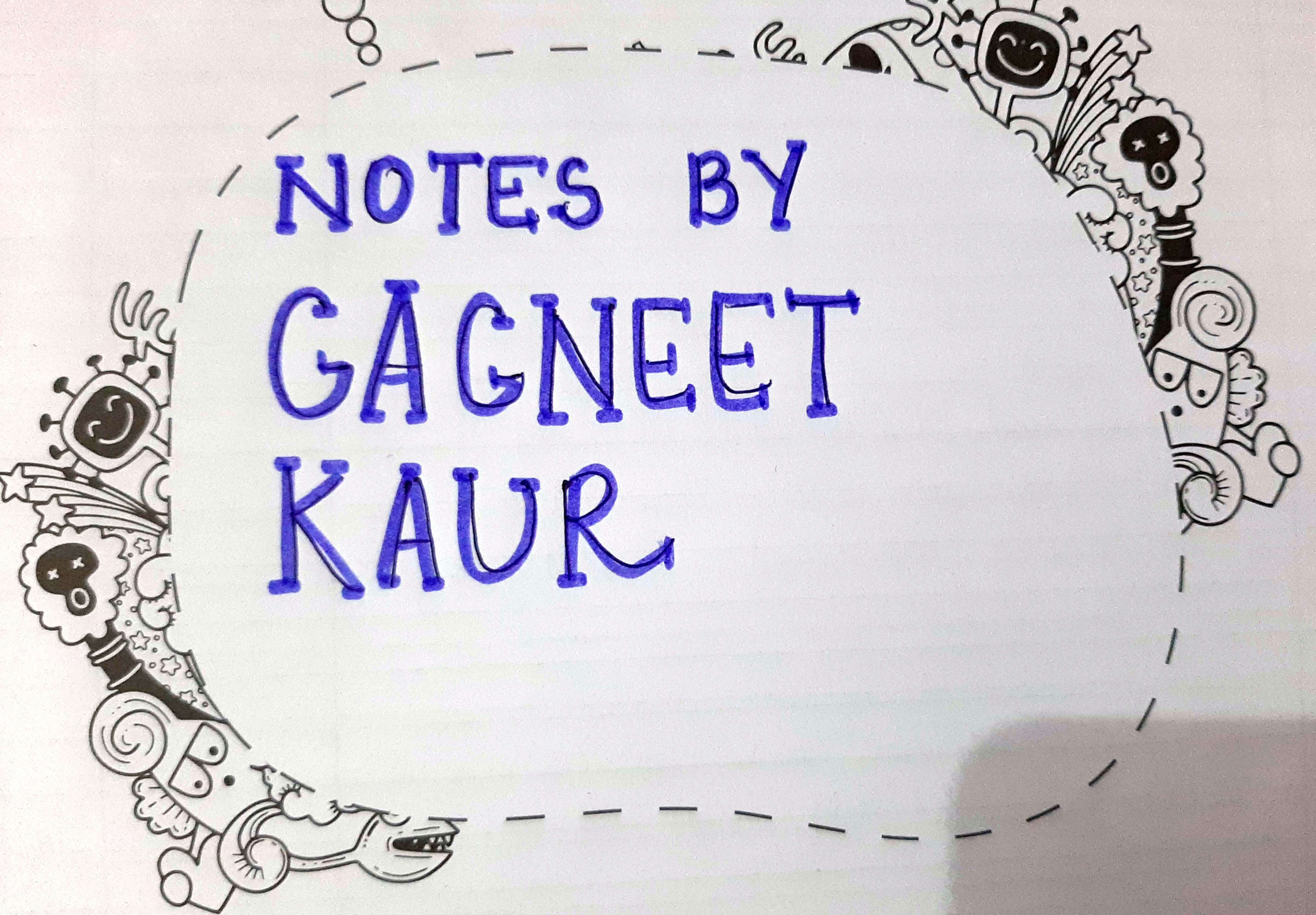

NOTES BY
GAGNEET
KAUR

January 25, 2023

apsara

Date: __/__/____

WEEK 1 - INTRO TO ML -

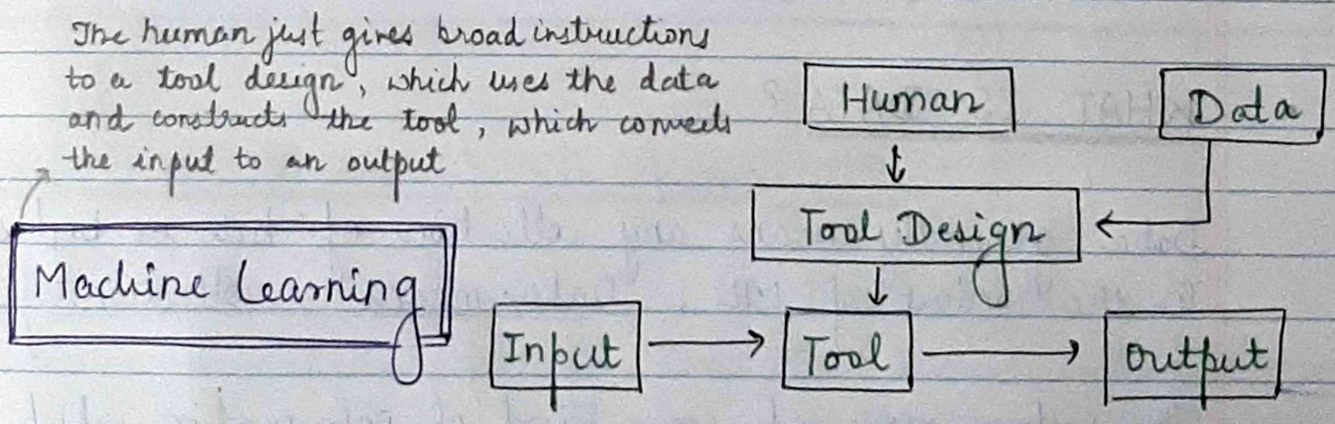
- ## What is Machine Learning ?

→ It is a study of algorithms that improve through the use of data

→ It is sub-field of AI that extracts patterns out of raw data.

→ It allows computers to learn from experience without being explicitly programmed or human intervention.

→ It is used for tasks where programming / human labor fails.



The human does not build the tool himself. He just use a broad blueprint to a tool design box, which constructs the tool.
It also uses an additional source of data along with human inputs to make the tool.

- **Why machine learning ?**

If we can use programming or human labor to do a task, we should not use machine learning and go with the former techniques.

BUT, there are several situations where programming or human labor will just fail.

REASONS WHY HUMAN LABOR FAILS :
not possible to achieve the task at the SCALE, SPEED or COST.

REASONS WHY PROGRAMMING APPROACH FAILS :
- we do not know the exact rules transforming the input to output OR we are unable to express rules.

- DATA , MODELS & ML TASK -

- **WHAT IS DATA ?**

Data generally means any collection of bits or bytes.
In the context of ML, Data means a collection of vectors.

The vectors represent some kind of information which we will call as "FEATURES" (the column vectors).

- **WHAT IS A MODEL ?**

We can define a model as a mathematical simplication of reality.

Basically a predictive model's goal is to predict on the future data.

PROBABILITY models essentialy score reality & you can give any configuration of reality.

apsara

Date: / /

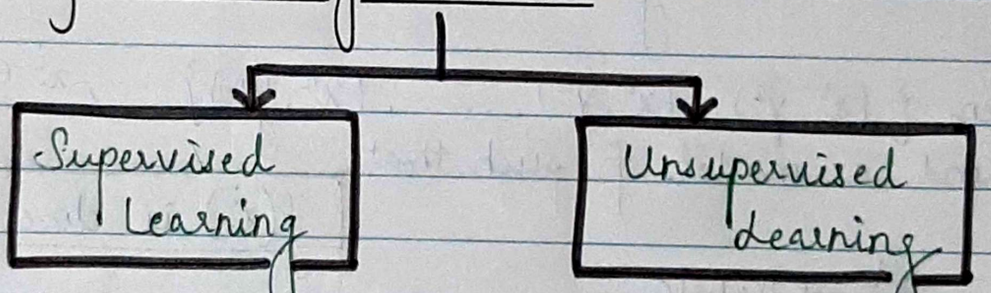
It is almost always representing reality, but it is simpler than reality, and it is more compact than reality.

Models are just approximations.

In ML, ~~they~~ there are two main types of models :

1. PREDICTIVE (You can use the model to make predictions on the unknown or unseen data)
2. PROBABILISTIC → goal is to evaluate how likely a certain event or a configuration is.

Types Of ML Algorithms



- Labels are present
- CURVE - FITTING

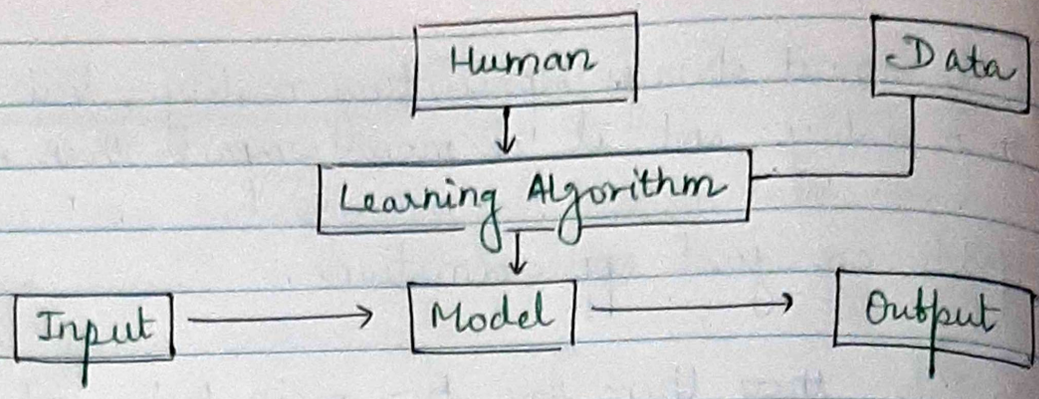
- Labels are not present
- UNDERSTANDING DATA

WHAT IS A LEARNING ALGORITHM ?

It is something that converts data into models.
Learning algorithms are the tools that convert data into model.

HOW ? → Choose from a collection of models with same structure, but different parameters.
→ They give the best model.

MACHINE LEARNING



— SUPERVISED LEARNING —

→ Supervised Learning is CURVE FITTING.

→ Given $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$, $x^i \in \mathbb{R}^d$
find a model f such that $f(x^i)$ is close to y^i

$f : \mathbb{R}^d \rightarrow$ labels

(1) REGRESSION (It is used when output variable is CONTINUOUS)

- Training Data : $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$
- $x^i \in \mathbb{R}^d$, $y^i \in \mathbb{R}$
- Algorithm outputs a model $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- LOSS FUNCTION $= \frac{1}{n} \sum_{i=1}^{n} \left(f(x^i) - y^i \right)^2$
- $f(x) = w^T x + b = \sum_{j=1}^{d} w_j x_j + b$

GOAL TO GET BEST MODEL
→ minimize the LOSS function

apsara
Date: __/__/____

(2) CLASSIFICATION

- Training Data : $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$
- $x^i \in \mathbb{R}^d$, $y^i \in \{+1, -1\}$
- Algorithm outputs a models $f : \mathbb{R}^d \rightarrow \{+1, -1\}$
- LOSS FUNCTION = $\frac{1}{n} \sum_{i=1}^{n} 1\left(f(x^i) \neq y^i\right)$ → fraction of misclassified instances.
- $f(x) = \text{sign}\left(w^T x + b\right)$ → linear separation

---

- Regression Model predicts a real valued variable. It lives on a continuum.
- Classification Model predicts a variable which is discrete in nature.

Page No.: ____

Date: January 26, 2023

apsara

Date: / /

— UNSUPERVISED LEARNING —

→ Unsupervised Learning is 'UNDERSTANDING DATA'.

→ Data : $\{x^1, x^2, \ldots, x^n\}$

"Unsupervised Learning typically is used as a pre-processing step."

→ $x^i \in \mathbb{R}^d$

→ Build models that compress, explain & group data.

(1) DIMENSIONALITY REDUCTION

→ Compression and simplification

→ Data : $\{x^1, x^2, \ldots, x^n\}$, where $x^i \in \mathbb{R}^d$

→ Encoder $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$

→ Decoder $f : \mathbb{R}^{d'} \rightarrow \mathbb{R}^d$

$\Big\}\ d' < d$

→ GOAL : $g(f(x^i)) \approx x^i$

→ LOSS FUNCTION $= \frac{1}{n} \sum_{i=1}^{n} \| g(f(x^i)) - x^i \|^2$

* Dimensionality reduction is the process of reducing the numbers of features (or dimensions) in a dataset while retaining as much as information as possible.

(2) DENSITY ESTIMATION

→ The output of a density estimation algorithm is a probabilistic model.

→ A density estimation model takes in several samples from a random source, and outputs a model that assigns a probability score to every possible instance.

∴ Data : $\{x^1, x^2, \ldots, x^n\}$, $x^i \in \mathbb{R}^d$

∴ Probability mapping $P : \mathbb{R}^d \to \mathbb{R}_+$ that sums to one.

∴ GOAL : $P(x)$ is large if $x \in$ Data, and low otherwise

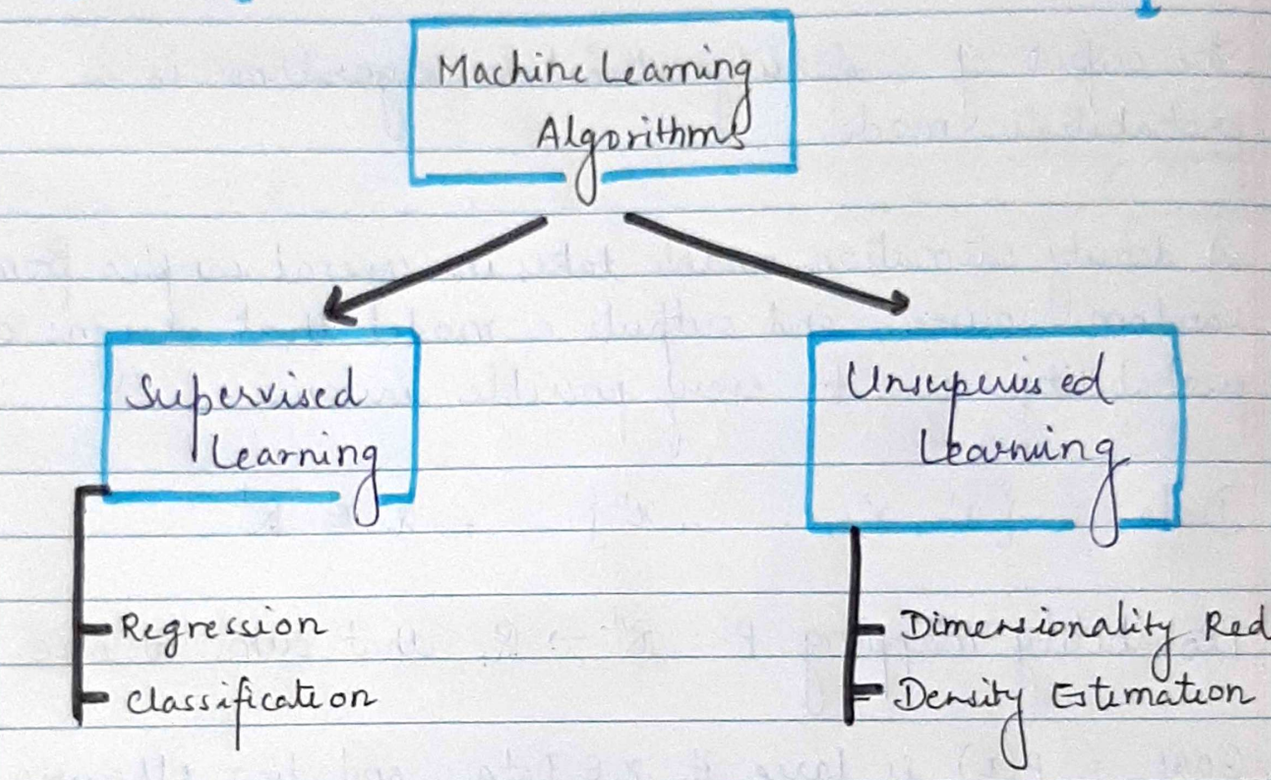
∴ LOSS FUNCTION $= \frac{1}{n} \sum_{i=1}^{n} -\log\left(P(x^i)\right)$ → negative log likelihood

---

Why Dimensionality Red^n?

→ to reduce the complexity of a model
→ to improve the performance of a learning algorithm
→ to make it easier to visualize data.

IMPORTANT CONCEPTS FROM W-1



LOSS FUNCTIONS

Regression Model

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(x^i)-y^i\right)^2$$

↑ Supervised ↓

Classification Model

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

Dimensionality Red^n

$$\frac{1}{n}\sum_{i=1}^{n}\left\| g(f(x^i))-x^i \right\|^2$$

↑ Unsupervised ↓

Density Estimation

$$\frac{1}{n}\sum_{i=1}^{n}-\log\left(P(x^i)\right)$$

Goal is to minimize the loss function to get best model.

TYPES OF DATA for Model Evaluation & Selection

- **Training Data**

→ This type of data builds up the machine learning algorithm. This is the actual dataset that we use to train the model. The model sees and learns from this data.

- **Validation Data**

→ The validation data is used to evaluate a given model, but this is for frequent evaluation. The model occassionally sees this data, but never does it "learn" from this.

- **Test Data**

→ The test data is used only once a model is completely trained (using training & validation data). The test set is generally what is used to evaluate competing models. It contains carefully sampled data that spans the various classes that the model would face, when used in real world.

NOTE : We evaluate our model on TEST DATA.
Model selection is done using VALIDATION DATA.