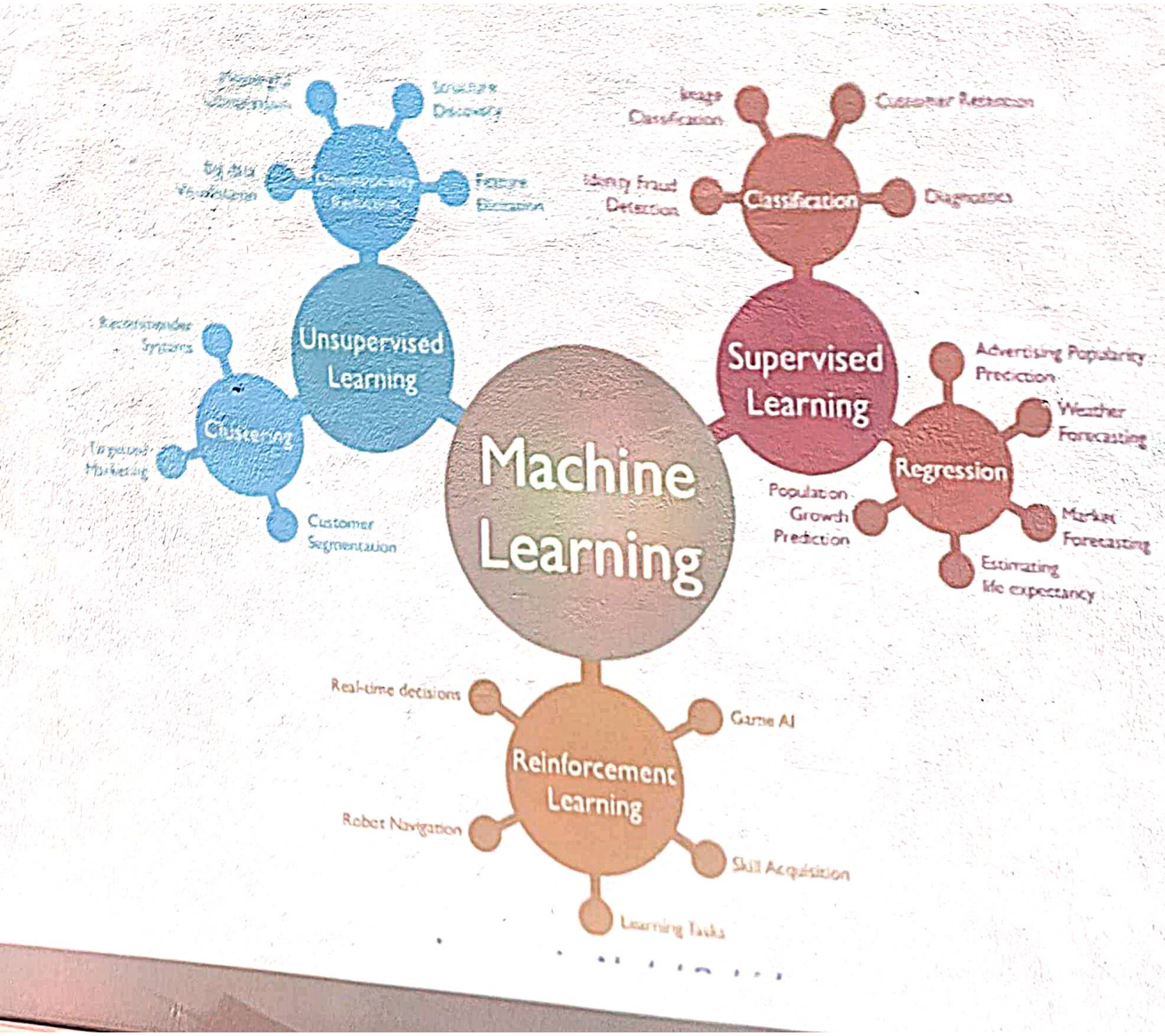
Machine Learning

- **Machine Learning (ML)** is a branch of Artificial Intelligence (AI) that focuses on creating systems that can **learn from data and improve their performance over time without being explicitly programmed.**
- Instead of writing step-by-step rules, we feed data to algorithms that automatically find patterns and make predictions or decisions.
- Traditional programming:
  Rules (program) + Data → Output
- Machine learning:
  Data + Output (examples) → Algorithm learns rules


Machine Learning
Unsupervised Learning
Clustering
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Machine Learning ...

- Can handle complex tasks where writing rules is impossible (like recognizing faces, understanding speech).
- Improves automatically as more data becomes available.
- **Examples**
- **Email spam filter**: Learns from examples of spam and non-spam emails.
- **Voice assistants (Siri, Alexa)**: Learn from speech data.
- **Netflix/YouTube recommendations**: Learn from your watch history.
- **Medical diagnosis**: Predict diseases from scans or patient data.

Supervised Learning

- The model is trained on a labeled dataset (input → output pairs). Th goal is to learn a mapping from inputs to known outputs.
- "Learn from examples with answers provided."
- **Examples**:
  - Predicting house prices (input: size, location → output: price).
  - Classifying emails (spam vs. not spam).
- **Techniques**: Linear regression, logistic regression, decision trees, random forests, SVMs, neural networks.

Unsupervised Learning

- The model works with **unlabeled data** (no predefined outputs). The goal is to find hidden patterns or structures in the data.
- "Learn from data without answers."
- **Examples**:
  - Customer segmentation in marketing.
  - Clustering similar documents or news articles.
  - Dimensionality reduction (PCA, t-SNE).
- **Techniques**: K-means clustering, hierarchical clustering, PCA, autoencoders.

Semi-Supervised Learning

- A mix of **small amounts of labeled data** + **large amounts of unlabeled data**. Useful when labeling is expensive or difficult.
- "Learn with a few answers and many questions."
- **Examples**:
  - Medical imaging (few labeled scans, many unlabeled).
  - Speech recognition (some transcribed audio + much raw audio).
- **Techniques**: Self-training, graph-based methods, semi-supervised neural networks.

Reinforcement Learning (RL)

- The model (called an agent) learns by interacting with an environment. It receives **rewards or penalties** based on its actions and tries to maximize long-term reward.
- "Learn by trial and error with feedback."
- **Examples**:
  - Game playing (e.g., AlphaGo beating human champions).
  - Robotics (teaching a robot to walk).
  - Recommendation systems (optimizing what to show users).
- **Techniques**: Q-learning, Deep Q-Networks (DQN), Policy Gradient methods.

Parameter Estimation

- In statistics and machine learning, we often assume data comes from some **probability distribution** (e.g., Normal distribution, Bernoulli distribution).
- This distribution has **parameters** (e.g., mean μ, variance σ^2).
- **Parameter estimation** means: from observed data, we try to **find the bes values of these parameters**.
- Parameter estimation helps us figure out the most likely values of thes unknown parameters from the data.
- The purpose of parameter estimation is to **learn about the underlyin population distribution**, so that we can **model, predict, and mal decisions** based on limited observed data.

Parameter Estimation

- In statistics and machine learning, we often assume data comes from some **probability distribution** (e.g., Normal distribution, Bernoulli distribution).
- This distribution has **parameters** (e.g., mean μ, variance σ^2).
- **Parameter estimation** means: from observed data, we try to **find the best values of these parameters.**
- Parameter estimation helps us figure out the most likely values of these unknown parameters from the data.
- The purpose of parameter estimation is to **learn about the underlying population distribution**, so that we can **model, predict, and make decisions** based on limited observed data.

Maximum Likelihood Estimation(MLE)

- **Probability Distribution Function (PDF) :** describes how probabilities are distributed over the possible values of a random variable.
- **$f(x;\theta)$** (sometimes written $f(x|\theta)$, it usually means the **probability distribution function** (PMF or PDF) of a random variable X, parameterized by some **parameter(s) θ**.
- $x \rightarrow$ the random variable's value.
- $\theta \rightarrow$ the parameters that define the shape/location/scale of the distribution.
- $f(x;\theta)$ is a **family of distributions**, and each choice of θ gives one specific distribution.

Examples

1. Bernoulli Distribution (discrete)

$$f(x;p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\},\ 0 < p < 1$$

Here,

- x = outcome (0 or 1)
- $\theta = p$ = probability of success

2. Normal (Gaussian) Distribution (continuous)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here,

- x = real number
- $\theta = (\mu, \sigma^2)$ = mean and variance

3. Exponential Distribution

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0,\ \lambda > 0$$

Here,

- x = waiting time
- $\theta = \lambda$ = rate parameter

✅ **In short:**

- $f(x;\theta)$ = the probability distribution function of X, parameterized by θ.
- θ controls the shape of the distribution.

3. Exponential Distribution

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0,\ \lambda > 0$$

Here,

- x = waiting time
- $\theta = \lambda$ = rate parameter

✅ In short:

- $f(x;\theta)$ = the probability distribution function of X, parameterized by θ.
- θ controls the shape of the distribution.

MLE ...

- MLE is a method to estimate the parameters of a probability distribution given observations
- MLE attempt to find the parameter values that maximize the likelihood function
- The resulting estimate is called Maximum Likelihood Estimate (MLE
- In Machine learning we have to find $P(\theta \mid \text{data})$.
  - Uses Bayes' theorem:

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

Likelihood function

Data are sample in the datase

- So data can be written as X
- We need to maximize $P(x \mid \theta)$
- The likelihood of sample X is a function of the parameter θ
- It can be defined as the chance that the parameter θ would generate the observed data

$$l(\theta) = P(x \mid \theta)$$

- Here X = {x1, x2, x3, ... xn} .... Assume they are independent

$$l(\theta) = P(x_1, x_2, \ldots, x_n \mid \theta)$$

- This is the likelihood function

$$l(\theta) = P((x_1, \ldots, x_n) \mid \theta) = P(x_1 \mid \theta)\, P(x_2 \mid \theta) \cdots P(x_n \mid \theta).$$

- We take logarithm for reducing the computational complexity
- We *define* a new function
- $$L(\theta) = \log l(\theta).$$

$$L(\theta) = \log l(\theta) = \log \left[P(x_1 \mid \theta)\, P(x_2 \mid \theta) \cdots P(x_n \mid \theta). \right]$$

Using the log property $\log(ab) = \log a + \log b$, we expand:

$$L(\theta) = \log P(x_1 \mid \theta) + \log P(x_2 \mid \theta) + \cdots + \log P(x_n \mid \theta).$$

- This function is called log likelihood function
- Now we have to maximize the the log likelihood function to estimate the parameter of the probability distribution

For continuous data (like heights, weights), we can't really talk about $P(X = x_i)$ (that probability is 0). Instead, we use the probability density function (pdf):

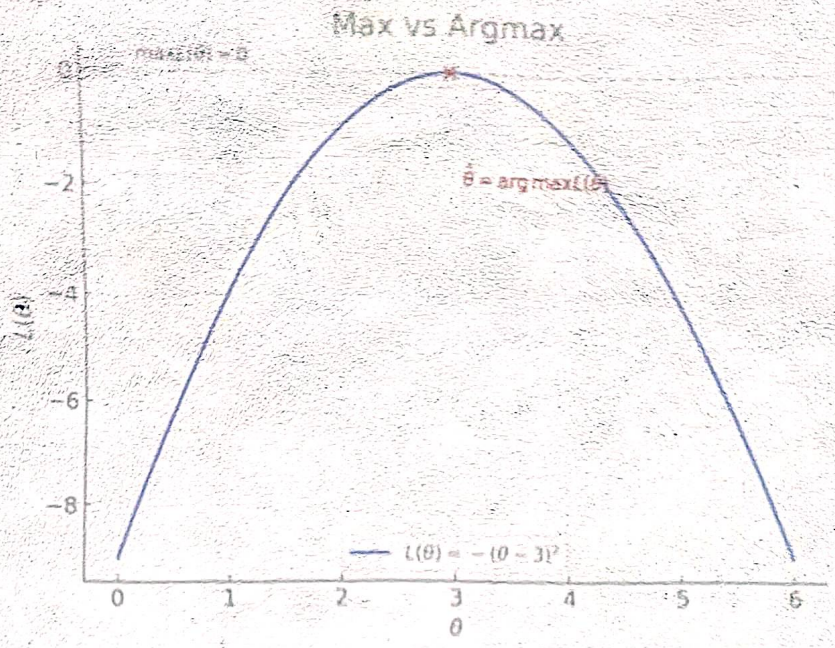
$$f(x_i \mid \theta).$$

So, the likelihood with densities looks like:

$$l(\theta) = f(x_1 \mid \theta)\, f(x_2 \mid \theta) \cdots f(x_n \mid \theta),$$

and the log-likelihood is:

$$L(\theta) = \log f(x_1 \mid \theta) + \log f(x_2 \mid \theta) + \cdots + \log f(x_n \mid \theta).$$



e's the visual:

- The blue curve is the log-likelihood $L(\theta)$.
- The red dot marks $\hat{\theta} = 3$, the arg max (the location of the maximum).
- The gray dashed line shows $\max L(\theta) = 0$, the value of the maximum.

- A maximum occurs at a **critical point** where the slope is zero.
  So why we set the derivative w.r.t θ t zero to find the maximum.

- This is the likelihood function

$$l(\theta) = P((x_1, \ldots, x_n) \mid \theta) = P(x_1 \mid \theta)\, P(x_2 \mid \theta) \cdots P(x_n \mid \theta).$$

- We take logarithm for reducing the computational complexity
- We *define* a new function

$$L(\theta) \equiv \log l(\theta).$$

$$L(\theta) = \log l(\theta) = \log \left[P(x_1 \mid \theta)\, P(x_2 \mid \theta) \cdots P(x_n \mid \theta). \right]$$

Using the log property $\log(ab) = \log a + \log b$, we expand:

$$L(\theta) = \log P(x_1 \mid \theta) + \log P(x_2 \mid \theta) + \cdots + \log P(x_n \mid \theta).$$

- This function is called log likelihood function
- Now we have to maximize the the log likelihood function to estimate the parameter of the probability distribution

For continuous data (like heights, weights), we can't really talk about $P(X = x_i)$ (that probability is 0). Instead, we use the probability density function (pdf):

$$f(x_i \mid \theta).$$

So, the likelihood with densities looks like:

$$l(\theta) = f(x_1 \mid \theta)\, f(x_2 \mid \theta) \cdots f(x_n \mid \theta),$$

and the log-likelihood is:

$$L(\theta) = \log f(x_1 \mid \theta) + \log f(x_2 \mid \theta) + \cdots + \log f(x_n \mid \theta).$$

Maximum Likelihood Estimation(MLE) - Examples

The Bernoulli distribution is the simplest probability distribution for a discrete random variable. It models a random experiment that has exactly two possible outcomes:

- Success (usually coded as 1)
- Failure (usually coded as 0)

Probability Mass Function (PMF) / Density Function

If $X \sim \text{Bernoulli}(p)$, then

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0,1\},\ 0 \leq p \leq 1$$

Where:

- $p = P(X = 1)$ is the probability of success,
- $1 - p = P(X = 0)$ is the probability of failure.

Bernoulli distribution ...

For a single observation $x \in \{0, 1\}$, with parameter θ (here $\theta = p$):

$$f(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \quad 0 \le \theta \le 1$$

Suppose we have independent samples:

$$x_1, x_2, \ldots, x_n \sim \text{Bernoulli}(\theta)$$

$$\overline{L(P)} = \log\left[\left[P^{x_1}(1-P)^{1-x_1}\right] \times \left[P^{x_2}(1-P)^{1-x_2}\right] \times \cdots \times \left[P^{x_n}(1-P)^{1-x_n}\right]\right]$$

$$= \log\left[P^{x_1}(1-P)^{1-x_1}\right] + \log\left[P^{x_2}(1-P)^{1-x_2}\right] + \cdots + \log\left[P^{x_n}(1-P)^{1-x_n}\right]$$

$$= \left[x_1 \log P + (1-x_1)\log(1-P)\right] + \left[x_2 \log P + (1-x_2)\log(1-P)\right] + \cdots + \left[x_n \log P + (1-x_n)\log(1-P)\right]$$

$$= (x_1 + x_2 + \cdots + x_n)\log P + \left[(1-x_1) + (1-x_2) + \cdots + (1-x_n)\right]\log(1-P)$$

Bernoulli distribution ...

- Suppose $Y = x_1 + x_2 + \cdots + x_n$

$$L(P) = Y \log P + (n - Y) \log(1 - P)$$

- To maximise
  - Derivative w.r.t. P: Equate to zero

$$\frac{Y}{P} - \frac{n - Y}{1 - P} = 0$$

Rearrange

$$\frac{Y}{P} = \frac{n - Y}{1 - P}$$

Cross multiply:

$$Y(1 - P) = (n - Y)P$$

$$Y - YP = nP - YP$$

$$Y = nP$$

$$P = \frac{Y}{n}$$

$$P = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

So the MLE of P is the sample mean.

Poisson Distribution

$$f(x, \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Likelihood for n i.i.d. samples (before log)

For observations $x_1, x_2, \ldots, x_n$:

$$= \left[\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\right] \times \left[\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right] \times \cdots \times \left[\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}\right].$$

Log-likelihood $L(\lambda)$

Take log (log of product = sum of logs). Expand term-by-term:

$$L(\lambda) = \log\left[\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\right] + \log\left[\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right] + \cdots + \log\left[\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}\right]$$

apply $\log\left(\frac{A}{B}\right) = \log A - \log B$ and $\log(AB) = \log A + \log B$, and $\log(e^{-\lambda}) = -\lambda$, $\log(\lambda^{x_i}) = x_i \log\lambda$. So each term becomes:

$$\log\left[\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] = \log(e^{-\lambda}) + \log(\lambda^{x_i}) - \log(x_i!) = -\lambda + x_i \log\lambda - \log(x_i!).$$

$$= \left[-\lambda + x_1 \log\lambda - \log(x_1!)\right] + \left[-\lambda + x_2 \log\lambda - \log(x_2!)\right] + \cdots + \left[-\lambda + x_n \log\lambda - \log(x_n!)\right].$$

Group common terms (use $Y = \sum_{i=1}^{n} x_i$ and constant $C = \sum_{i=1}^{n} \log(x_i!)$)

$$f(\mathbf{x}, \lambda) = -n\lambda + Y \log\lambda - C, \quad \text{where } Y = \sum_{i=1}^{n} x_i,\ C = \sum_{i=1}^{n} \log(x_i!).$$

Poisson Distribution

$$\boxed{f(x,\lambda)} = \frac{e^{-\lambda}\lambda^{x}}{x!}.$$

Likelihood for n i.i.d. samples (before log)

For observations $x_1, x_2, \ldots, x_n$:

Log-likelihood $L(\lambda)$ $$= \left[\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\right] \times \left[\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right] \times \cdots \times \left[\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}\right].$$

Take log (log of product = sum of logs). Expand term-by-term:

$$L(\lambda) = \log\left[\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\right] + \log\left[\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right] + \cdots + \log\left[\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}\right]$$

apply $\log\left(\frac{A}{B}\right) = \log A - \log B$ and $\log(AB) = \log A + \log B$, and $\log(e^{-\lambda}) = -\lambda$, $\log(\lambda^{x_i}) = x_i \log\lambda$. So each term becomes:

$$\log\left[\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] = \log(e^{-\lambda}) + \log(\lambda^{x_i}) - \log(x_i!) = -\lambda + x_i\log\lambda - \log(x_i!).$$

$$= \left[-\lambda + x_1\log\lambda - \log(x_1!)\right] + \left[-\lambda + x_2\log\lambda - \log(x_2!)\right] + \cdots + \left[-\lambda + x_n\log\lambda - \log(x_n!)\right].$$

Group common terms (use $Y = \sum_{i=1}^{n} x_i$ and constant $C = \sum_{i=1}^{n} \log(x_i!)$)

$$\boxed{f(x,\lambda) = -n\lambda + Y\log\lambda - C, \quad \text{where } Y = \sum_{i=1}^{n} x_i,\ C = \sum_{i=1}^{n} \log(x_i!).}$$

Differentiate, set to zero, and solve

Derivative w.r.t. λ:

$$\frac{\partial f}{\partial \lambda} = -n + \frac{Y}{\lambda}.$$

Equate to zero:

$$-n + \frac{Y}{\lambda} = 0 \implies \frac{Y}{\lambda} = n \implies \boxed{\hat{\lambda} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

- Hw –

- Gemetric distribution $P(X = x) = P\,(1 - P)^{x-1}$.

- Gaussian Distribution $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2} \right)$.