ASSIGNMENT - 1

M Ramani Priya
2451-18-733-001

1. What is Data Science? How is it different from Data Analysis and Buissiness Intelligence?

Data Science : Data science is a field in which information and knowledge are extracted from the data as by using various scientific methods, algorithms and processes. It can be defined as combination of various mathematical tools, algorithms, statistics and machine learning techniques which are thus used to find the hidden patterns and insights from the data which helps in decision making process.

Data Analysis	Buissiness Intelligence
1. Data analytics refers to modifying the raw data into a modified format.	1. Buissiness intelligence refers to the information required to enhance buissiness decision making activities.
2. The prime purpose of data analytics is to model, cleanse, predict and transform the data as per the buissiness needs.	2. The prime purpose of buissiness intelligence is to provide support in decision making & help the organisations to grow their buissiness.
3. Data analytics can be implemented using various storage tools available in the market. Data analytics can also be implemented using BI tools but it depends on the approach or strategy	3. Buissiness intelligence can be implemented using various BI tools available in the market. BI is implemented only on historical data stored in data warehouses or data marts.

designed by an organisation.

2451-18-733-001

4. Data analytics can be debugged via the proposed model to convert the data into meaningful format.	4. BI mechanism can be debugged only through historical data provided & the end user requirements

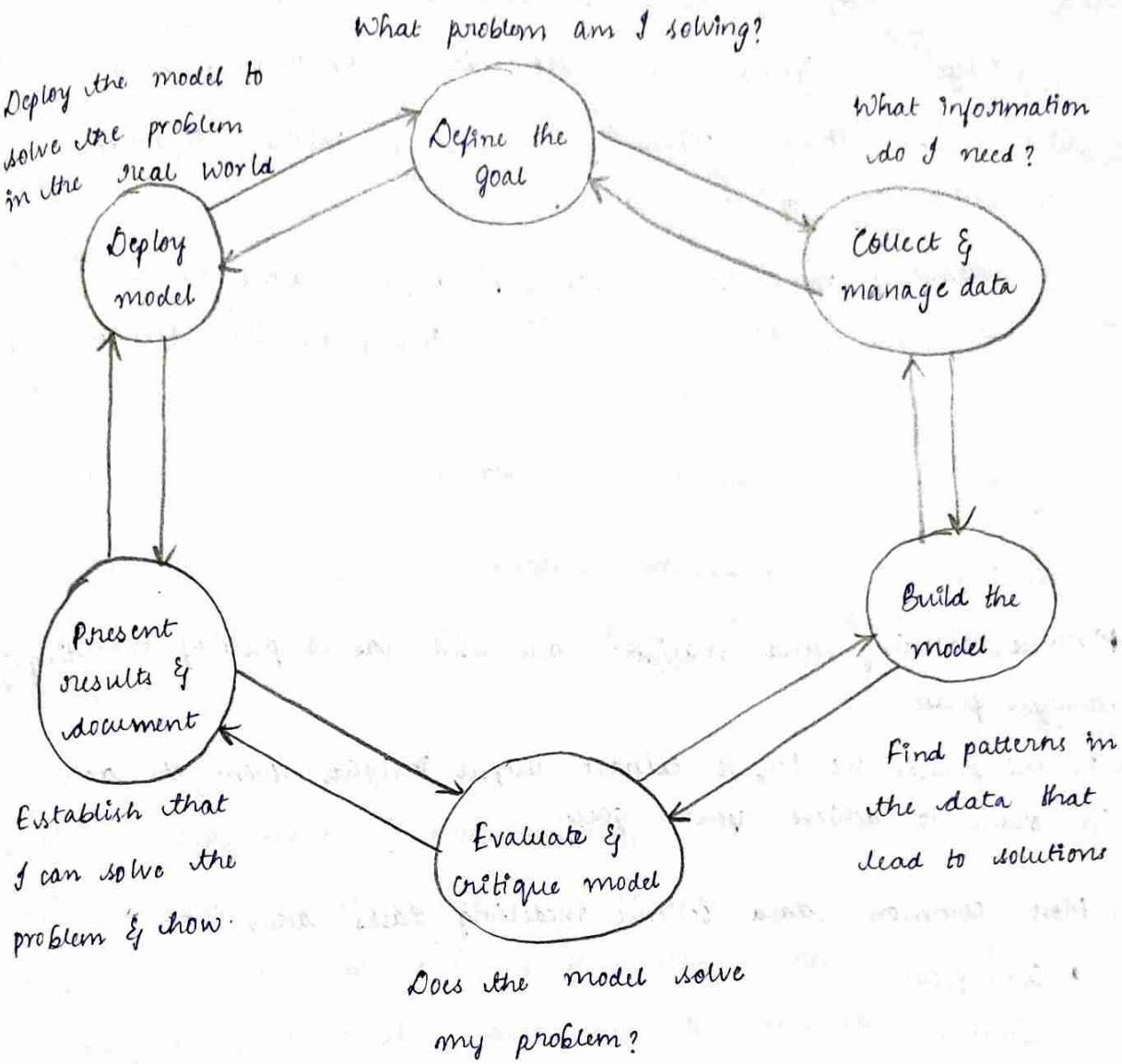
→ Data Analytics is a process that helps the enterprise users to transform the raw or unstructured data into a meaningful format.

→ Buissiness Intelligence is implemented accross many organisations to enhance their decision-making capabilities, analyze the buissiness data, perform data mining, develop reports & improve operational capabilities.

→ The major difference between buissiness intelligence & data analytics is that analytics is geared more toward future predictions & trends, while BI helps people make decisions based on past data.

6. Draw & explain the lifecycle of Data Science. 2451-18-733-001
2. What is Data Science process? Explain.

The lifecycle of data science project.



1. Define the goal.

→ The first task in data science project is to define a measurable and quantifiable goal.

→ The goal should be specific & measurable.

→ A concrete goal begets concrete stopping conditions & concrete acceptance criteria. The less specific the goal, the likelier that the project will go unbounded.

→ The goal needs to come up with a candidate hypothesis. These hypothesis can then be turned into concrete questions or goals for a full ~~time~~ scale modeling project.

2. Data collection and management.

→ This step encompasses identifying the data you need, exploring it, and conditioning it to be suitable for analysis.

→ This stage is often the most time consuming step

→ This is the stage where you conduct initial exploration & visualization of the data.

→ Data cleaning needs to be done: repair data errors & transform variables as needed

3. Modelling.

→ Machine learning and statistics are used as a part of modelling & analysis phase.

→ In this phase, we try to extract useful insights from the data in order to achieve your goals.

→ Most common data science modelling tasks are,

1. Classification
2. Scoring
3. Ranking
4. Clustering
5. Finding relations
6. Characterization

4. Model Evaluation and Critique.

→ Once a model is built, we need to determine if it meets our goals.

Mesu

→ Some of the measures used for evaluation are,

→ Recall

→ Precision.

→ false positive value

→ accuracy etc.

5. Presentation and Documentation.

→ Once the model meets the success criteria, the results of project are presented to sponsor & other stakeholders.

→ The model needs to be documented for those in organisation who are responsible for using, running & maintaining the model once it has been deployed.

→ Different audiences requires different kinds of information.

→ A presentation for the model's end users would instead emphasize how the model will help them do their job better.

6. Model deployment & maintainence.

→ Finally, the model is put into operation.

→ It should still be ensured that the model will run smoothly & won't make disastrous unsupervised decisions.

3. List out the applications of Eigen values & Eigen vectors in Data Science?

Eigen ~~value~~ vector:

In linear algebra, an eigen vector or characteristic vector of a square matrix is a vector that doesnot changes its direction under the associated linear transformation.

In other words, if V is a vector that is not zero, then it is an eigen vector of a square matrix A if AV is a scalar multiple of V.

$$AV = \lambda V$$

V - eigen vector

λ - eigen value

Eigen value:

In above equation, λ is eigen value or characteristic value associated with eigen vector V.

$$|A - \lambda I| = 0$$

Applications of Eigen values & Eigen vectors.

1. Used in dimensionality reduction technique - Principal Component Analysis (PCA), these concepts help in reducing dimensionality of data (curse of dimensionality) resulting in the simpler model which is computationally efficient & provides greater generalization accuracy.

→ The concept of Eigenvectors & Eigenvalues are used to determine a set of important variables along with scale along different dimensions (key dimensions based on variance) for analysing data in better manner.

→ Feature extraction algorithms such as Prime PCA depend on the concepts of eigen values & eigenvectors to reduce the dimensionality of data (features) or compress the data (data compression) in form of principal components while retaining the most of the original information

4 Why Linear algebra is significant in Data science?

→ Linear algebra is one of the foundational blocks of Data Science.

→ With the understanding of Linear algebra, we will be able to develop a better inhition for machine learning & deep learning algorithms. This would allow you to choose proper hyperparameters & develop a better model.

→ Linear algebra is important for machine Learning. Most machine learning models can be expressed in matrix form. A dataset is often represented as matrix.

→ Linear algebra is used in data preprocessing, data transformation & model evaluation.

→ In data preprocessing, the shape of dataset refers to the number of features & number of observations.

→ In data visualization, the features are defined as column matrices.

→ Covariance matrix - provides information about co-movement (correlation) between features.

→ Eigenvalues & Eigen vectors.

→ Cumulative variance

→ Linear Regression matrix

→ Linear Discriminant analysis matrix

→ Linear Algebra applications for data scientists

1. Machine Learning: loss functions & recommender systems.
2. Natural language processing: word embedding
3. Computer vision: Image convolution

5. How Linear Algebra is applied in Data Science?

→ Linear algebra is the heart to almost all areas of mathematics & its concepts are crucial prerequisite for understanding the theory behind Data Science.

Applications of Linear Algebra in Data Science:

1. Machine Learning
2. Dimensionality reduction
3. Natural Language Processing
4. Computer vision.

1. Linear algebra in Machine Learning.

(i) Loss functions

→ It is difficult to predict calculate how different prediction is from the expected output. This issue can be resolved using loss function.

→ A loss function can simply be a vector norm - magnitude.

→ 2 of the vector norms :

1. L1 Norm or Manhattan distance or Taxicab norm

2. L2 Norm or Euclidean distance

(ii) Regularization

It is a technique used to prevent models from overfitting.

→ A model is said to overfit when it fits the training data too well. Such model doesnot perform well with new data because it has learnt even the noise in training data.

→ Regularization penalizes overly complex methods models by adding the norm of the weight vector to the cost function. Since the cost function is needed to be minimized, the regularization norm needs to be minimized This causes unrequired components ~~to~~ of weight vector to reduce to 0 & prevents the prediction from being overly complex.

→ types of regularization - 1. Lasso regression - L1 regularization
2. Ridge regression - L2 "

(iii) Covariance Matrix.

→ Bivariate analysis is an important step in data exploration to study relationship between pairs of variables.

→ Correlation is standardized value of covariance that tells the strength & direction of linear relationship.

→ expression for covariance matrix : $Cov = X^T X$

X - standardised data matrix containing all numerical features

(iv) Support vector machine classification.

→ It is an application of concept of vector spaces in linear algebra

→ It is a supervised learning algorithm. SVM is a discriminative classifier that works by finding a decision surface

→ A hyper plane is a subspace whose dimensions are one less than its corresponding vector space, so it would be a straight line for 2D vector space, a 2D plane for 3D vector space, ...

Vector norm is used to calculate the margin.

2. Linear Algebra in Dimensionality Reduction

(i) Principle Component Analysis.

(ii) Singular value Decomposition.

(i) Principal of Component Analysis (PCA) is an unsupervised dimensionality reduction technique that finds the directions of maximum variance & projects the data along them to reduce the dimensions. These directions are eigenvectors of the covariance matrix.

(ii) Singular value Decomposition (SVD)

SVD is an amazing technique of matrix decomposition.

In Truncated SVD.,

Start with mxn data matrix,

decompose into 3 matrices - Choose k singular values based on diagonal matrix & truncate (trim) the 3 matrices.,

Finally multiply the truncated matrices to obtain the transformed matrix A_k. It has dimensions $m \times k$ - it has k features ($k < n$)

$A = U D V^T$

- U: Left singular vectors
- D: singular values
- V^T: right singular vectors.

3. Linear Algebra in Natural Language Processing.

1. Word Embeddings
2. Latent Semantic analysis.

1. Word Embeddings.

is a way of representing words as low dimensional vectors of numbers while preserving their context in the document.

→ These representations are obtained by training different neural networks on a large amount of text which is called a corpus.

They also help in analysing syntactic similarity among words.

2. Latent Semantic analysis (LSA)

is one of the techniques of topic modeling.

Latent means 'Hidden'. It is an application of singular value Decomposition.

→ LSA attempts to capture the hidden themes or topics from the documents by leveraging the context around the words.

→ First generate the Document-Term matrix of data

Use SVD to decompose matrix into 3 matrices:

→ Document Topic matrix

→ Topic Importance Diagonal matrix

→ Topic term matrix

→ Truncate the matrices based on the importance of topics

4. Linear Algebra in Computer vision.

→ Deep learning methods can achieve state of art results on challenging computer vision problems such as image classifications, object detection & face recognition.

(i) Image representation as Tensors

(ii) Convolution & Image processing.

(i) Image representation as Tensors.

A computer doesnot process images as humans do. Machine Learning (ML) algorithms need numerical features to work with.

A digital image is made up of small indivisible units called pixels.

→ grayscale image (zero) — 8×8 - 64 pixels

each pixel has value of range - 0 to 255.

↑ black pixel ↑ white pixel

→ an m×n grayscale image can be represented as a 2D matrix with cells containing respective pixel values.

→ A colored image is stored in RGB system.

each image is thought of being represented as three 2D matrices each one for R, G, B.

In every channel 0 pixel value - 0 intensity

255 - full intensity of color.

Each pixel value is a combination of values in 3 channels

→ A tensor is used instead of using 3 matrices.
A tensor is generalized n dimensional matrix.

→ RGB image uses 3rd ordered tensor.

(i) Convolution & Image Processing.

→ 2D convolution is a very important operation in image processing

Steps:

→ Start with a small matrix of weights, called a kernel or a filter

→ Slide this kernel on the 2D input data, performing element wise multiplication.

→ Add the obtained values & put the sum in a single output pixel.

→ This function is used for performing various image processing operations like sharpening & blurring the image & edge detection.

2. What is Data Science process ? Explain

Data Science process helps data scientists use the tools to find unseen patterns, extract data & convert information to actionable insights that can be meaningful to the company.

The roles in data science process.

→ Data science project is a collaborative effort that draws on a number of roles, skills & tools

Project roles:

1. Project sponsor - Represents the business interests, champions the project.
2. Client - Represents end users interests, domain expert.
3. Data scientist - Sets & executes analytic strategy, communicates with sponsor & client
4. Data architect - Manages data & data storage, sometimes manages data collection.
5. Operations - Manages infrastructure, deploys final project results.

Stages in data science project.

→ An ideal data science environment is one that encourages feedback & iteration between the data scientist & other stakeholders.

Steps in data science life cycle:

1. Define the goal: The first task in a data science project is to define a measurable & quantifiable goal.

Once you have good idea of projects goals, you can focus on collecting data.

2. Data collection & management

This step encompasses identifying the data you need, exploring it, & conditioning it to be suitable for analysis.

3. Modelling.

Here is where you try to extract useful insights from the data in order to achieve your goals using machine learning & statistics.

common modelling tasks:

classification, scoring, ranking, clustering, finding relations, characterisation.

4. Model evaluation & critique.

Once model is built we need to evaluate if it meets our goals.

We use statistical measurements like precision, recall, accuracy etc.

5. Presentation & Documentation.

The results of the project needs to be presented to the project sponsor & other stakeholders.

The model must be documented for those in organisation who are responsible for using running & maintaining model once deployed.

6. Model deployment & maintainence.

The model is put into operation & should ensure that the model will run smoothly & won't make unsupervised decisions.

→ Setting expectations.

2451-18-733-001

(i) Determining lower & upper bounds on model performance

a) The NULL Model: A lower bound on performance

b) The BAYES rate: An upper bound on model performance