V-1 (Data Science Honours)

DOMS | Page No. | Date / /

* Need for AI -

- Handling massive data
- Automating tasks
- Improving accuracy
- Working 24/7
- Innovation across fields
- Enhanced decision making
- Personalized experiences
- Augumented human capabilities.
- Addressing global challenges
- Exploration & discovery

* Applications -

- Image & speech recognition
- Recommendation system
- Fraud detection
- Self driving cars
- Customer service
- Health care
- Manufacturing
- Agriculture

DOMS | Page No.
Date / /

Logic programming (solving problems)

- Based on formal logic
- allows you to solve problems by declaring facts & rules.
- Facts → represent basic truths about world.
- Rules → define relationships b/w facts.

Benefits →
- Declarative (Focus on what to solve)
- Readable → clear code resembling logic statements.
- Flexible → easy to add new facts or rules
- well suited for specific problems

Search

Heuristic Techniques -

- clever way to find good solutions in complex problems
- Weak Methods → (don't apply great deal of knowledge).
- Not domain or problem specific
- Heuristic Function → applied to a state in search space indicate a likelihood of success if state is selected.
- given a search space, a current state & a goal state.
- generate all successor states & evaluate each with our heuristic function.
- select move yields best heuristic value.
- faster than exhaustive search, especially for large problems.
- finds good solutions even if optimal one is hard to find.

eg → game playing (eg. chess), 8-puzzle, navigating robots or in games & scheduling problems.

DOMS | Page No.
Date / /

Constraint satisfaction problem -

→ consists of 3 components - V, D, C

- V: Variable {N, NW, NE, M, MW, ME, S, SE}
- D: Domains.
- C: constraints — adjacents regions must have diff colors.

→ finding soln that meets the constraints.
→ V set of variables $\{V_1, V_2, V_3 \ldots V_n\}$.
→ D is set of domains $\{D_1, D_2, D_3 \ldots D_n\}$ one for each variable.
→ C is set of constrains specify allowable combination of values.
$C_p = (scope, rel)$ $\{C_1 C_2 C_3\}$
→ scope is set of variable that participate in constraint.
→ rel is relation that define values that variable to take.

→ Properties of CSPs -
- specific partial info.
- non-directional
- declarative
- additive
- independent.

→ example - Map Coloring, Sudoku, scheduling problems, configuration problems etc.
→ Variables & Constrains. find an assignment of values to all variables such as all constraints are satisfied.

DOMS | Page No.
Date / /

Local Search Techniques -

- Algorithms that iteratively explore solution space by making incremental changes to current solution, aiming to improve gradually.

Characteristics →
- Iterative Improvement → Adaptability
- Greedy Approach → Scalability
- May not guarantee optimality → Heuristic based.

Types →
- Hill Climbing
- simulated annealing
- Genetic Algorithms

Applications →
- Travel Salesman Problem
- Graph Coloring.
- Scheduling.

Limitations →
- Local optima
- sensitivity to initial solution.
- Difficulty in Escaping local optima.
- Limited Exploration.
- can get stuck in local optima.
- effectiveness depends on problem

Greedy search →
- simple & intuitive algo used to solve optimization problems.
- Hope of finding optimum.

Characteristics →
- Local optimization
- Simpl Implementations
- efficiency
- applicability
- simple & efficient
- good for approx. sol^n.

Limitations →
- No Backtracking
- May not find global optimum
- Depends on Heuristics.
- non optimal for some
- suboptimal sol^n problems.
- problem dependent

DOMS | Page No.
Date / /

Big Data Learning (V2)
Honors Data Science

Characteristics of Big data -

Big data → large vol. of data available at various sources in varying degree of complexity, generated at different speed ie. velocities & technologies, processing modules.

- volume →
  - Big data characterized → enormous value → ~~potentiality~~
  - large data needs specialized tools & technologies to store process & analyze
  - observe & tracks data from various sources.
- velocity →
  - data streams → high speed & dealt timing
  - processing of data → streamed data → real time results → fast
  - insights generated → relevant & actionable.
  - speed of generation of data.
- variety →
  - heterogeneous sources -
  - nature of data → structured & unstructured
  - social media data → sensors, audio, video, etc.
  - data → insights → manage & analyze.
- value →
  - business value → Big Data.
  - generate some sort of value → company doing analysis
  - insights →
    - → formal decisions
    - → optimize operations
    - → gain competative edge in marketplace.
- veracity → (led to)
  - inconsistency / uncertainity in data
  - challenges in data quality & Reliability
  - Insights → accurate & trustworthy.

DOMS | Page No.
Date / /

* Types of Data

	Structure	Unstructure	Semi - Structured
	predefined schema (like tables)	No predefined format.	some internal structure
eg →	Databases, spreadsheets.	Text documents, images, videos.	Log files, JSON, XML.
Analysis	• Easiest (SQL, statistics).	• Requires specialized technique	• Easier than unstructured
storage	Relational databases, spreadsheets.	File systems, cloud storage.	• file system, cloud storage NO SQL databases.
Uses	Transactions, reporting ML	Text analytics, img recognition.	• configuration, log analysis data exchange.
Table?	Yes (directly)	No (needs processing)	Maybe (after processing)

Supervised ML	Unsupervised ML
• Learns from labelled data with input-output pairs.	• Learns patterns & structures from unlabelled data.
• Labelled training data -	• doesn't require labelled
• Predicts output based on input data	• discovers hidden patterns or structure in data.
• Receives feedback during training.	• Doesn't receives feedback.
• can test our model	• desired output is not given
• Desired output is given	• clustering
• Also called classification.	• K-Mean clustering, PCA.
• eg → classification, regression, object detection.	

* Regression Analysis -

- statistical technique used to estimate rel" b/w a dependent variable & one or more independent variable.
- identify general trend b/w variables
- use model to predict dependent variable based on value of its independent variables.

DOMS | Page No.
Date / /

* Process → • Data Collection • Parameter Estimation
• Model selection • Evaluation.
(Linear regression, multiple etc).

* Types of Regression.

1) Linear Regression → ML algorithm → supervised learning
→ predict dependent variable based on independent.
→ regression line is best fit line for model.
x → Independent variable Y → output.

2) Simple linear Regression → one independat (or input) variable.

3) Multiple linear Regression → more than one indepent variables.
→ multivariate regression.

4) Logistic Regression → used to model & estimate probability of an event
→ estimate probability of an event occuring
→ Is a curve
→ sigmoid curve or S-curve in short
→ 0 → totally uncertain & 1 → certain

DOMS | Page No.
Date / /

Clustering -

cluster → a no. of similar things that occur together.

→ technique in which data points are arranged in similar groups dynamically without any pre-assignment of groups.

partioning a set of data in a set of meaningful subclasses.

- properties of a cluster -
  - all data points in a cluster should be similar to each other
  - data points from diff clusters should be as different as possible.

Types → Hard clustering → each data point → only one cluster
→ soft clustering → each data point → separate cluster probability or likelihood of data.

Applⁿ →
- Customer Segmentation
- Marketing
- Seismology
- Image Processing
- Insurance
- Land Use
- Recommendation Engine
- Healthcare
- Urban Planning

K-Means

- heuristic method.
- unsupervised learning algo.
- solve clustering problems.
- groups are unlabelled dataset in diff clusters.
- k → define no. of prefined clusters.

adv →
- efficient in computation.
- easy to implement

disadv -
- applicable only when mean is defined
- need to specify K.
- Trouble with noisy data.
- not suitable to discover clusters.

Hierachial Clustering

- hierarchial cluster analysis or HCA.
- method of cluster analysis
- data points are arranged in hierarchy

Agglomerative	Divisive
• bottom up approach	• top down approach.
• each item → cluster	• all item → one cluster
• Iteratively ↳ merged together	• large clusters ↳ successively divided

DOMS | Page No.
Date / /

Data dimensionality –

no. of attributes or features used to represent each data point in a dataset.

Type → High → many features relative to no. of observations.
→ Low → Features relative to no. of observations.

Impact → • curse of dimensionality
• overfitting
• Interpretability
• Computational complexity

Dimensionality Reduction
→ Feature selection
→ Feature Extraction

Techniques → PCA (Principal Component Analysis)
→ t-distributed Stochastic Neighbor Embedding
→ Linear discriminant analysis
→ Autoencoders

Appl^n → Image & Video Processing
→ Text Mining & NLP
→ Bioinformatics
→ Financial Modeling

Spark Programming Model

→ Built for distributed computing on big data
→ Lazy evaluation of transformations & actions.
→ Fault tolerance through lineage info.
→ In memory processing for high performance.

Benefits
→ Scalability
→ High Performance
→ Fault Tolerance
→ Ease of Use
→ Unified Framework

Adv
→ Flexibility
→ Interactivity
→ Integration
→ more approachable for programers

Disadv
→ Complexity
→ Resource Management Overhead
→ Overhead in memory Processing
→ Debugging & Monitoring

DOMS | Page No.
Date / /

* MLib Library -

- algorithm for Big Data - offers various ML algo to handle large datasets
- provides easy to use tools for ml workflows high level APIs.
- ML library for Apache Spark.

Adv →
- Scalability
- Variety of algo
- ease of use
- Integration with Spark
- Active community.
- flexibility.
- simplified development

Disadv →
- Learning Curve
- complexity for Beginners
- limited out of box features
- dependency on Spark ecosystem.

* Content based recommendation System -

- suggest items based on features of items & user preferences.
- don't relay on user-item interaction but analyze item attributes.

Adv →
- Interpretability
- Cold start Mitigation.
- User Personalization.
- Transparency.
- User Independence

Disadv →
- Data Sparsity
- Echo chambers
- limited context consideration
- lack of diversity in recommendations

Appln →
- E-commerce
- Media streaming
- News Aggregation.