1) Question → question tab
2) Repeat in the chat.
3) TA ✓ → Calling
4) slack, whatsapp (8240991339), titas.de_1@scaler.com

Policy

1) Attendance (Live + Recorded) ≥ 90%
2) HW & Assignments ≥ 75%
3) Participation in Quizzes & Polls ≥ 95%

ML 1 & ML 2 → 2 transition tests

ML 1 → 2 months → Code from scratch, Research oriented topics

ML 2 → 1 month → After 1 month, After 2 months

Agenda

Summary → learnt so far

Overview → 10,000 ft of ML is all about.

- ML ? ML VS SDE
- ML Tasks
- Types of learning in ML

Application

EDA → Exploratory Data Analysis (if time permits)

Already Learnt

1) Python, DSML libraries MUST
2) Probability & Statistics
3) Co-ordinate Geometry. (Linear Algebra)
4) Calculus & Optimisation.

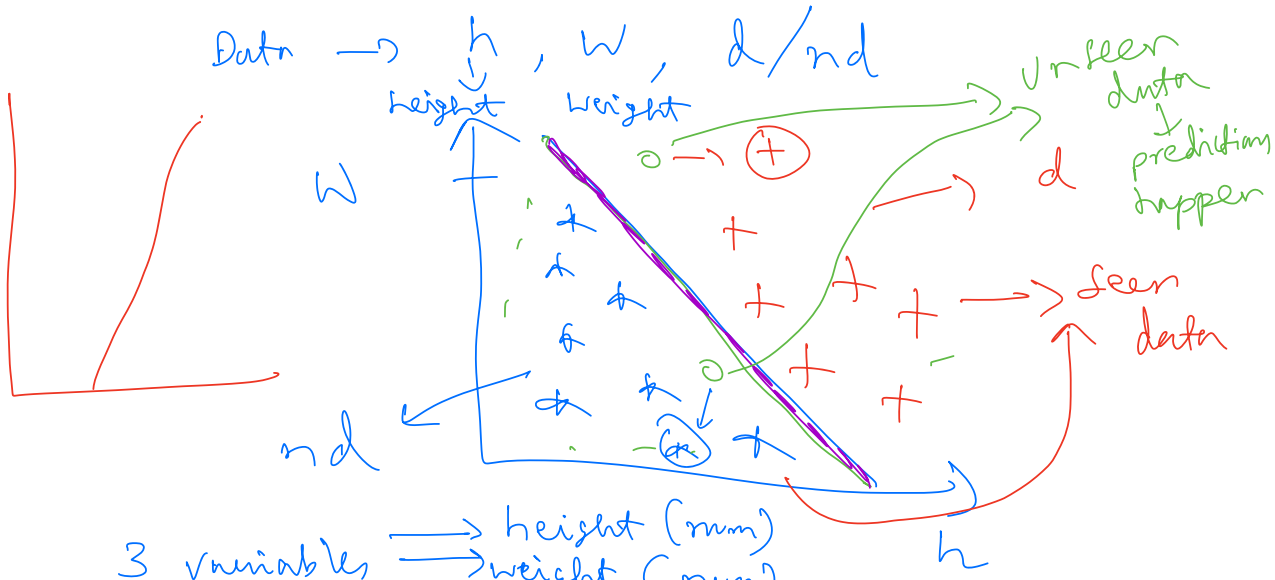
IV → Numerical → histogram, KDE
→ categorical → bar, count plot, pie

2 V → num - num → scatter plot, line plot

cat - cat → stacked bar plot

num - cat → heat plot, clustermap, box plot, violin plot

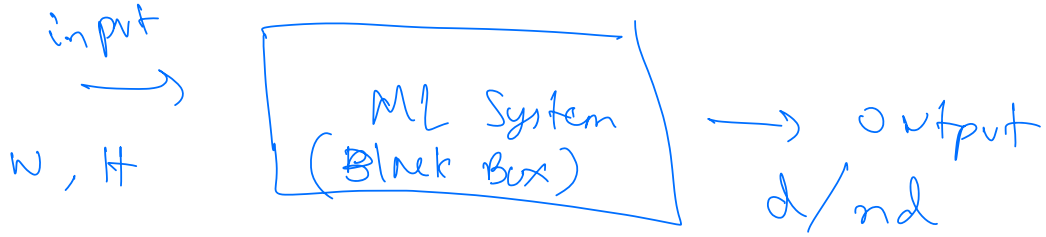
List of patients → predict is diabetic ?



3 variables → height (num)
→ weight (num)
→ label (categorical → d/nd)

input → height & weight (features) →

output → label (d/nd)



Predicting whether diabetic or non-diabetic

H	W	L
6'5"	180	D
6'	190	D
5'11"	150	ND
5'5"	100	ND

Memory

H ≥ 6'

& W ≥ 150 lbs

↓

D

Spam vs Non-Spam Email Classification / Categorization

Spam

Won a lottery

Register to get 1L cashback

Call me for quick 5L loan

Non-Spam →

Memory

List of keywords

"won", "lottery", "quick loan", etc

"lottery" → | l o t t e r y | → unseen

spam

"won" → won

achieved, hit the jackpot

Challenges : → infinite variations of known keywords
→ infinite keywords

ML approach → subject, body
↓
understand patterns
↓
Categorization

l0tt3ry
lottery → same
↓
[ML system]

generalization
↓
broader set of patterns

lottery
l0tt3ry

✓ good night
✓ gud n8

Tom M Mitchell
→ Father of Machine Learning

(T) Task → Spam vs Non-Spam
(P) Performance → Accuracy → % of correct predictions
(E) Experience → Data which is given to you

If with more Experience, your performance is improving with respect to Task (T), then your ML system is actually Learning (L)

Performance → Seen data
Unseen data
↓
performance at unseen data
↓
decision boundary & learning of Patterns

ML Tasks
- classification / Categorization
- regression
- clustering
- recommendation
- Forecasting

Types of ML
- supervised
- Unsupervised
- Reinforcement
- Semi-supervised
- Weakly supervised → Research oriented

Supervised Learning

Classification (categorization)

- finite number of labels present

Data → Label

input → output

1) Spam vs Non-spam
2) Diabetic vs non-diabetic

- labels provided
- only 2 categories/classes possible

Regression → infinite number of labels/values

Predict price of a house

Area (sq ft), No. of bedrooms, No. of bathrooms, which floor, facing which direction

→ infinite number of values

Area	BedRooms	BathRooms	which floor	direction	Price
1080	3	2	4	NE	50L
1800	3	3	3	NE	1.5 Cr
600	1	2	1	W	10L

0 to 2

50 L, 10 L, 1.5 Cr

↓

Regression task

50, 10, 150

↓ ↓

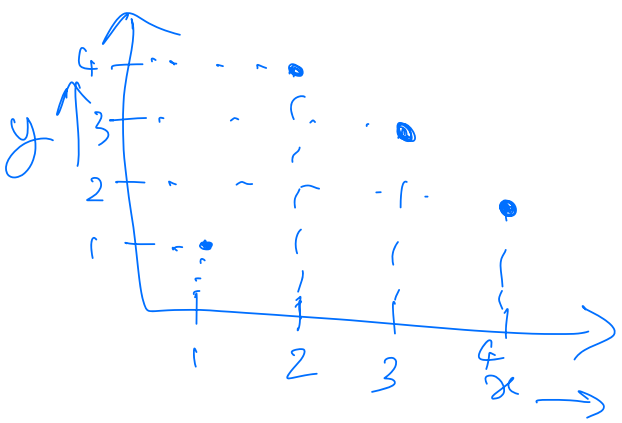
26.2 68.73

↓

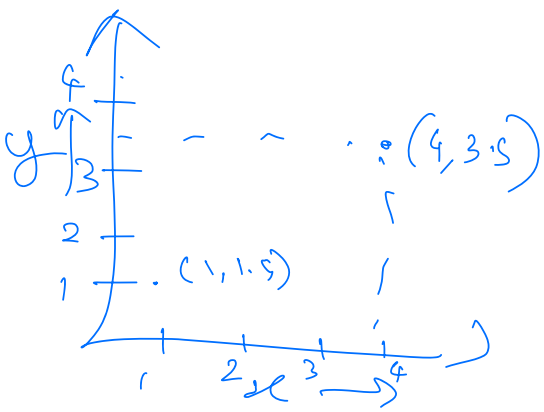
Continuous values

↓

non-discrete



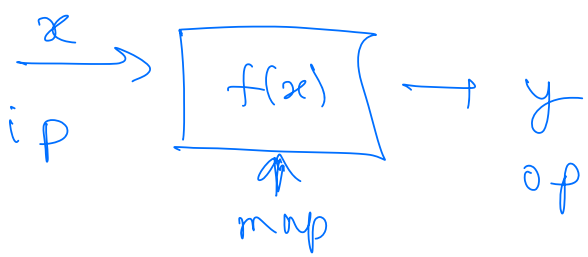
Discrete example



Continuous example

Regression

Classification → (ip, op) pair



(ip1, 5.5)
(ip2, 6.4)
(ip3, 0.1)
(ip4, 1.5)

Seen data

Unseen data

(ip5, ?)

ip5 = ip1,

(ip5, 5.5)

Classification ?

or

Regression ?

10L, 15L, 28.3L, 34.7L, 122L, 105L, ...

0 – 10L, 10L – 20L, 20L – 30L, ..., 140 – 150L

15 ranges

↓

15 classes / categories

Classification → Discrete → 1, 2, 3,

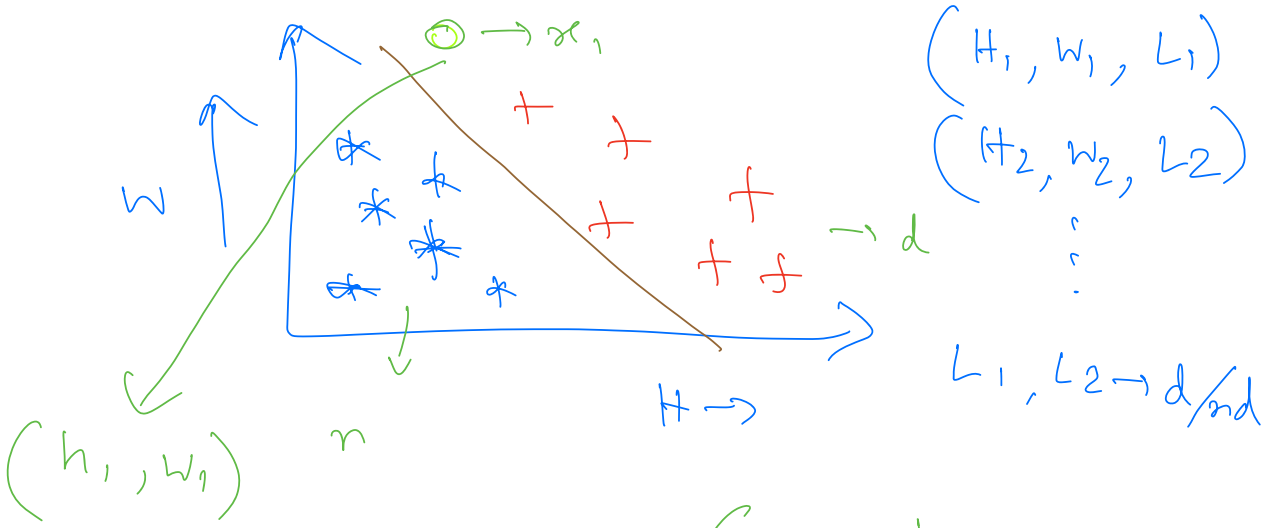
Continuous → 1.11, 2.75, 3.9999

↓

Real-valued ∈ I, R

Rational →

Irrational → $\sqrt{3}$, $\sqrt{2}$



$x_1 = d > 0.5$

diabetic patient

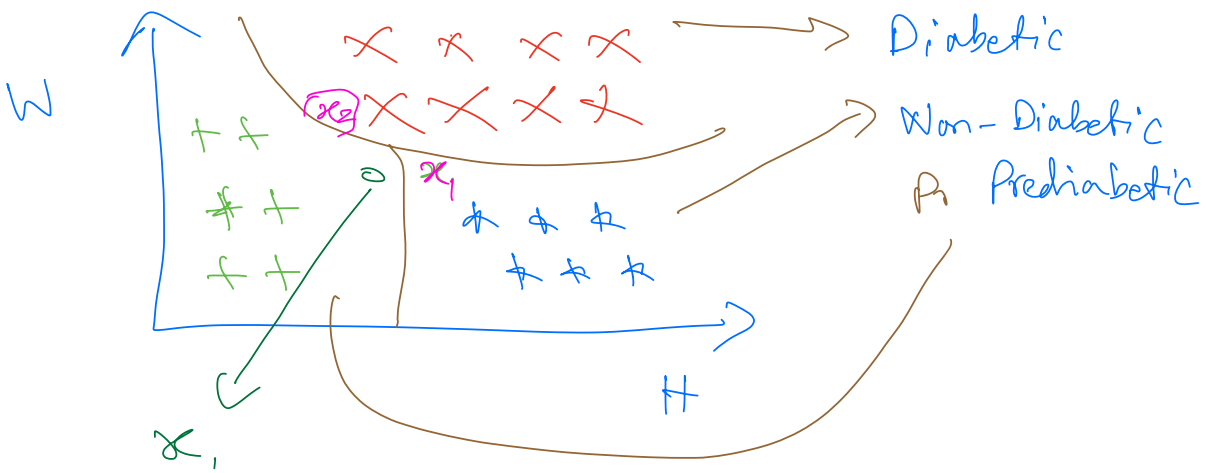
$p(x_1 = d \mid h_1, w_1) = 0.9$

$p(x_1 = nd \mid h_1, w_1) = 0.1$

2 categories → Binary classification

More than 2 categories → ?

Multi-class classification



$$P(x_1 = d \mid H_1, W_1) = 0.3$$

$$P(x_1 = nd \mid H_1, W_1) = 0.6$$ ↳ non-diabetic

$$P(x_1 = pd \mid H_1, W_1) = 0.1$$

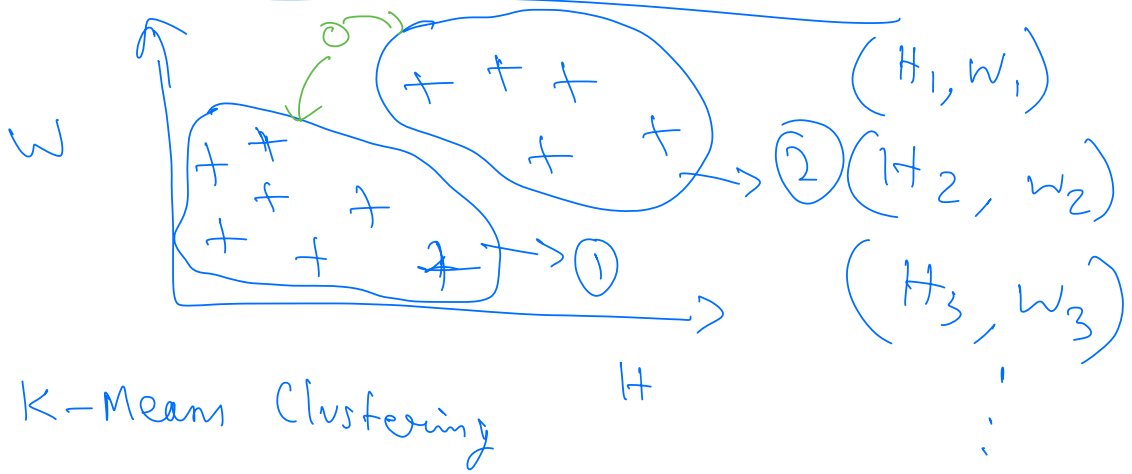
$P(x_2 = d \mid H_1, W_1) = 0.7$

$nd = 0.2$

$pd = 0.1$

↳ Conditional probability

Unsupervised Learning



K-Mean Clustering

Predicting a new value / label

↓

infinite range of possibilities

Clustering (x_i, y_i)

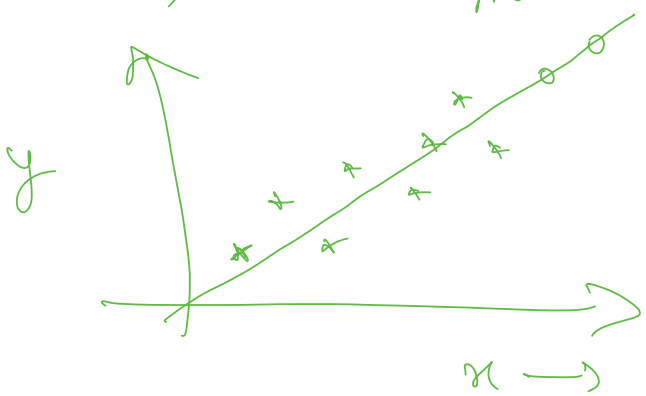
x_1
x_2
x_3

✗ Classification

✓ Regression

Regression

→ Linear Regression

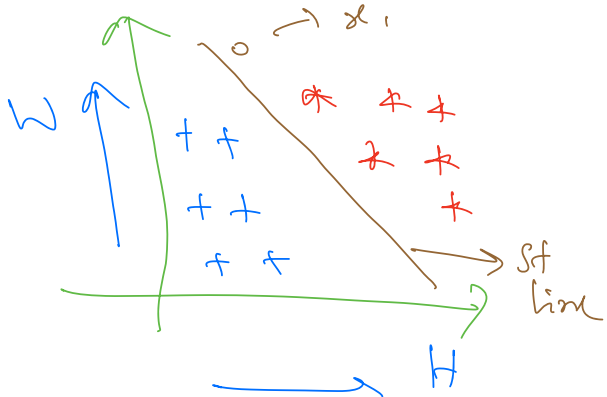


$y = mx + c$

$y_i = mx_i + c$

Classification

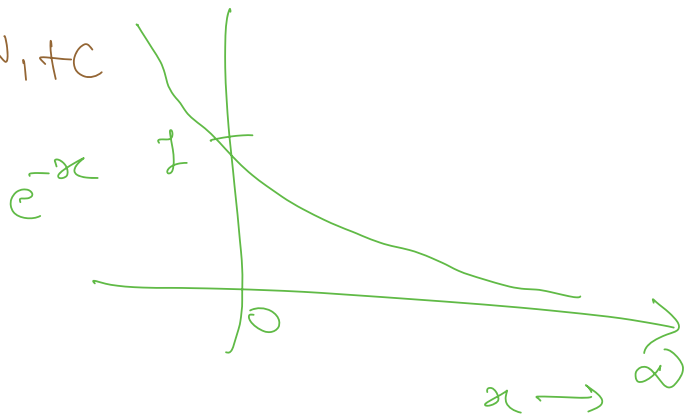
Logistic Regression



$P(x_1 = d \mid H_1, W_1)$

↳ $\frac{1}{e^{-(m_1 H_1 + m_2 W_1 + C)}}$

$y_i = m_1 H_1 + m_2 W_1 + C$



fake real

↓

classification

↓

Logistic Regression