DEFINITION [ARTHUR SAMUEL 1959 – PIONEER OF AI, COINED THE TERM MACHINE LEARNING]

MACHINE LEARNING GIVES THE COMPUTERS THE ABILITY TO LEARN WITHOUT EXPLICITLY BE PROGRAMMED.

DEFINITION (TOM MITCHEL 1998 – WRITTEN ONE OF THE FIRST BOOK OF ML)

TOM: "THE PREVIOUS DEFINITION IS NOT WELL POSED!"

A MACHINE LEARNING ALGORITHM IS SAID TO LEARN FROM EXPERIENCE **T** WITH TO RESPECT A TASK **H** AND SOME MEASURE OF PERFORMANCE **P** IF ITS PERFORMANCES ON **H**, AS MEASURED BY **P**, IMPROVE WITH EXPERIENCE **T**.

WHAT "EXPERIENCE" MEAN?
WHAT "TASK" MEANS?
WHAT "PERFORMANCE" MEANS?
] HOW THESE THREE ARE RELATED EACH OTHER?

TYPE OF LEARNING ALGORITHMS

SUPERVISED LEARNING

$x^{(i)} \in \mathbb{R}^d \quad y^{(i)} \in D \quad i = 1, \dots, m$

$x^{(i)} \rightarrow [H] \rightarrow y^{(i)}$

$D = \mathbb{R} \rightarrow$ REGRESSION

$D = \{0, 1, \dots K\} \rightarrow$ CLASSIFICATION

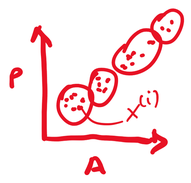
QUESTIONS: WHO IS **T**? WHO IS **E**? WHO IS **P**? WHO IS **x**? WHO IS **y**?

UNSUPERVISED LEARNING

FIND **H** WHICH IS ABLE TO MODEL DATA

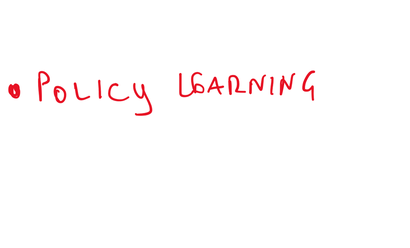
$x^{(i)} \in \mathbb{R}^d \quad i = 1, \dots, m$

- CLUSTERING



How many S, M, L, XL?

REINFORCEMENT LEARNING



- POLICY LEARNING

SELF-SUPERVISED

A MIX OF SUPERVISED AND UNSUPERVISED LEARNING

MOTIVATION: PRODUCING LABELED DATA IS EXPENSIVE (FOR SOME DOMAIN VERY HARD, EG. MEDICAL)

LEAR HOW INFANTS: LEARNING IN AN UNSUPERVISED WAY EXPLOITING SUPERVISION GIVEN BY DATA

EG. PUZZLE → RELATIVE POSITION

KEY INGREDIENTS OF ML ALGORITHM

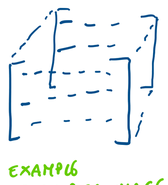
- MODEL = PARAMETRIZED FUNCTION (HYPOTHESIS) **H**
- COST FUNCTION **J**
- PARAMETERS ϑ
- TRAINING ALGORITHM / LEARNING PROCEDURE **A**
- TRAINING SET X_{TRAIN}
- VALIDATION SET X_{VAL}
- TEST SET X_{TEST}
- EVALUATION MEASURES **P**

] $X = X_{TRAIN} \cup X_{VAL} \cup X_{TEST}$ (EXPERIMENTAL DATASET)

$X_{TRAIN} \cap X_{VAL} = \emptyset \quad X_{VAL} \cap X_{TEST} = \emptyset$

$X_{TRAIN} \cap X_{TEST} = \emptyset$

NOTATION

INPUT $x^{(i)} = [x_1^{(i)}, \dots, x_d^{(i)}]^T$; $X^{(i)} = \begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \dots & x_{1,n}^{(i)} \\ \vdots & & & \vdots \\ x_{c,1}^{(i)} & x_{c,2}^{(i)} & \dots & x_{c,n}^{(i)} \end{bmatrix}$; $X^{(i)} =$ 

EXAMPLES: A PATIENT MEDICAL RECORD

EXAMPLE: A GRAY SCALE IMAGE

EXAMPLE
- AN RGB IMAGE
- A SERIES OF GRAY SCALE IMAGES OVER TIME

TENSOR

EXAMPLE: A SERIES OF RGB IMAGES OVER TIME

$\vartheta = [\vartheta_0, \vartheta_1, \dots \vartheta_d]^T$ PARAMETERS OF THE MODEL

$x_j^{(i)} \quad i = 1, \dots, m \quad j = 1 \dots d$

- m: SIZE OF THE TRAINING SET
- j: INPUT DIMENSION
- $x_j^{(i)}$: INPUT COMPONENT OF A ONE DIMENSIONAL VECTOR

FOR MATH CONVENIENCE, WE WILL USUALLY EXTEND $x^{(i)}$ ADDING $x_0^{(i)} = 1$

$x^{(i)} = [x_0^{(i)=1}, x_1^{(i)}, \dots, x_d^{(i)}]^T$ → ASSOCIATED TO ϑ_0

HOW CAN WE REPRESENT THE EXPERIENCE **T**?

$$\begin{bmatrix} x_0^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_1^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & & \vdots \\ x_d^{(1)} & x_d^{(2)} & \dots & x_d^{(m)} \end{bmatrix}$$

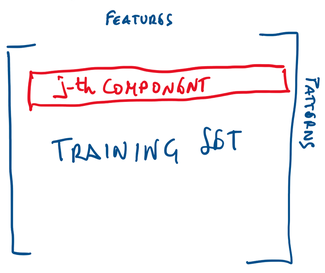
TRAINING SET
(IN CASE OUR PATTERNS ARE VECTORS)

DATA NORMALIZATION

$\hat{x}_j = \frac{x_j - x_j^{min}}{x_j^{max} - x_j^{min}} \in [0, 1]$ MAPPING

$x_j^{min} = \min \{ x_j^{(i)} \mid i = 1, \dots m \}$

$x_j^{max} = \max \{ x_j^{(i)} \mid i = 1, \dots, m \}$



MOST USED: ZERO MEAN NORMALIZATION

$\hat{x}_j = \frac{x_j - \mu_j}{\tilde{\sigma}_j}$ DATA WILL HAVE ZERO MEAN AND STANDARD DEVIATION 1

$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$ MEAN

$\tilde{\sigma}_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2}$

How "SHAPE" OF MY DATA DISTRIBUTION CHANGE? WHY THIS IS USEFUL?