MDL

Grading :

Assn × 4	:	25%	(4+7+7+7)
Quiz × 2	:	15%	(7.5+7.5)
Mid sem	:	25%	
End sem	:	35%	

Instructor :

Prof. Sujit Gujar +
Prof. Praveen Paruchuri

Textbooks :

1. Python ML by Example by Yuxi Liu
2. AI : A Modern Approach (4th Edition) by Russel and Norvig

04/01/2025

Future Technology

* Space Elevator in Japan, Metaverse
* Cloud-computing, IoT, Blockchains, High speed Internet (5G)

↓ ENABLES

Analytics/IT for agri, Smart City/Home, Intelligent healthcare, E-governance

* Data-driven ML plays a vital role

Artificial Intelligence

AI is the sci & engg of making machines intelligent & enabling them to take actions & decisions that humans typically do.

Machines

* A machine is a physical system that uses power to apply forces and control movement to perform an action making lives easier.
* Complex machines are combinations of 2 or more simple machines. Eg: sewing machines, cars, cranes, computers, robots etc.

Future of AI

"Robots will keep us safer, especially from disasters." — Pieter Abeel

"Very smart computers could solve all our problems including climate change." — Stuart Russell

Is AI a threat?

* During Industrialisation, machines replaced a lot of human jobs. Comps replaced lot of accountants/book-keeping
* (Check Ray Zinn's opinion on Forbes in May 2017)

Proliferation of AI

* Autonomous driving, robotics, Llama, computer vision.
* In general, AI assisted decision-making
* Used in healthcare (diagnostic, personalised medicines), finance (fraud detection, algorithmic trading) & daily life (online shopping, digital assistants etc)

04/01/2025

AI for Social Good

* With the global population inc (7.7 – 10B by 2050), there is an inc in agri production to meet food demand.
* Computer vision, data analytics for farmers, efficient supply chain management, AI for public & wild life safety, use of Game Theory, ML & AI, predictive diagnostics, IoT based framework.

What is Intelligence?

The ability to understand, learn and think. Also the capacity for abstraction, logic, understanding, self-awareness, learning, problem solving, emotional knowledge, reasoning, planning, creativity, critical thinking

Intelligence in Humans: We think, learn from observations, act/make decisions.

Two dimensions to AI → always doing the right thing

(i) Human v/s Rational (ii) Act v/s Think

Acting Humanly ← Turing Test	Machines that act human	Machines that act rationally → Agent based modelling: Agent should operate autonomously, perceive the environment, adapt to change
Cognitive modelling ←	Machines that think like humans	Machines that think rationally → Right-thinking, logic-based approach. Need to involve probability to model uncertain things

Acting Humanly Turing Test

A human interrogator should not be able to distinguish between a computer & human after posing some written questions. Total Turing Test involves interaction w/ objects, people, moving objects around etc. Additionally need computer vision and robotics.

Capabilities required by computers: NLP, Knowledge representation, Automated reasoning, Machine learning

04/01/2025

What is Machine Learning?

Construction of algorithms that can learn from and make predictions on data. Data is similar to features.

* Supervised: Classification, Regression
* Unsupervised: Clustering, Anomaly detection

Cricket ball v/s Tennis ball

Can be solved using Perceptron (Look up Perceptron: Threshold Activation)

$$y = f(z) = \begin{cases} 1 & ; z \geq 0 \\ 0 & ; z < 0 \end{cases}$$

$$f(z) = w_0 + w_1 \underbrace{x_1}_{\text{weight}} + w_2 \underbrace{x_2}_{\text{diameter}}$$

Let $w_0 = -100$

Here $w_2 = 0$

Neural networks and Deep Learning

Neural networks replicate neurons in human brains.

Complexity of neural networks:

(i) PaLM – 540B parameters (ii) MT-NLG – 530B parameters

(iii) GPT3 (May 2021) – 175B parameters (iv) GPT4 (March 2023) – 1.76T parameters

* Intelligence of humans is achieved not by purely reflex mechanisms but by processes of reasoning that operate on internal representations of knowledge.
* AI is about building machines and systems that can reason, learn & solve problems similar to how humans do. In many AI applications, logical reasoning plays a crucial role allowing systems to make decisions based on given conditions & facts.
* Leads to knowledge-based agents approach for AI.

Knowledge base is a set of sentences represented in knowledge representation language.

↳ expressing knowledge explicitly in a computer-tractable way.

08/01/2025

Logic

It is the basis of all mathematical reasoning & automated reasoning

Formal	Informal
* Uses formal language	* Natural language
* Study of logical truths	* Studies informal fallacies, critical thinking
* Symbolical & mathematical logic	

Why formal languages?

Natural languages exhibit ambiguity which makes reasoning difficult and incomplete

Eg: (i) Our shoes are guaranteed to give you a fit.

(ii) The boy saw a girl with a telescope

Formal languages promote rigour and thereby reduce possibility of human error; help reduce implicit/unstated assumptions by removing familiarity with subject matter; and achieve generality due to possibility of finding alternative interpretations for sentences & arguments.

Formal logic

Uses syllogisms to make inferences & examines how conclusions follow from premises based on the structure of arguments.

→ Symbolic logic: Uses symbols to accurately map out valid & invalid arguments.

→ Mathematical logic: Uses math symbols to prove theoretical arguments

→ Propositional logic

Why is it important?

* Logic is the core of AI

→ possibility of automating reasoning draw inferences on knowledge

→ answer queries

→ discover facts that follow from the knowledge base

* In AI, propositional logic is essential for knowledge representation, reasoning and decision making processes.

08/01/2025

Impact of Logic in AI

1. "Computer Math Proof Shows Reasoning Power" — Gina Kolata (Dec 10, 1996)
2. AI Robot Scientist helps in finding cure for malaria.

Logical arguments

(A) All humans have 2 eyes.
(B) Sujit is a human.
] (P) Sujit has 2 eyes.
(correct deduction and holds logically)

(C) All humans have 4 eyes.
(D) Sujit is a human.
] (Q) Sujit has 4 eyes.
(correct deduction but logically incorrect)

(E) All humans have 2 eyes.
(F) Kishore has 2 eyes.
] (R) Kishore is a human.
(not necessary)

(G) No human has 4 eyes.
(H) Kishore has 2 eyes.
] (S) Kishore is not human
(not necessary, no connection btwn Kishore & human)

Fallacy: Conclusion may be correct. The reasoning is incorrect.

Propositional logic

Deals with propositions which are either true or false. Also known as propositional calculus or zero-order logic.

Order of 'Logic'

The order of a logic refers to the degree of quantification that can be performed over sets:

First order: Quantities only over individuals. Also known as predicate logic/calculus or quantificational logic.

Second order: Quantities over sets.

Third order: Quantities over sets of sets.

Higher order: Union of first, second, third & higher order logic. Allows quantification over sets that are nested arbitrarily deep.

* Propositions are declarative statements.

08/01/2025

* **Atomic propositions**:
→ simple indivisible statements, cannot be broken down further.
→ each of them represents a basic fact or condition.
→ Eg: "The door is closed."

* **Compound propositions**:
→ Multiple atomic propositions combined using logical connectives (AND, OR, NOT)
→ Eg: "The door is closed and the heater is on."

AND	$S: P \wedge Q$	T if P & Q are T
OR	$S: P \vee Q$	T if any of P, Q is T
NOT	$S: \neg P$	T if P is F
IMPLIES	$S: P \rightarrow Q$ $S: \neg P \vee Q$	T if P implies Q
IFF	$S: P \leftrightarrow Q$ $S: \neg P \text{ XOR } Q$	T if P & Q are T or F together

(i) $P \rightarrow Q$ is false if P is True and Q is false.
(ii) $P \rightarrow Q$ is True if P is a sufficient condition for Q and Q is a necessary condition for P.
(iii) $P \leftrightarrow Q$ is True ⇒ P and Q are equivalent. P & Q are necessary and sufficient conditions for each other.

Truth table: Breakdown of all the possible truth values returned by a logical expression.

Exercise:
A & B are True; X & Y are false:
(i) $\neg(A \vee X)$ = false
(ii) $A \vee (X \wedge Y)$ = True
(iii) $A \wedge (X \vee (B \wedge Y))$ = false
(iv) $(X \wedge Y) \wedge (\neg A \vee X)$ = false
(v) $[(A \wedge X) \vee \neg B] \wedge \neg [(A \vee X) \vee \neg B]$ = false
(vi) $[(X \wedge Y) \rightarrow A] \rightarrow [X \rightarrow (Y \rightarrow A)]$ = True

Tautology: A sentence is valid if it is True under all possible assignments of True or false to its propositional variable (Eg: $P \vee \neg P$)

11/01/2025

SAT: A sentence is satisfiable iff there is some assignment of T/F to its propositional variables for which the sentence is True. A sentence is unsatisfiable iff it is not satisfiable (Eg: $P \wedge \neg P$)

How to write a proof?

Step 1 Intend to formally capture the notion of proof that is commonly applied in other fields.

Step 2 A proof of a formula from a set of premises is a sequence of steps in which any step of the proof is:
→ an axiom or premise
→ a formula deduced from previous steps of the proof.
→ using some rule of inference

Step 3 The last step of the proof should deduce the formula we wish to prove.

Step 4 We say that S follows from α (premises) to denote that the set of formulae α prove the formula S.

Semantic notion: It concerns the notion of Truth.
$\alpha \rightarrow \beta$ (whenever all the formulae in set α is T, β is also T)

Soundness, Completeness and Decidability

* A logic is sound if it preserves the truth.
→ If a set of premises are all T, any conclusion drawn from those premises must also be T.
* A logic is complete if it is capable of proving any valid consequence.
→ (Syntatic) completeness requires every sentence or its negation to be provable.
* A logic is decidable if there is a mechanical procedure to prove any given consequence in the theory.
→ There should be a decisional finite procedure for the theorem.

11/01/2026

Inference rules:

(i) Modus Ponens: $P, P \to Q \Rightarrow Q$.

(ii) Modus Tollens: $P \to Q, \neg Q \Rightarrow \neg P$

(iii) Hypothetical syllogism: $P \to Q, Q \to R \Rightarrow P \to R$

(iv) And-elimination: $P_1 \wedge P_2 \wedge \ldots \wedge P_n \Rightarrow P_i$

(v) And-introduction: $P_1, P_2, \ldots, P_n \Rightarrow P_1 \wedge P_2 \wedge \ldots \wedge P_n$

(vi) Or-introduction: $P_i \Rightarrow P_1 \vee P_2 \vee \ldots \vee P_n$

(vii) Double-negation elimination: $\neg\neg P \Rightarrow P$

(viii) Unit resolution: $P \vee Q, \neg Q \Rightarrow P$

(ix) Resolution: $P \vee Q, \neg Q \vee R \Rightarrow P \vee R$

Eg: 1. α: $\{A \vee (B \to D), \neg C \to (D \to E), A \to C, \neg C, \neg A, B \to D\}$

β: $B \to E$

(i) $\neg C$ (ii) $\neg C \to (D \to E)$

(iii) $D \to E$ [Modus Ponens; (i), (ii)]

(iv) $B \to D$ (v) $B \to E$ [Hypothetical syllogism; (iv), (iii)]

2. α: $\{C \to N, N \to I, I \to S, (C \to S) \to (N \to C), \neg C\}$

β: $\neg N$

(i) $C \to N$ (ii) $N \to I$ (iii) $I \to S$

(iv) $C \to I$ [Hypothetical syllogism; (i), (ii)]

(v) $C \to S$ [Hypothetical syllogism; (iv), (iii)] (vi) $(C \to S) \to (N \to C)$

(vii) $N \to C$ [Modus Ponens; (v), (vi)] (viii) $\neg C$

(ix) $\neg N$ [Modus Tollens; (vii), (viii)]

3. If I study, I make good grades. If I do not study, I enjoy myself. Therefore, either I make good grades or I enjoy myself.

α: $\{S \to G, \neg S \to E\}$

β: $G \vee E$

S ≡ I study

G ≡ I make good grades

E ≡ I enjoy myself

(i) $S \to G$ (ii) $\neg S \to E$

(iii) $\neg S \vee G$ (iv) $\neg\neg S \vee E$ (v) $S \vee E$

(vi) $G \vee E$ [Resolution; (v), (iii)]

11/01/2025

Conjunctive Normal Form (CNF)

* Literal: propositional letter or the negation of a propositional letter.
* Clause: Disjunction of literals
* CNF: Conjunction of clauses. Eg: $(P \vee Q \vee \neg R) \wedge (\neg S \vee \neg R)$
* Disjunctive Normal form (DNF): Disjunction of conjunctions of literals.
  Eg: $(P \wedge Q \wedge \neg R) \vee (\neg S \wedge \neg R)$
* Every propositional logic formula can be converted to CNF & DNF

Converting to CNF

* Rewrite $P \leftrightarrow Q$ as $(P \rightarrow Q) \wedge (Q \rightarrow P)$.
* Rewrite $P \rightarrow Q$ as $\neg P \vee Q$.
* Use De Morgan's laws:
  - Rewrite $\neg(P \vee Q)$ as $\neg P \wedge \neg Q$.
  - Rewrite $\neg(P \wedge Q)$ as $\neg P \vee \neg Q$.
* Rewrite $\neg\neg P$ as P.
* Use distributive laws:
  - Rewrite $(P \wedge Q) \vee R$ as $(P \vee R) \wedge (Q \vee R)$
  - Rewrite $(P \vee Q) \wedge R$ as $(P \wedge R) \vee (Q \wedge R)$

Eg: Convert $\neg(P \rightarrow (Q \wedge R))$ to CNF.

$\neg(P \rightarrow (Q \wedge R)) = \neg(\neg P \vee (Q \wedge R))$

$= P \wedge \neg(Q \wedge R) = \underline{P \wedge (\neg Q \vee \neg R)}$

MACHINE LEARNING

* Scientific study of algorithms and statistical models that computer systems use to:
  → perform a specific task effectively w/o using explicit instructions.
  → rely on patterns & inference instead.
* Involves:
  → building a mathematical model based on sample data known as "training data" and evaluating its performance

15/01/2025

w/ the help of "test data".
→ no explicit programming done to perform the task.

* The term Machine Learning was coined around 1960.
* Lots of patterns for an activity/event. Events can be dynamic, for eg, automated cars. Data is increasing ~~dynamically~~ exponentially and in various formats. Higher quality data due to cheaper storage.

1940s : Idea of neural networks existed
1960s : Computers w/ reasonable computing ~~power~~ power, simulation of neural networks.
1990s - 2000s : Shift from manual survey data results to datasets sourced from internet for putting into neural nets. Cost effective to train models hence came Machine Learning.

* Can be broadly classified into 3 categories : Unsupervised, Supervised and Reinforcement Learning.

<u>Unsupervised Learning</u>
Takes a set of data that contains only inputs and finds structure in data. Eg: grouping or clustering of data points.

Examples :-
→ <u>Marketing</u> : Finding grps of customers w/ similar behaviour given a large database of customer data containing their properties and past buying records.
→ <u>Biology</u> : Classification of plants and animals given their features
→ <u>Earthquake Studies</u> : Clustering observed earthquake epicenters to identify dangerous zones.
→ <u>World Wide Web</u> : Clustering weblog data to discover groups of similar access patterns.

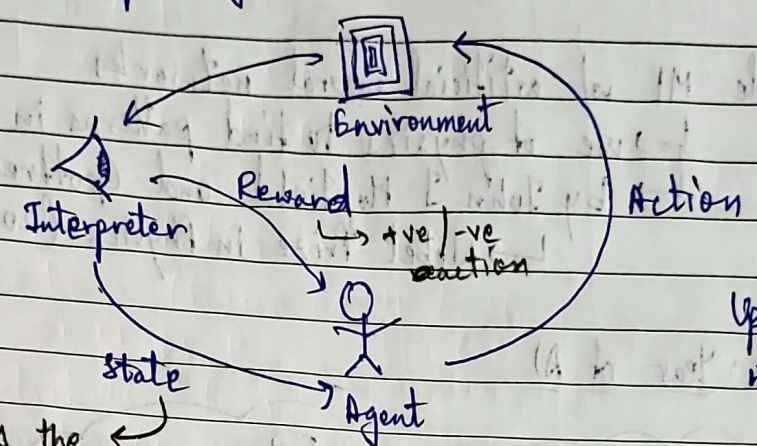
15/01/2025

Supervised Learning

Builds mathematical model using data set that has both inputs and desired outputs. Eg: Classification and Regression tasks.

Reinforcement Learning

Concerned w/ how s/w agents should take actions in an environment to maximise cumulative reward. Eg: autonomous vehicles, computer games etc. Also called real-time learning.



Updates knowledge when feedback is received.

informs the state of the world

Labelled : Input X, Output Y → Supervised Learning
Unlabelled : Input X, Output Y is not given → Unsupervised Learning

Alpha Go

AlphaGo → defeated a 9-dan professional player & World No. 1
- → uses Monte Carlo Tree search algo
- → based on knowledge learned by deep learning method
- → trained on games played by top players, self-play model

~~Alpha Go Zero~~

Alpha Go Zero → defeated AlphaGo (100-0)
- → no manual training, only self-play

Alpha Zero → generalised version of Alpha Go Zero

Mu Zero → masters games w/o knowing rules.

19/01/2025

Protein-folding problem: No. of ways a protein could theoretically fold before settling into its final 3D structure
- ↳ Cyrus Levinthal: $\sim 10^{300}$ possible conformations
- ↳ attempts to solve this using nuclear magnetic resonance and X-ray crystallography..
- ↳ finally solved using an AI Model developed by Demis Hassabis and John Jumper (AlphaFold).
  - ↳ Nobel Prize in Chemistry (2024)

Enable ML w/ artificial neural networks
- ↳ use of physics to find patterns in information
- ↳ by John J Hopfield and Geoffrey Hinton
  - ↳ Nobel Prize in Physics (2024)

2023: Year of AI

Jan
Microsoft → $10B funding in OpenAI

Feb
Google → Bard
Microsoft → Bing chatbot

March
limited access to Bard
Adobe → Firefly
Canva → virtual design assistant
APIs for ChatGPT
Whisper → text-to-speech model
GPT-4

April
Google Research + DeepMind
Sberbank (Russia) → Gigachat
Hugging face
↳ Hugginlhat

May
Bard chatbot
NVIDIA Neuralink
Windows 11: AI assistant

June
Apple → Vision Pro (AI powered AR headset)
EU AI Act
Gen AI can add $4.4T in value to global economy ~ McKinsey

19/01/2025

July

Meta → Llama 2
↳ open-source LLM
↳ trained on publicly available data

can drive other chatbots

MS 365 Copilot → paid service

Anthropic → Claude 2

August

Google Duet AI for Workspace
↳ paid service

OpenAI → Poe

September

Amazon → $4B investment in Anthropic

OpenAI → Canva plug in for ChatGPT

October

Executive order on Safe, Secure & Trustworthy Development and Use of AI

November

xAI (Elon Musk) → Grok

AWS → Amazon Q, Stability AI → Pika 1.0

~~November~~ OpenAI → Sam Altman fired, comes back as a new observer for Microsoft

GENERALISATION AND GOOD FIT

Generalisation: how well the concepts learned by an ML model generalises to specific examples or data not yet seen by the model

Goodness of fit: how well a model fits for a set of obs.
→ Overfitting and Underfitting

Training data: 70% of the original dataset collected is used to build & train the model.

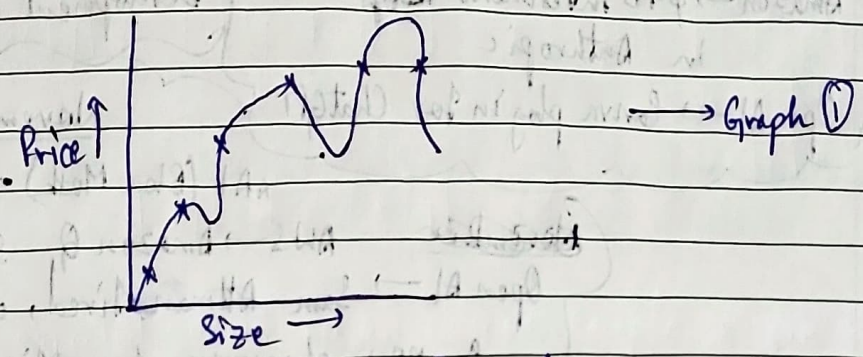
Test data: Remaining 30% of the dataset is used to analyse performance of models.

OVERFITTING

Phenomenon of extracting too much information from training sets or memorisation can cause overfitting.

22/01/2025

* Low bias: as the model works well w/ training data
* High Variance: as it does not generalise well or derive patterns and performs poorly on test datasets.
* Ends up making predictions based on noise if the algo is too complex or too flexible.
* A well functioning ML algo seperates signal from noise.
* Extreme form of overfitting: memorisation
* When you use too few examples or have too complex features.



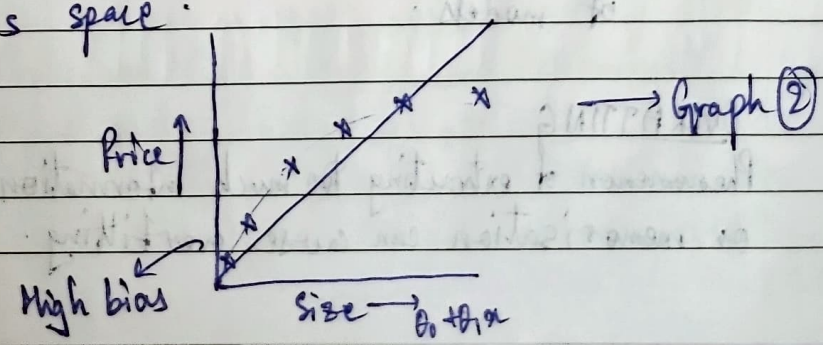
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

↳ High variance

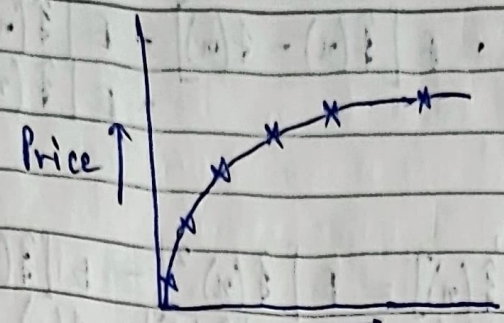
UNDERFITTING

Model is underfit if it does not perform well on training sets and hence will not do so on test sets.

* High bias: as the model does not work well w/ training data
* Low variance: as it performs consistently w/ training data as well as test data. (and not in a good way here).
* Occurs when enough data is not used to train or the wrong model is tried to fit to the data.
* May need to increase no. of features since it expands the hypothesis space.



22/01/2025



→ Graph ③

$\theta_0 + \theta_1 x + \theta_2 x^2$

↳ Just right

Graph ①, ② & ③ use the same data.

BIAS - VARIANCE TRADEOFF

* Model is too simple & has very few parameters → high bias & low variance
  Model has large no. of parameters → high variance & low bias
* Good balance required
* As more parameters are added to a model
  → Complexity of the model rises
  → Variance becomes primary concern while bias falls steadily.

Bias → error due to incorrect assumptions in learning algo
Variance → error due to small fluctuations in training set.

Suppose a training set consists of points $x_1, x_2, \ldots, x_n$ and real values y_i associated w/ each point x_i.
Assume a function $y = f(x) + \varepsilon$
where noise ε → zero mean, variance σ^2

$\hat{f}(x)$ → approximates $f(x)$ as well as possible.
↳ generalises to points outside of our training set.

$$Var(X) = E[(X-\mu)^2] = E[(X-E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2]$$
$$= E[X^2] - 2E[X]E[X] + E[X]^2$$
$$= E[X^2] - E[X]^2$$

29/01/2025

$$\text{Bias}(\hat{f}(x)) = E[\hat{f}(x) - f(x)] = E[\hat{f}(x)] - E[f(x)]$$
$$= E[\hat{f}(x)] - \underbrace{f(x)}_{\text{as } f \text{ is deterministic}}$$

So we have,

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$$
$$\text{Bias}(\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$

Given $y = f + \varepsilon$, $E[\varepsilon] = 0$, $\text{Var}[\varepsilon] = \sigma^2$

$$E[y] = E[f+\varepsilon] = E[f] + E[\varepsilon] = E[f] = f$$
$$\text{Var}[y] = E[(y - E[y])^2] = E[(y-f)^2] = E[\varepsilon^2]$$
$$= \text{Var}[\varepsilon] + E[\varepsilon]^2$$
$$= \text{Var}[\varepsilon] = \sigma^2$$

$$\therefore E[y] = f, \ \text{Var}[y] = \sigma^2$$

The expected error on an unseen sample x can be decomposed as:

$$E[(y - \hat{f})^2] = E[(f + \varepsilon - \hat{f})^2] = E[(f + \varepsilon - \hat{f} + E[\hat{f}] - E[\hat{f}])^2]$$
$$= E[(f - E[\hat{f}])^2] + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2]$$
$$+ 2E[(f - E[\hat{f}])\varepsilon] + 2E[\varepsilon(E[\hat{f}] - \hat{f})]$$
$$+ 2E[(f - E[\hat{f}])(E[\hat{f}] - \hat{f})]$$

$$= E[(f - E[\hat{f}])^2] + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2]$$
$$\cancel{+ 2(f - E[\hat{f}])E[\varepsilon] + 2E[\varepsilon]E[E[\hat{f}] - \hat{f}] + 2E[E[\hat{f}] - \hat{f}](f - E[\hat{f}])}$$

$$= E[(f - E[\hat{f}])^2] + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2]$$
$$= \text{Bias}[\hat{f}]^2 + \text{Var}[y] + \text{Var}[\hat{f}]$$
$$= \text{Bias}[\hat{f}]^2 + \sigma^2 + \text{Var}[\hat{f}]$$

$$\therefore \boxed{E[(y - \hat{f})^2] = \text{Bias}[\hat{f}]^2 + \sigma^2 + \text{Var}[\hat{f}]}$$

01/02/2025

How To Avoid Overfitting?

1. Cross-validation

* Conventional validation → training, validation and test set
* The validation ~~used~~ to set is used to experiment & tune parameters of learnt model, compare performances of prediction algorithms and choose the best algo out of them.

* Model evaluation method checks how well a model generalises to a new data.

One round:

→ Data divided into 2 subsets
→ One for training and one for testing
→ Testing performance is recorded

Multiple such rounds are performed under diff partitions. Testing results from all rounds are averaged to generate a more accurate estimate of model prediction performance.

* Non-exhaustive methods

(i) k-fold cross validation

Step1 Original data split into k equal-sized folds.

Step2 In each trial, one fold becomes testing set and rest become training set.

Step3 Repeat the process k times ~~by~~ ~~training~~ ~~or~~ such that each fold is assured to be left out atleast once by the end of the process

(ii) Holdout method

→ Split data into training & testing set numerous times.
→ Issue: some samples may never end up in testing set while others may be selected several times

01/02/2025

* Exhaustive method

i) Leave-out-one cross validation

→ Let each datum be in the testing set once (^nC_k partitions)

→ Dataset of size n so LOOCV needs n rounds of cross-validation

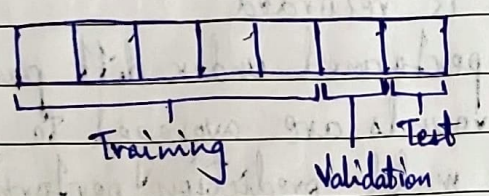
→ Results in a more robust estimate of model performance as each row of data can represent the entirety of test dataset

~~* Exhaustive method involves testing the model in all possible ways by dividing the original data set into training and validation sets using~~

* Nested cross-validation

→ About fine-tuning parameters of an algorithm.

→ One version: k-fold cross validation w/ validation & test set.



Say parameter X needs tuning.
Possible values of X: 10, 20, 30, 40, 50
Here, k = 7

Division of data sets:

Test = 1 set Validation = 1 set Training = (k-2) sets

For picked test set, perform k-fold cross validation on training & ~~test~~ validation data sets ~~(6 dat~~ (k = 6), compute avg training error for each value of X and pick the best X.
Repeat for each possible test set (i.e. 7 times)
Pick X that was returned max times to outer loop.

2. Regularisation

* Unnecessary complexity of model causes overfitting
* Add extra parameters to the error function that is to be minimised in order to penalise complex models.

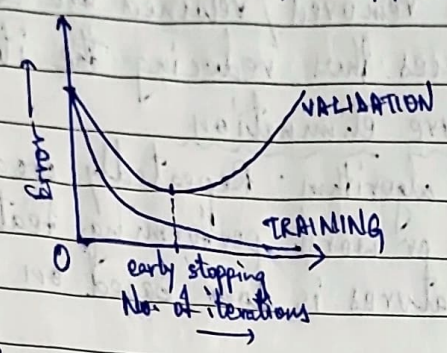
Let $\hat{f}_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

01/02/2025

We want to minimise MSE

where $MSE = \frac{1}{m} \cdot \min_{\theta_0, \theta_1, \theta_2, \theta_3} \sum_{i=1}^{m} (f_\theta(x^i) - y^i)^2$

* Occam's Razor: Preference should be given to the simplest hypothesis that fits the data.
* Stopping training procedure early can also be a form of regularisation. It is called Early Stopping and produces simpler models.



3: Feature selection

* Process of picking a subset of significant features for use in better model construction. Not all features are useful and end up adding randomness to our results.
* For n features, we have 2^n feature sets.
* Either start w/ all features & then iteratively remove them or start w/ no features & then iteratively add them.

Filter methods

Features selected on the basis of their scores in various statistical tests for their correlation w/ their outcome variable.
For example, Chi Square Test is used for categorical variables and Pearson correlation coefficient scores for continuous variables.
Companies adopt algorithms and tailor them for specific tasks.
Most effort goes in data collecting, gathering, cleaning and model testing & development.

most imp step

05/02/2025

If two features are highly correlated and do not influence outcome variable much, then you can remove one of them. However, if one of the above mentioned highly correlated features ~~corr~~ have high correlation w/ the outcome variable as well then the feature must be retained.

Wrapper methods
Features to be removed / retained are decided based on previous inferences thus reducing ~~the~~ it to a search problem
Recursive feature elimination
Follows greedy algorithm. Repeatedly creates models & keeps aside the best or worst performing feature at each iteration. Ranking of features is done based on the order of their elimination.
Essentially 80% of the work can be done w/ 20% of the effort

Embedded methods:
Features are selected based on how best they contribute to the accuracy of the model while the model is being created.

4. Dimensionality reduction
Transform features w/ high dimensions to low dimensional data with minimal loss of information/data.

DATA PREPROCESSING:
Data mining:
* Data mining is predecessor to data science.
* The most popular methodology used is Cross Industry Standard Process for Data Mining (CRISP-DM).

CRISP-DM has the following phases:
→ Business understanding: Business person (domain experts) formulate

08/02/2025

a business problem.

→ Data understanding: Technical analyst gets involved w/ domain experts to understand complicated data. Also called exploration phase.

→ Data preparation: Creation of training & test data sets. Also called pre-processing phase.

→ Modelling: Formulate a model to fit data and is usually associated w/ ML.

→ Evaluation: Evaluate the model & data to check if the business problem is solved.

→ Deployment: Setting up the system in a production environment.

* There is no one way of cleaning data, often involves manual work.
* Data can be of various types → grid of numbers, audio, text etc.
* For the purpose of visualising data, it is convenient to have numerical features.

FEATURE ENGINEERING

More art than science. Features can be of following types:

* Distributions: normal, binomial, poisson etc.
* Binaries: yes/no, +/-, true/false etc.
* Categorical: instance of continents, high/medium/low etc.
* Quantitative: temperature in degree, price in dollars etc.

It is the process of creating or improving features.

Features are created based on common sense, domain knowledge and prior experience.

Data can have missing values for some features. In such cases, the following happens:

→ the missing values are ignored.

→ the missing values are imputed w/ fixed values (can be arithmetic mean, median or mode).

Humans work w/ various types of values but ML algorithms

12/02/2025

work best w/ numerical values. Hence label encoding is done to convert labels into * numeric form.

One-hot-encoding or one-of-K

Refers to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains '0' or '1' corresponding to which column it has been placed.

Example: Original data:

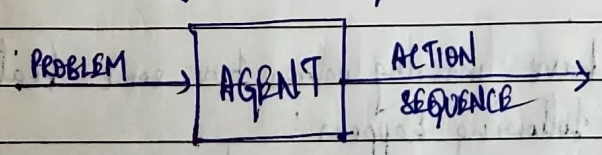
Fruit	Categorical value of fruit	Price
Apple	1	5
Mango	2	10
Apple	1	15
Orange	3	20

After one-hot encoding:

Apple	Mango	Orange	Price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

Search algorithm

Process of looking for a sequence of actions that reach the goal.



19/02/2025

DATA SCIENCE

It is an interdisciplinary academic field that uses statistics, scientific computing, methods, processing & visualisations, algorithms and systems to extract or extrapolate knowledge from potentially noisy, structured or unstructured data.

Raw data			Business insights
* Sales data * Customer feedback * Web logs	→	Data Science Process	→ * Sales prediction * Product optimisation * User behaviour

DATA MINING

* Process of discovering valuable insights, patterns & information from vast datasets by employing various data sets by employing various techniques, algos and tools.
* Technology that blends traditional data analysis methods w/ sophisticated algos for processing large volumes of data.

Statistics: sampling, estimation, hypothesis testing

Machine Learning: search algo, modelling techniques, learning theories from AI, pattern recognition

Types of patterns:

(i) Association: Coffee buyers usually also buy sugar.

(ii) Clustering: Segments of customers requiring different promotional strategies

(iii) Classification: Determining if a bank customer who is applying for a loan will be a defaulter.

Association

Identifies relationships between items in a data set.

19/02/2023

Example:
1. Bread Butter Milk
2. Bread Butter Salt
3. Bread Butter Milk
4. Bread Butter Sugar

We can cluster 1 & 3 together.

Bread → Butter 100%
(Bread, Butter) → Milk 50%
(Bread, Butter) → Salt 25%
(Bread, Butter) → Sugar 25%

~~Rules~~: Association rules:
~~Those~~ There's always an antecedent & a consequent in an association rule.

No. of association rules possible $= \sum_{i=1}^{n} {}^{n}C_i \left(2^{n-i} - 1\right)$

AR: $X \rightarrow Y$

$\{x_1, x_2, \ldots, x_n\} \rightarrow \{y_1, y_2, \ldots, y_m\}$

* Itemset(AR) = list of all items in the antecedent & consequent
$= X \cup Y$

* Support (AR) $= P(X \cap Y)$
$= \dfrac{\#\text{ transactions containing } X \cap Y}{\#\text{ transactions in database}}$

* Confidence (AR) $= P(Y \mid X)$
$= \dfrac{\#\text{ transactions containing } X \cap Y}{\#\text{ transactions in } X}$

* Lift (AR) $= \dfrac{P(Y \mid X)}{P(Y)}$
$= \dfrac{\dfrac{\#\text{ transactions containing } X \cap Y}{\#\text{ transactions in } X}}{\dfrac{\#\text{ transactions in } Y}{\#\text{ transactions in database}}}$

If Lift < 1, then you can ignore it.

19/02/2025

Applications: Market basket analysis, recommendation system, fraud detection, healthcare & medical

Example:

Transaction ID	Items
1	tomato, potato, onion
2	tomato, potato, brinjal, pumpkin
3	tomato, potato, onion, chilli
4	tamarind, lemons.

AR: (tomato, potato) → onion

⋮

etc.

Itemset: {tomato, potato, onion, brinjal, pumpkin, chilli, tamarind, lemons}

Apriori algorithm

Invented by Rakesh Agrawal & Ramakant Srikanth (1994)

Apriori: acknowledges the prior knowledge

→ If any itemset is not frequent, then its superset cannot be frequent

→ An itemset is frequent only if all its subsets are frequent.

Step 0: Create 1-size frequent itemsets list that meet threshold support $k=1$.

Step 1: Expand the itemsets list by combining overlapping sets from k-sized itemsets to k+1-sized itemsets list.

Step 2: Prune the expanded itemsets list using apriori property, $k=k+1$.

Step 3: Remove infrequent item sets from the list.

Repeat steps 1, 2, 3 till no ~~more~~ further expansion is possible.

Example:

1	1,2,3,4	5	2,3
2	1,2,4	6	3,4
3	1,2	7	2,4
4	2,3,4		

22/02/2025

k=1	{1}	{2}	{3}	{4}		
	3	6	4	5		
k=2	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
	3	1	2	3	4	3
k=3	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}		
	1	2	1	2		
k=4	{1,2,3,4}					
	1					

Frequent itemsets : {1}, {2}, {3}, {4}, {1,2}, {2,3}, {2,4}, {3}

Pseudocode :

```
Apriori (T, ε) {
    C = All '1'-sized itemsets
    k = 1
    while (|C| > 0) {
        find frequency of all itemsets in C
        L = sets in C w/ frequency count ≥ ε

        C = Apriori Generate (L, k)
        k++
    }
}
Apriori Generate (L, k) {
    ∀ x, y ∈ L
        if |X ∩ Y| = k-1 :
            Z = X ∪ Y
        if any (k-1)-sized subset of Z not in L :
            do not add Z to L
        else :
            add Z to L
    return L
}
```

22/02/2025

* Different types of association rules (categorical, hierarchical, cyclic).
* <u>Eclat algorithm</u>
  Equivalent Class Transformation: a DFS strategy
* <u>FP growth</u>
  Frequent Pattern growth: a compact data structure called the frequent Pattern tree is used to compress the dataset

<u>Classification</u>

We collect different requirements/facts/data about certain features

$X = (x_1, x_2, \ldots, x_d) \subseteq \mathbb{R}^d$ → data set

$Y = \{1, 2, \ldots, K\} = [K]$ → class set (label space)

$K = 1$ → no need for classification

$K = 2$ → binary classification

$K > 2$ → multi-class

<u>What is a classifier?</u>

x_{new} $\quad y_{new}$ $\quad \hat{y}_{new}$

y_{new} → True label

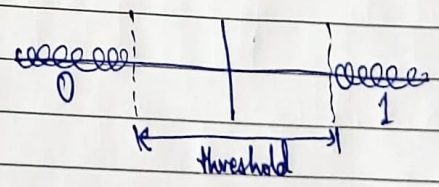
$\hat{y}_{new}$ → Predicted label

Classifier is a function

$$f: X \to Y$$

<u>Example</u>: $f(x) = 1$; $x \geq$ threshold

0 ; otherwise



Goodness of a classifier: $P(\hat{y}_{new} = y_{new})$

$$\text{Accuracy} = \frac{\#\ x_i \text{ correctly classified}}{L}$$

22/02/2025

y_i \ $\hat{y}_i$	0	1
0	TN	FP
1	FN	TP

$P = FN + TP$

$N = TN + FP$

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

$$\text{Recall (also called sensitivity)} = \frac{TP}{P} = \frac{TP}{FN + TP}$$

→ also called True Positive Rate (TPR)

$$\text{Precision} = \frac{TP}{TP + FP} \longrightarrow \text{Positive Predictive Value (PPV)}$$

$$FNR = \frac{FN}{P} \qquad TNR = \frac{TN}{N} \qquad FPR = \frac{FP}{N}$$

$$TNR + FPR = 1 \qquad FNR + TPR = 1$$

F_1 score = harmonic mean of recall & precision

$$= \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = \frac{2}{\frac{FN+TP}{TP} + \frac{FP+TP}{TP}} = \frac{2TP}{2TP + FN + FP}$$