
3. Prof Issouf Fofana (University of Quebec at Chicoutimi, Quebec Canada)
4. Dr. Benoit Duglas (Thales, Canada)
5. Dr. Roshan Jain (Startup in AI, Waterloo, Ontario, Canada)
6. Mr. Abhishek (Manager, software developer, Accenture, Quebec, Canada)
7. Ms. Dhavni Sharma (Working at International air transport association (IITA, Montreal, Canada)
8. Mrs. Aakansha Chawla, MBA (Business analyst, IITA, Montreal, Canada)
9. Prof. Hitesh Upreti (Professor Shivnadar University, Greater Noida)

AI fundamentals

Artificial Intelligence: Simulation of human intelligence in machines enabling them to perform tasks typically require human thinking. Chat bot ELISA (developed in mid 1960s and could mimic human like conversation to an extent). It's a very broad terms encompassing several techniques.

AI's ability to learn and adapt has the potential to transform entire industries, create new innovations and ultimately benefit society as a whole.



.

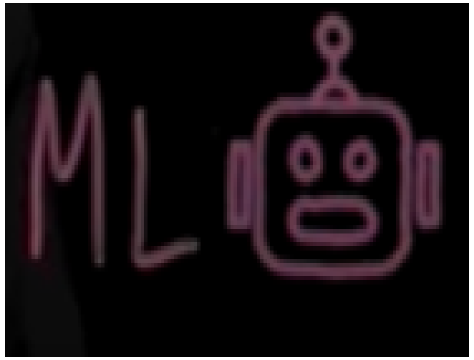
Cont.

Machine Learning: A subfield of AI, focussing on developing algorithms that allow computers to learn from and make decisions based upon data rather than being explicitly programed to perform a specific task.

These algorithms uses statistical techniques to learn patterns in data and make predictions or decisions without human intervention. It's again a broad terms and uses traditional statistical methods and complex neural networks (Categories: SL, UL, RL).

Or

ML is subset within AI, more focussed on the use of various self-learning algorithms that derive knowledge from data in order to predict outcomes



Cont.

Deep Learning: Artificial neural networks with multiple layers (nodes and connections). ML can extract simpler patterns in data while DL excels at handling vast amounts of big data (unstructured data) like images or natural language.

Foundation models: popularized in 2021 by researchers at the Stanford institute and provide more generalized and scalable AI solutions. These models are large scale neural networks trained on vast amount of data and they serve as a base for a multitude of applications. So, instead of training a model from scratch for each specific task, you can take a pretrained foundation model and fine tune it for a particular application (save both resources and time). Can perform task ranging from language translation to content generation to image recognition.

Cont.

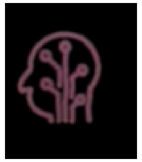
Large Language Model: Process and generate humanlike text. L stands for large scale (billions or millions of parameters). Next, L stands for language that designed to understand and interact using human languages as they are trained on massive datasets. LLMs can grasp grammar, context, idioms and even cultural references. Last letter stands for computational model, a series of algorithms and parameters working together to process input and produce output. It can handle a broad spectrum of task such as answering questions, translating or even creative writing.
Vision model: It can see in and quotes, interpret and generate images.
Scientific models: are used in biology where there are models for predicting how proteins fold into 3D shape.
Audio model: for generating human sounding, speech or composing the next fake drake hit song.

Generative AI: Models and algorithms specifically crafted to generate new content. Foundation models provide the underlying structure and understanding, GI is about harnessing that knowledge to produce something that is new. It's a broad field of AI that uses algorithms to create new content like text, images, videos, audio, code, and simulations.
Example?

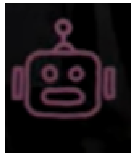
AI and Augment AI

AI: is the ability for leveraging computers or machines to mimic the problem solving and decision-making capabilities of human mind. It can perform task and make decisions that normally require human intelligence, such as reasoning, natural communication and problem solving. Basically, replaces the need of humans.

Augmented intelligence: m/c and humans both work together by enhancing each other's efforts when completing tasks. It augment human abilities, such as screen reader for blind, voice navigation or in-car collision avoidance system or blind spot detection system. They complement our own capabilities.
So, AI or Augment AI?



+



Reinforcement learning in AI

Reinforcement learning in AI is when machines learn to make better decisions by trying things out and getting feedback. For example, it can be used to teach a robot how to navigate in a room. When robot perform an action, such as stopping, turning around or moving forward, it then receives a reward or penalty based on how well it did. The robot uses this feedback to learn and improve its decision-making abilities and over time it gets better at navigating in the room.

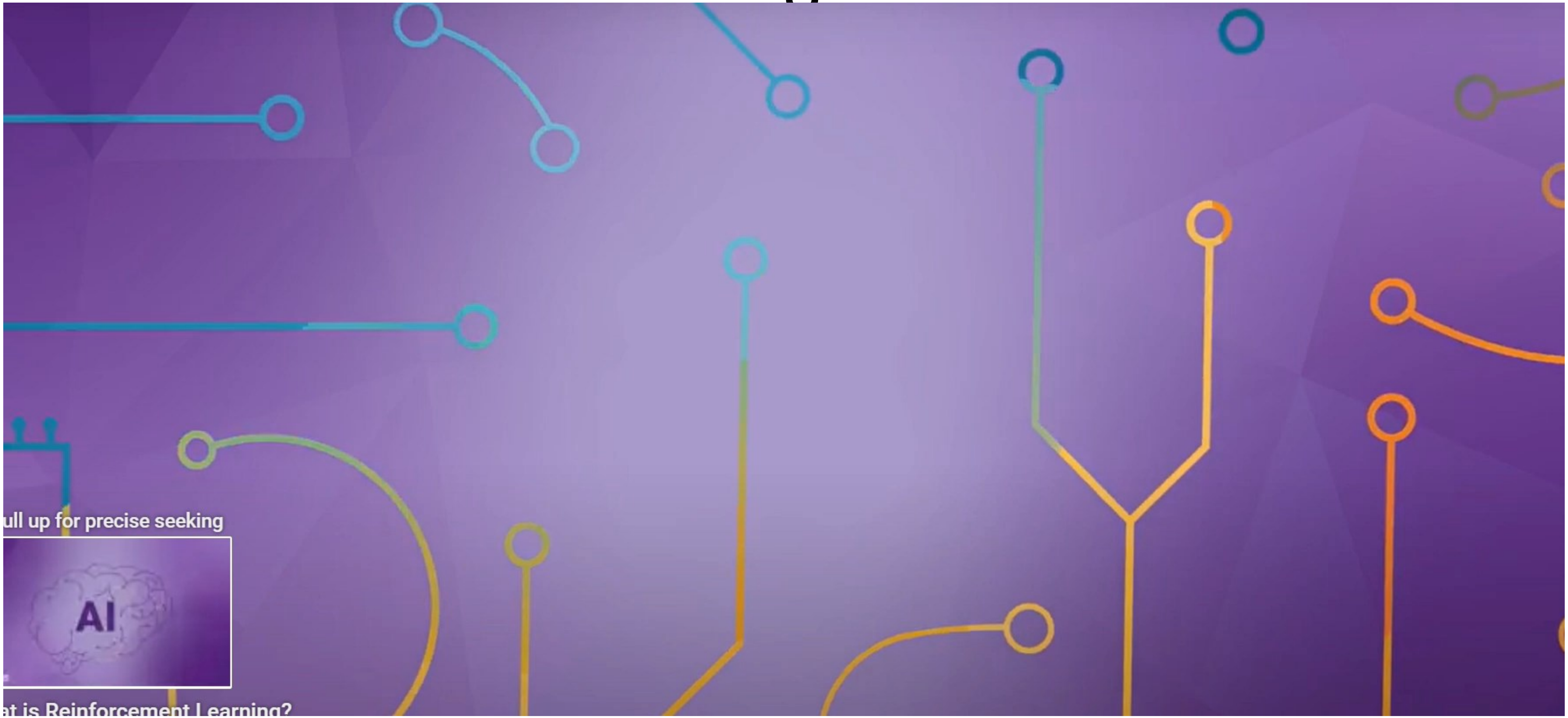
Use cases: Robotics, gaming, autonomous vehicles and recommendation systems use reinforcement learning to improve performance. The ability to learn from mistakes and get better over time makes reinforcement learning a critical tool in AI.

Reinforcement learning

Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results. It mimics the trial-and-error learning process that humans use to achieve their goals, through a feedback system, the agent learns from its environment and optimizes its behaviors.

During training, model perceive and interpret its environment, take actions and learn through trial and error. E.g., such as a feature in a video game or a robot in an industrial setting and recommendation systems.

Reinforcement Learning



Data Analytics and Data Science

Data analytics involves examining data to extract meaningful insights, while data science encompasses a wider scope, including data collection, cleaning, analysis, and machine learning modeling for predictive insights and decision-making.

Data analytics focuses more on analyzing the past data or historical data (explaining the past) to predict or forecast future, outcome or decision making. E.g., Amazon product sale or temperature prediction.

Introduction to Data Mining and Analysis

Machine Learning
Summer 2024

Data Mining

- Data mining is the process of discovering *insightful*, *interesting*, and *novel patterns*, as well as deriving descriptive, understandable, and predictive models from large-scale data.

- At the heart of data mining is *data itself*.

- We begin this course by looking at basic properties of data modeled as a data matrix

Data Matrix

- Data can often be represented or abstracted as an *n x d* data matrix, with n rows and d columns, where rows correspond to entities in the dataset, and columns represent attributes or properties of interest

$$\boldsymbol{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \boldsymbol{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \boldsymbol{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

Data Matrix

- **Rows:** Also called instances, examples, records, transactions, objects, points, feature-vectors, etc. Given as a d-tuple

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

- **Columns**: Also called attributes, properties, features, dimensions, variables, fields, etc. Given as an n-tuple

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Attribute Classification

Discrete Attribute
Has a finite or countably set of values
Examples: Zip codes, click counts, set of words in a collection of documents (often represented as integer values)
Binary attribute is a special case of discrete attribute

Continuous Attribute
Has real numbers as attribute values
Examples: temperature, height, or weight
Continuous attributes are typically represented as floating-point variables

Attributes

Attributes may be classified into two main types

- **Numeric Attributes**: real-valued or integer-valued domain
  - Interval-scaled: only differences are meaningful, e.g., temperature
  - Ratio-scaled: differences and ratios are meaningful, e.g., Age
- **Categorical Attributes**: set-valued domain composed of a set of symbols
  - Nominal: only equality is meaningful e.g., domain(Sex) = { M, F}
  - Ordinal: both equality (are two values the same?) and inequality (is one value less than another?) are meaningful e.g., domain(Education) = { High School, BS, MS, PhD}

Attribute Types: Qualitative

- **Nominal:** categories, states, or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, letter grades, army rankings

Iris Dataset Extract

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Data: Algebraic and Geometric View

- For numeric data matrix *D*, each row is a d-dimensional data point (i.e., a vector with d attributes):

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}^T \in \mathbb{R}^d$$

whereas each column is an n-dimensional attribute vector (i.e., a vector with n data points).

$$X_j = \begin{pmatrix} x_{1j} & x_{2j} & \cdots & x_{nj} \end{pmatrix}^T \in \mathbb{R}^n$$

Data: Algebraic and Geometric View

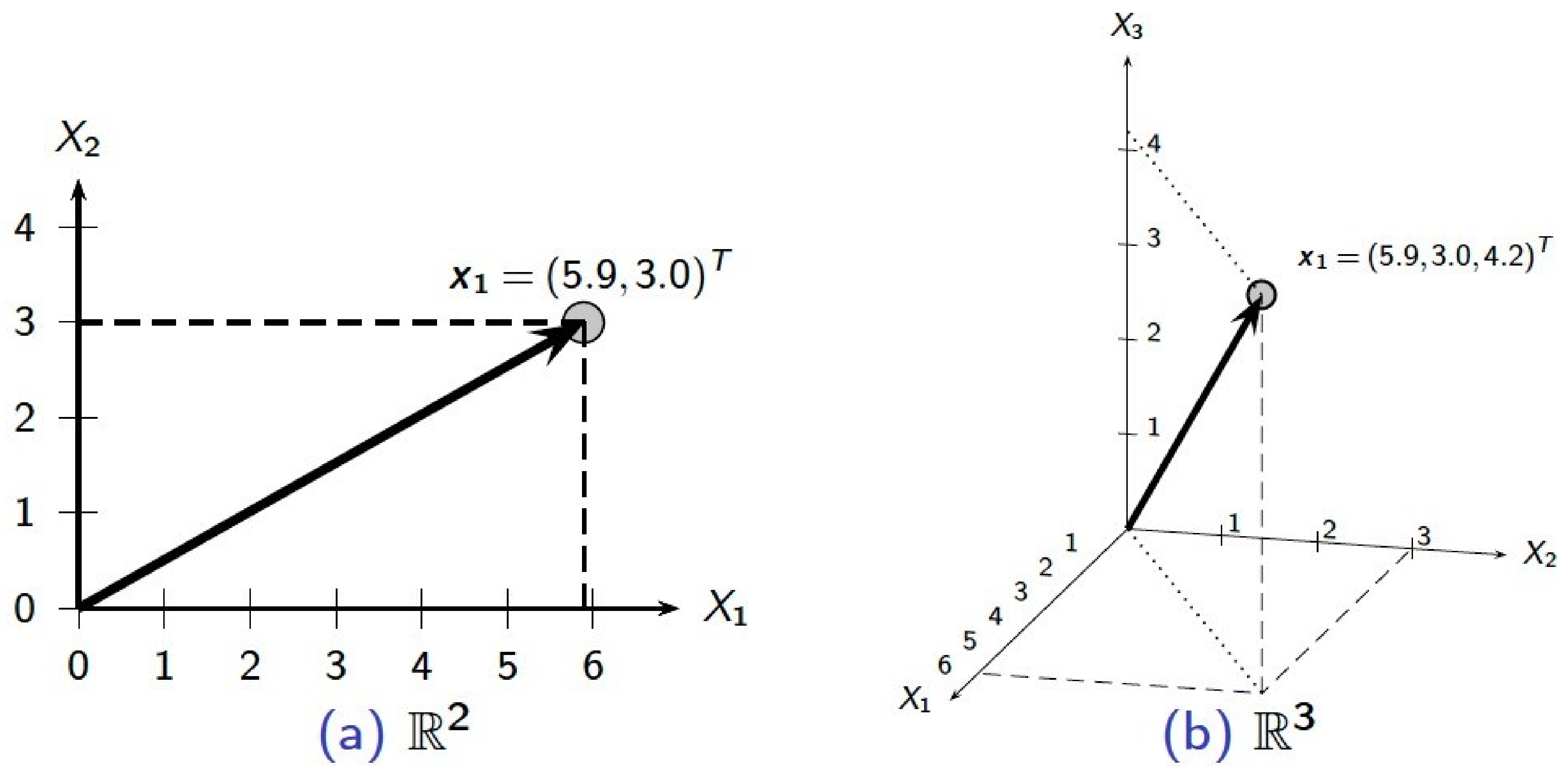
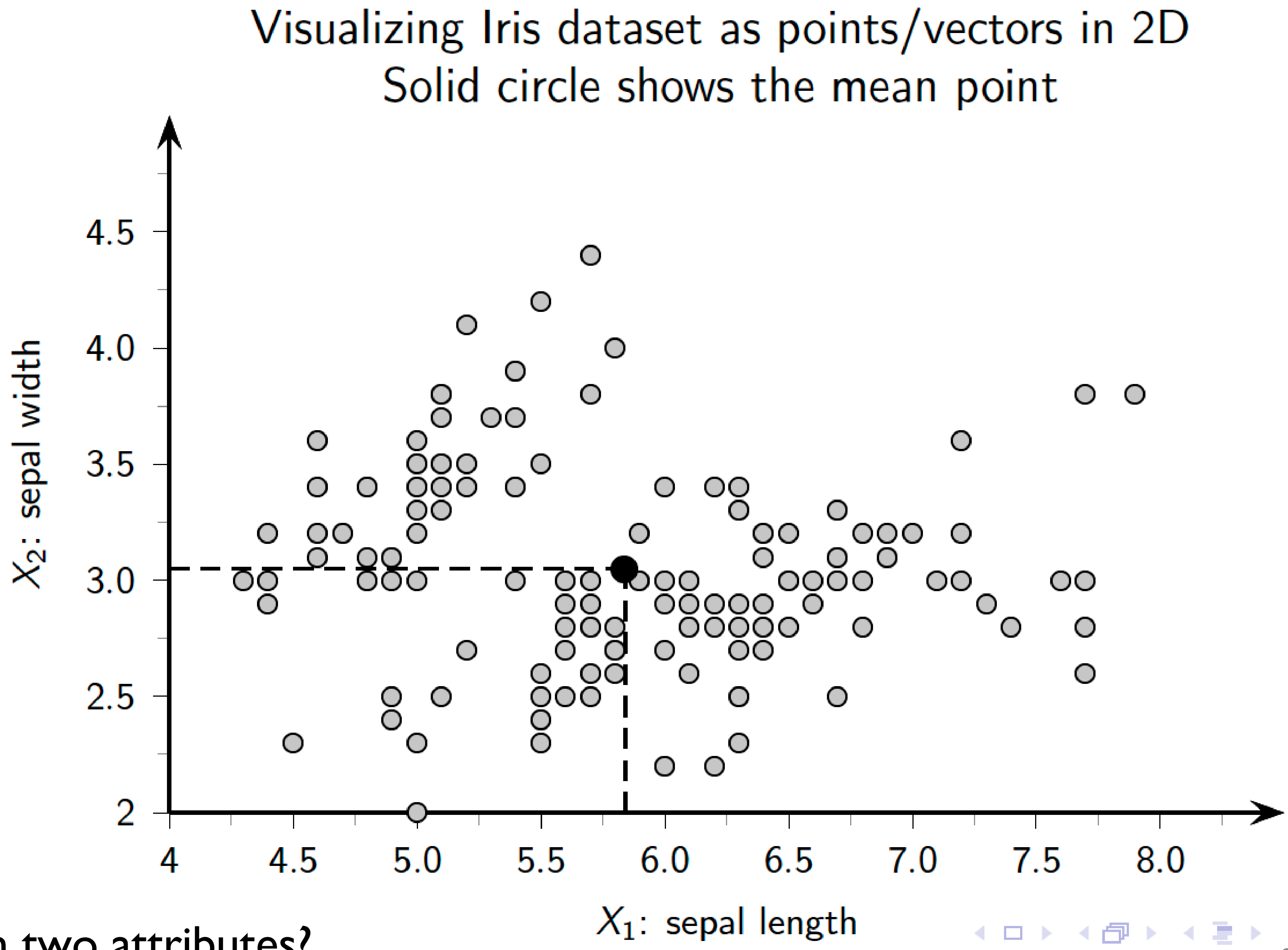


Figure: Projections of $x_1 = (5.9, 3.0, 4.2, 1.5)^T$ in 2D and 3D

Scatterplot: 2D Iris Dataset *sepal length* versus *sepal width*.



What about more than two attributes?.

Numeric Data Matrix

- If all attributes are numeric, then the data matrix D is an n x d matrix, or equivalently a set of n row vectors $x_i^T \in R^d$ or a set of d column vectors $X_j \in R^n$

$$\boldsymbol{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} — \boldsymbol{x}_1^T — \\ — \boldsymbol{x}_2^T — \\ \vdots \\ — \boldsymbol{x}_n^T — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

- The **mean** of the data matrix D is the average of all the points:

$$mean(\boldsymbol{D}) = \boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

Numeric Data Matrix

- The **centered data matrix** is obtained by subtracting the mean from all the points:

$$
\boldsymbol{Z} = \boldsymbol{D} - \boldsymbol{1} \cdot \boldsymbol{\mu}^T = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}^T \\ \boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T - \boldsymbol{\mu}^T \\ \boldsymbol{x}_2^T - \boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{x}_n^T - \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_1^T \\ \boldsymbol{z}_2^T \\ \vdots \\ \boldsymbol{z}_n^T \end{pmatrix}
$$

where $\boldsymbol{z}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$ is a centered point, and $\boldsymbol{1} \in \mathbb{R}^n$ is the vector of ones.

Norm, Distance and Angle

Given two points $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, their *dot product* is defined as the scalar

$$\begin{aligned} \boldsymbol{a}^T \boldsymbol{b} &= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m \\ &= \sum_{i=1}^{m} a_i b_i \end{aligned}$$

Distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given as

$$\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

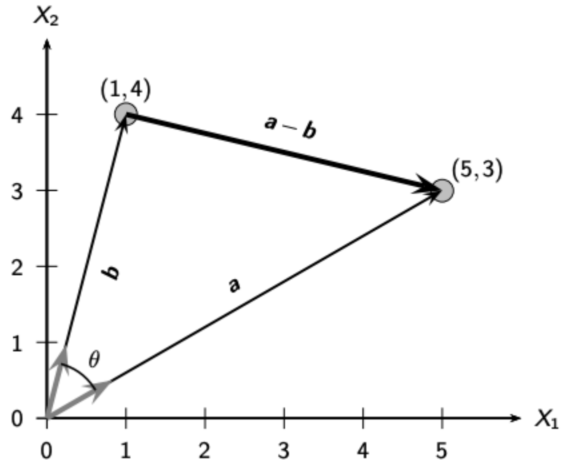
Norm, Distance and Angle

The *Euclidean norm* or *length* of a vector $\boldsymbol{a}$ is defined as

$$\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^T\boldsymbol{a}} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

Angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given as

$$\cos\theta = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\|\boldsymbol{a}\|\,\|\boldsymbol{b}\|} = \left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}\right)^T\left(\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right)$$



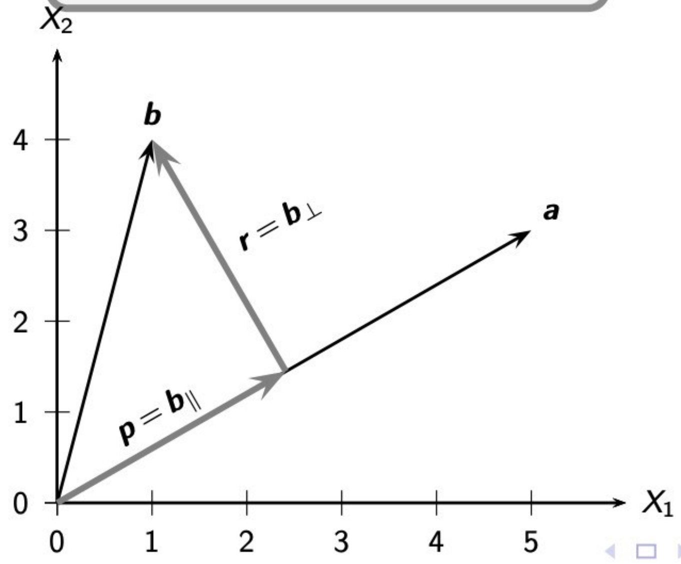
Orthogonal Projection

Two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are *orthogonal* iff $\boldsymbol{a}^T\boldsymbol{b} = 0$, i.e., the angle between them is 90°. Orthogonal projection of $\boldsymbol{b}$ on $\boldsymbol{a}$ comprises the vector $\boldsymbol{p} = \boldsymbol{b}_{\|}$ parallel to $\boldsymbol{a}$, and $\boldsymbol{r} = \boldsymbol{b}_{\perp}$ perpendicular or orthogonal to $\boldsymbol{a}$, given as

$$\boldsymbol{b} = \boldsymbol{b}_{\|} + \boldsymbol{b}_{\perp} = \boldsymbol{p} + \boldsymbol{r}$$

where

$$\boxed{\boldsymbol{p} = \boldsymbol{b}_{\|} = \left(\frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\right)\boldsymbol{a}}$$



DATA: PROBABILISTIC VIEW

- The probabilistic view of the data assumes that each numeric attribute X is a *random variable*, defined as a function that assigns a real number to each outcome of an experiment.

- Formally, X is a function $X : O \rightarrow R$, where O, the domain of X, is the set of all possible outcomes of the experiment, also called the sample space, and R, the range of X, is the set of real numbers.

X (O: all possible outcomes or sample space, R= Range)

- If the outcomes are numeric, and represent the observed values of the random variable, then $X: O \rightarrow O$ is simply the identity function: $X(v) = v$ for all $v \in O$.

DATA: PROBABILISTIC VIEW

- The distinction between the outcomes and the value of the random variable is important, as we may want to treat the observed values differently depending on the context

- A random variable X is called a <u>discrete</u> random variable if it takes on only a finite or countably infinite number of values in its range, whereas X is called a <u>continuous</u> random variable if it can take on any value in its range.

Example

- Consider the sepal length attribute ($X1$) for the Iris dataset in.
- All $n = 150$ values of this attribute lie in the range $[4.3, 7.9]$, with centimeters as the unit of measurement.
- Let us assume that these constitute the set of all possible outcomes O.

- By default, we can consider the attribute $X1$ to be a continuous random variable, given as the identity function $X1(v) = v$, because the outcomes (sepal length values) are all numeric.

Example Cont.,

- On the other hand, if we want to distinguish between Iris flowers with short and long sepal lengths, with long being, say, a length of 7 cm or more, we can define a discrete random variable A as follows:

$$A(v) = \begin{cases} 0 & \text{if } v < 7 \\ 1 & \text{if } v \geq 7 \end{cases}$$

- In this case the domain of A is [4.3,7.9], and its range is {0,1}.

Probability Mass Function

- If X is discrete, the probability mass function of X is defined as

$$\boxed{f(x) = P(X = x) \qquad \forall x \in \mathbb{R}}$$

- Intuitively, for a discrete variable X, the probability is concentrated or massed at only discrete values in the range of X, and is zero for all other values.

Next…

Data Exploration