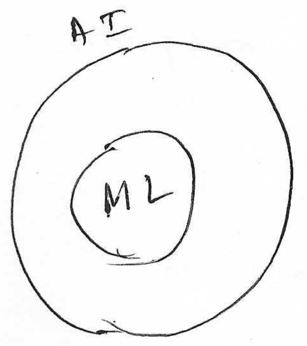
Introduction



$\boxed{AI}$ ⇒ The broad goal
Building machines that can perform tasks that normally require human intelligence.
Ex: Reasoning & problem solving, understanding language, acting in an environment (robots, agents).

$\boxed{ML}$ ⇒ A subset of AI.
Techniques that allow machines to learn patterns from data instead of being explicitly programmed.
Ex: Learning to classify images, translating languages, recommender systems.

[ALL ML is AI, but not all AI is ML]

Before ML dominated:
- AI systems were rule based.
- Ex: medical expert systems
  If symptoms A AND B → disease X.
- Problems: hard to scale, fragile, required human experts to encode knowledge.

↓

ML shift: Instead of rules, learn from examples.
Ex: Given 1 million labelled cases, learn disease prediction.

What is learning?
- The acquisition of knowledge or skills through experience, study, or being taught.

What is ML?
- Field of study that gives computers the ability to learn without being explicitly programmed.

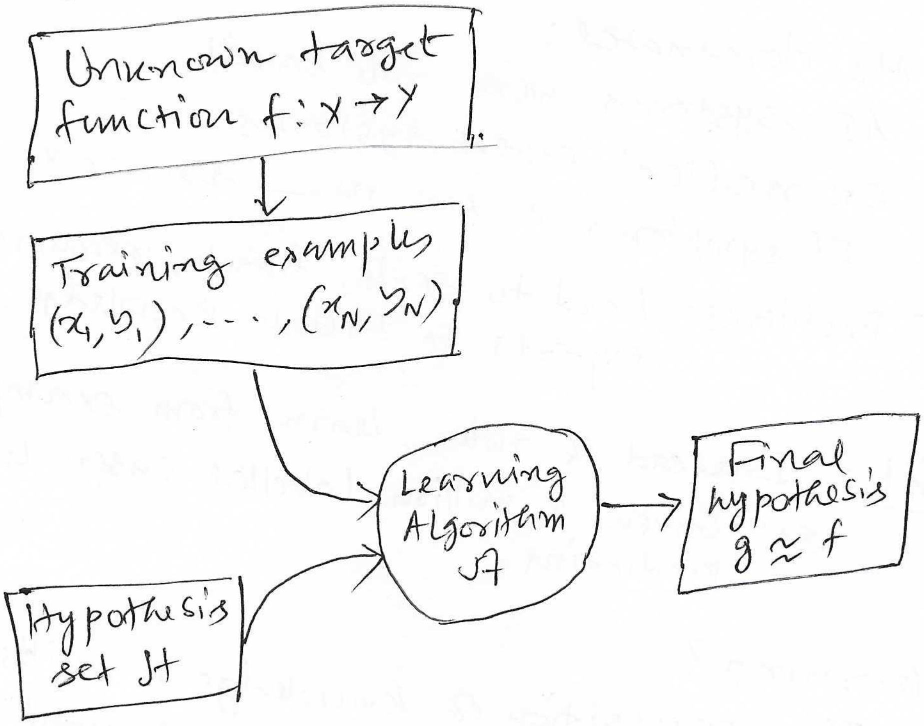
Types of Learning

- Supervised : training data includes desired outputs.
- unsupervised : " does not include " "
- weakly/semi supervised : " " includes a few " "
- Reinforcement : Rewards from sequence of actions.

Supervised learning

Given examples of a function $(x, f(x))$, predict the function $f(x)$ for new examples. The function can be:

- Discrete → classification (e.g. spam/not spam)
- Continuous → regression (housing price, weather)
- probability → probability estimation (e.g. patient symptom)



Training vs Testing

What do we want? good performance on training data?
- good performance on unseen test data.
- No, good performance on unseen test data.

Training data → given to us for Learning the function f

Testing data → used to see if you have learnt anything

Usually a dataset is split into train & test

train → annotate

test → Check performance.

Probability Review.

Consider a nondeterministic event A (boolean variable)

What does $P(A)$ mean?

Frequentist view: Limiting frequency of a repeating non-deterministic event.

$$\lim_{N\to\infty} \frac{\#A \text{ is true}}{N}$$

Bayesian view: $P(A)$ is your "belief" about A.

Axioms of probability.

- $0 \leq P(A) \leq 1$
- $P(\text{empty set}) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Sample space → space of events

Random Variables → Mapping from events to numbers (set of possible values from a random experiments)

discrete, continuous.

Probability distribution → List all possible outcomes of a random variable along with their corresponding values. Ex: $\{H, T\}$, $\{1: 1/6, 2: 1/6, \ldots, 6: 1/6\}$

Probability Mass function (PMF) → Probability that a discrete random variable equals a specific value.

$$P(X = x) = p(x)$$

Probability Density function (PDF) → How dense probability is around values (in the continuous case, probability at an exact point is zero).

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Expectations → Long run average outcome (over many repeatations). weighted average, where weights are probabilities.

$$E_p[f(x)] = \sum_{x=1}^{k} p(x) f(x) \quad \text{(discrete)}$$

$$= \int_{-\infty}^{\infty} p(x) f(x) dx \quad \text{(continuous)}$$

Joint distribution (in discrete space) & Marginalization:

Ex: Suppose we observe 2 random variables.

W = weather : Rainy, Sunny

U = whether a person carries an umbrella : Yes, No.

Qn: what is the probability of a specific combination of weather & umbrella choice?

Ex:- rainy AND carries an umbrella

- sunny AND does NOT carry an umbrella

Let's say after observing many days, we get:

W	U	P(W, U)
Rainy	Yes	0.4
Rainy	No	0.1
Sunny	Yes	0.1
Sunny	No	0.1

← This table is the joint distribution P(W, U)

Marginalization

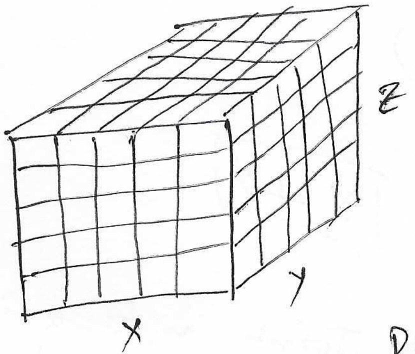
Suppose we ask what is the probability that it is rainy, regardless of the umbrellas?

Then we marginalize out (ignore) the umbrella variable.

P(W= Rainy) = P(Rainy, Yes) + P(Rainy, No) = 0.4 + 0.1 = 0.5

similar for P(W= Sunny) as well.

This is called marginal distribution of weather.



$P(X,Y)$

$\downarrow$ marginalize out Y)

$$P(X=x) = \sum_y P(X=x, Y=y)$$

$\uparrow$ Sum of all values where $X=x$ occurs with all possible values of Y

$P(X,Y,Z)$

$\downarrow$ We want to compute $P(X)$

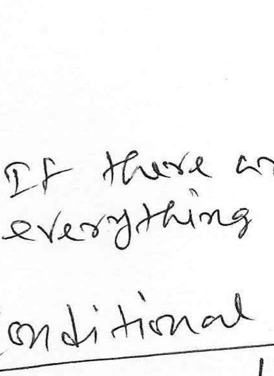
$$P(X,Y) = \sum_z P(X,Y,Z)$$

$$P(X) = \sum_y P(X,Y)$$

If there are n variables, we want marginalization of everything other than a single variable, we need $(n-1)$ summations.

<u>Conditional probabilities</u>:

$P(X=x \mid Y=y)$: What do you believe about $Y=y$, if I tell you $X=x$?



$p(X,Y|Z)$ (conditioning on a variable is taking that slice out)

$$p(x,y \mid Z=z) = \frac{p(x,y,z)}{p(z)} \leftarrow \frac{\text{Joint}}{\text{Marginal}}$$

<u>Chain rule</u>

$$P(Y=y \mid X=x) = \frac{P(Y=y, X=x)}{P(X=x)}$$

$$\Rightarrow P(Y=y, X=x) = P(Y=y \mid X=x)\, P(X=x)$$

For d-dimensional case of joint distribution, we need to recursively apply the chain rule:

$$P(X_1=x_1, \dots, X_d=x_d) = P(X_2=x_2, \dots, X_d=x_d \mid X_1=x_1)\, P(X_1=x_1)$$

$$= P(X_3=x_3, \dots X_d=x_d \mid X_2=x_2, X_1=x_1)\, P(X_2=x_2 \mid X_1=x_1)\, P(X_1=x_1)$$

$$\vdots$$

$$= \prod_{j=1}^{d} P(X_j=x_j \mid X_1=x_1, \dots, X_{j-1}=x_{j-1})$$

Conditional independence

$$p(Y=y, X=x) = p(Y=y \mid X=x)\, p(X=x)$$

India wins World cup

today is sunny

$\boxed{X \perp Y}$

$\Downarrow$

$$= p(Y=y)\, p(X=x)$$

← Joint is factorized into marginal.

Baye's rule

From chain rule,

$$p(Y=y, X=x) = p(Y=y / X=x)\, p(X=x)$$

↓ swap

$$p(X=x, Y=y) = p(X=x / Y=y)\, p(Y=y)$$

$$\Rightarrow \boxed{p(Y=y / X=x) = \frac{p(X=x / Y=y)\, p(Y=y)}{p(X=x)}}$$

Prior (what do you believe before seeing any data)

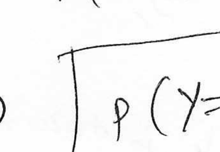
posterior (what do you believe of the classes, after you see the data)

Likelihood (how much does a certain hypothesis explain the data)

Normalization

Entropy: Measures amount of uncertainty in a distribution.

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$



low entropy

high entropy.

KL divergence: An asymmetric measure of distance between two distributions.

$$KL[P \| q] = \sum_x P(x)[\log P(x) - \log q(x)]$$

$KL > 0$ unless $p = q$ (in that case $KL = 0$)