Class XII

ANSWER KEY

Artificial Intelligence

Part-B: Subject Specific Skills

1. Capstone Project

Exercise

Unsolved Questions

A. 1. test 2. hyperparameters 3. MSE 4. deployed 5. Data cleansing

B. 1. False 2. False 3. True 4. True 5. False

C. 1. a. Empathize: The first stage of the design thinking process is to empathise. Design teams do research to gain a better understanding of their users' demands. They set aside assumptions in order to gain insights into the users' environment through observation and consultation with users. They can better comprehend users' experiences, motives, and challenges this way.

b. Ideate: Ideation is a creative process in which designers produce ideas in groups (e.g., brainstorming, worst possible idea). It is the third level of the Design Thinking method. Participants gather with open minds in order to generate as many ideas as possible in response to a problem statement in a supported, judgment-free setting.

c. Prototype: A prototype is a simple experimental model of a proposed solution that is used to rapidly and cheaply test or validate ideas, design assumptions, and other parts of its conceptualisation, allowing the designer/s involved to make necessary improvements or possible changes in direction.

2.

Training Set	Test Set
It is a subset of our actual dataset that is fed into the machine learning model in order for it to detect and learn patterns. It trains our model in this way.	After you've developed your machine learning model (using your training data), you'll need unseen data to test it. This data is referred to as testing data, and it may be used to assess the performance and development of your algorithms' training and to alter or optimise it for better outcomes.
Training data is typically larger than testing data	This dataset is new, "unseen" data, typically 20% of the total dataset.

IMPRINT 1

3. A capstone project is comprehensive, independent, and final project undertaken as part of the curriculum designed to assess the skills, knowledge, and expertise a student has acquired. Such a project often involves researching a topic, evaluating a new technique or method, developing a health plan, researching a character or event in history, or even the composition of a sketch or a play.

   Examples:

   - Studying images to diagnose disease
   - Forecasting student results
   - Creating a chatbot for the school admin department or counsellor to handle parents' students' queries using IBM Watson, Google Dialogue Flow
   - Image Classifier

4. Every project starts with a business understanding. Business sponsors, who need analytical solutions, play the most important role at this time in defining the problem, project objectives, and solution requirements from a business perspective. This first step lays the foundation for successful business problem solving and is perhaps the hardest. To help ensure project success, sponsors should be involved throughout the project to provide expert knowledge, review interim conclusions and ensure that work remains on track to produce the intended solution.

5. There are mainly two types of validation methods which are Train Test Split Evaluation and Cross Validation.

   The train test split technique can be used to test machine learning algorithms for classification and regression applications. The technique divides the given dataset into two subsets: Training dataset - it is used to train the algorithm and to fit the machine learning model; testing dataset - the algorithms create predictions using the input element from the training data in the test dataset.

   Cross-validation is a strategy for verifying model efficiency that involves training the model on a subset of input data and testing it on a previously unknown subset of input data.

D. There are numerous loss functions available for evaluating regression models. Choosing the right loss function is critical, because what makes one appealing depends on the data. Each function has its own set of attributes. Many factors influence the optimal choice of a loss function, including the algorithm employed, outliers in the data, whether the function should be differentiable, and so on.

MSE: One of the most used regression loss functions is MSE. We determine the error in Mean Squared Error, also known as L2 loss, by squaring the difference between the predicted and actual values and average it throughout the dataset. Squaring the error gives outliers more weight, resulting in a smooth gradient for minor errors. This penalization for significant errors benefits optimization algorithms by assisting in the determination of optimal parameter values. Because the errors are squared, MSE can never be negative. The error value varies

from 0 to infinity. The MSE grows exponentially as the error grows. An MSE value close to zero indicates a good model. It is especially useful in removing outliers with substantial errors from the model by giving them additional weight.

RMSE: The square root of MSE is used to calculate RMSE. The Root Mean Square Deviation (RMSE) is another name for the Root Mean Square Error. It is concerned with the deviations from the real value and measures the average magnitude of the errors. A RMSE value of 0 implies that the model is perfectly fitted. The model and its predictions perform better when the RMSE is low. A greater RMSE indicates a substantial discrepancy between the residual and the ground truth. RMSE can be used with a variety of features to determine whether or not the feature improves the model's prediction. Because of the square root, RMSE penalises errors less than MSE. The RMSE increases as the size of test sample increases.

2. Problem Decomposition entails breaking down a complex problem or system into smaller, more manageable, and easier-to-understand bits. Because smaller portions are easier to work with, they can be inspected and solved individually.

   Problem decomposition steps are as follows:

   a. Understand the problem and express the problem in your own words:
      - Understand the required inputs and outputs
      - Ask questions for clarity (in class, these questions may be directed to the teacher, however, you can also ask yourself or your colleagues)

   b. Break down the problem into several big parts. Write them down on paper.

   c. Divide any larger complicated part into smaller pieces. Continue this until all parts are small.

   d. Code the smaller parts one by one. Use the following methodology:
      - Analyse how to implement the code.
      - Write the code/query.
      - Test each code individually.
      - Fix the problem(s), if any.

3. Business Understanding, Analytic Approach, Data Requirements, Data Collection, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment, Feedback.

4. A loss function, at its most basic, is a measure of how well your prediction model predicts the expected outcome (or value). The learning problem is transformed into an optimization problem, a loss function is defined, and the method is optimised to minimise the loss function.

   Importantly, the loss function you choose is intimately tied to the activation function you choose in your neural network's output layer. These two design aspects are linked together. In machine learning, there is no such thing as a one-size-fits-all loss function. The type of machine learning method used, the ease of calculating derivatives, and, to some extent, the number of outliers in the dataset, all play a role in selecting a loss function for a certain task.

Once the business problem is clearly stated, the data scientist can define an analytical approach to solving the problem. This step is to represent a problem in the context of statistical techniques and machine learning so that the organization can determine the most appropriate for the desired outcome.

For example,

- If the goal is to predict an answer such as "yes" or "no", then the analytical method can be defined as the building, testing, and execution of a classification model.
- If the goal is to determine the probability of action, then predictive modelling can be used.
- If the goal is to show relationships, a descriptive approach may be necessary.

Δi In Life

Can have multiple solutions

Δi Deep Thinking

Can have multiple solutions

Δi Ready

1. Overfitting occurs when the model we trained has trained "too well" and is now, well, too tightly fitted to the training dataset. This frequently occurs when the model is very complex (i.e., there are too many features/variables in comparison to the amount of data). This model will be extremely accurate on training data but will most likely be extremely inaccurate on untrained or new data.

   In contrast to overfitting, underfitting occurs when a model does not fit the training data and hence misses the data's trends. It also indicates that the model cannot be applied to new data. This is typically the result of a very simplistic model (insufficient predictors/independent variables). It might also happen if we fit a linear model (such as linear regression) to nonlinear data. It practically goes without saying that this model will have low predictive power (it will only work on training data and will not generalise to other data).
2. We divide our data into k separate subgroups in K-Folds Cross Validation (or folds). We train our data using k-1 subsets and leave the final subset (or fold) as test data. The model is then averaged against each of the folds before being finalised. The more folds we have, the less error we have owing to bias but the more error we have due to variance; clearly, the computational price goes up as well — the more folds you have, the longer it takes to compute and the more memory you need. We reduce the error owing to variance by reducing the number of folds, while the error due to bias increases. It would also be less expensive in terms of computing. Therefore, in big datasets, k=3 is usually advised.

3. In K-fold cross validation, the data set is partitioned into k subsets. Each time, one of the k subsets is utilised as the test set, while the remaining k-1 subsets are combined to form the training set. The average error for all k trials is then computed. The advantage of this strategy is that it is less important how the data is separated. Every data point appears exactly once in a test set and k-1 times in a training set.

   Leave-one-out cross validation is a logical extension of K-fold cross validation, with K equal to N, the number of data points in the set. That is, the function approximator is trained on all of the data except for one point, and a forecast is made for that point N times. As previously stated, the average error is calculated and used to evaluate the model.

4. A hidden layer is positioned between the algorithm's input and output in neural networks, where the function assigns weights to the inputs and guides them through an activation function as the output. Hidden layers allow a neural network's function to be broken down into specific data modifications. Each hidden layer function is tailored to generate a certain result. For example, a hidden layer function that identifies human eyes and ears may be used in concert with succeeding layers to detect faces in photos. While the functions to detect eyes alone are insufficient to distinguish things independently, they can interact together within a neural network.

2. Model Lifecycle

Exercise

Unsolved Questions

A. 1. Modelling 2. the 4Ws Problem Canvas
3. government websites, cameras 4. test data 5. Who

B. 1. True 2. False 3. False 4. True 5. True

C. 1. a. Data Exploration: After gathering data, processes such as:

- data cleaning to locate missing values,
- eliminating worthless data (erroneous samples and outliers), and
- performing basic statistical analysis such as drawing graphs (or any other visual representation) and comparing different properties of the data set are carried out.

The initial insights gained help to get an understanding of the data and later on, help in algorithm selection, metrics choice, etc. This complete procedure is called "Exploratory Data Analysis". It is useful to see which elements are more essential and what the overall trend of the data is.

b. Modelling: Modelling is the process through which several models based on graphical data can be constructed and even tested for advantages and disadvantages. ML engineers go through multiple models to determine the best model configuration. Hence, the design phase is an iterative process. Hyperparameter fine-tuning provided by most ML frameworks helps to narrow down the number of feasible solutions. These approaches assess performance for many configurations, compare them, and inform of the best ones.

2. Deployment is the process of integrating a machine learning model into an existing production environment in order to make real data-driven business choices. It is one of the final steps of the machine learning life cycle and might be one of the most time-consuming. The deployed model's performance is monitored to ensure that it continues to function at the level required by the business.
3. Model evaluation is an essential step in the model development process. It aids in determining the optimal model to represent our data and how well the chosen model will perform in the future.

   Once a model has been created and trained, it must be properly tested to calculate the model's efficiency and performance. As a result, the model is evaluated using Testing Data (which was extracted from the acquired dataset during the Data Acquisition stage) and the model's efficiency is assessed.

   The set of measurements will differ depending on the problem you're working on. For regression problems, for example, MSE or MAE are commonly used. On the other hand, for a balanced dataset, accuracy may be a useful choice for evaluating a classification model. Imbalanced sets necessitate the use of more advanced metrics. In such instances, the F1 score is useful.

   There are a few other things considered during this stage too:

   - The volume of test data can be huge, which provides data complexities.
   - Human biases in picking test data might have a negative impact on the testing phase; thus, data validation is critical.
4. Problem Scoping is the process of comprehending an issue and determining the different aspects that affect it, as well as defining the project's goal. Scoping a project allows you to understand the users, the product, and the problem you're solving so that if something goes wrong, you'll know exactly what to modify. Example: an AI system that predicts the presence of a specific tumour based on CT images. So, how do your current business constraints look?

   - Your model must detect patterns in images.
   - False forecasts can have disastrous effects. Errors can be fatal in such a case.
   - It may take up to an hour to reveal the outcome so the motto "Take your time, but be precise." must be followed.

5.

OUR	[stakeholders] General public	WHO
HAS/HAVE PROBLEM THAT	[**issue, problem, need**] To detect COVID virus in a person.	WHAT
WHEN/WHILE	[**context, situation**] Problem is that a quick, dependable, widely available, and economical diagnostic method/system is required	WHERE
AN IDEAL SOLUTION	[benefit of solution to them] An effective AI solution should be accurate as early detection will save lives and help in restraining the spread of COVID 19.	WHY

Ai In Life

Can have multiple solutions

Ai Deep Thinking

Can have multiple solutions

Ai Ready

I. 1. Problem identification entails:

- a clear identification of the problem; and
- the discovery of the root cause of the problem.
- Determine the underlying, 'true' problem.
- Correctly framing the problem. (Framing is a Structural Representation of a Problem or Issue; it includes identifying and clarifying the context of the problem to aid understanding.)

2. Primary Data: Primary data is information that is gathered directly from a data source. It is typically collected specifically for a study project and may be shared openly for use in future research. Primary data is frequently trustworthy and authentic. Primary data is also real-time because it is collected and stored as needed.

   Secondary Data: Secondary data is data that has previously been acquired by someone else and made available for use by others. Because secondary data is released openly, it is usually widely available to academics and individuals. This, however, implies that the

data is typically general and not specific, and that this form of data may be out of date and unsuitable for the individual's usage.

3. No, the model is not of any use. We need to go back to Problem Scoping and check if we have defined/identified the problem correctly. We also need to go back to Data exploration to identify areas or patterns to dive into and dig more. This allows for a deeper, more comprehensive, and better understanding of the data.

II. Can have multiple solutions

3. Storytelling Through Data

Exercise

Unsolved Questions

A. 1. Data storytelling 2. entertained
3. Data Visualizations 4. pinpoint 5. Narrative

B. 1. True 2. False 3. False 4. True 5. False

C. 1. The primary advantage of data storytelling is that it generates intelligent, actionable insight from data. A compelling narrative will bring the data to life. It can result in a "Aha" moment or profound insight. In a corporate environment, the data's story can be adapted to the intended audience, making it more meaningful and relevant.

The advantages of data storytelling include:

- Give vital insights
- Showcase new viewpoints
- Complicated information should be interpreted
- Motivate people to act

2. Data storytelling uses a structured approach to delivering data insights that always includes a combination of three main elements: data, graphics, and narrative. When a narrative is backed by data, it helps to explain to the audience what is happening in the data and why a specific insight was developed. When visuals are applied to data, they can enlighten the audience to insights that they would not have noticed otherwise, such as charts or graphs. Finally, when narrative and images are combined, an audience can be engaged or even entertained. When the proper graphics and narrative are combined with the correct data, you have a data story that has the potential to impact and drive change.

3. Three aspects are used in persuasive data visualisation storytelling:

a. Data b. Narratives c. Visuals

Data and narratives enlighten. Providing context for your data, such as where it came from, why it's significant, and what you did with it, helps your audience understand what it is and why it's essential.

Data and graphics enlighten. Data by itself, just statistics, frequently leaves people perplexed. However, when you design a strong visualisation that clearly depicts the data and what it implies, your audience will have a "Aha!" moment. They've been awakened!

Narratives and visuals captivate. Clear and succinct communication, along with effective visualisation, engages your audience, draws them in, and captures their attention.

Change is brought about by the confluence of all of these factors.

4. The following are the steps involved in telling an effective data story:
   - Recognizing the audience
   - Choosing the appropriate data and visualisations
   - Highlighting important information
   - Creating a narrative
   - Keeping your audience interested
5. The structure consists of the following elements: characters, setting, storyline, conflict, and resolution. These crucial parts keep the story moving forward and allow the action to unfold in a logical manner that the reader can follow. The characters are the people that are the focus of the story.

D. 1. Narrative employs language in a format that is tailored to our specific needs, enhancing our full comprehension of new information. Visualisations and data are vital proof points in a story, which is a key vehicle for conveying ideas. The significance of a narrative stems from the fact that it explains what is happening in the data set. It provides context and meaning, as well as relevance and clarity. A story directs the audience's attention and informs them of what not to overlook. It also holds the audience's attention.

2. Can have multiple solutions

3. Can have multiple solutions

4. • *Context is what adds value to the insights created by data. Data* scientists can assist provide important context and successfully explain their ideas to everyone in the room by using their storytelling talents. A solid storytelling narrative may make data and analytics much more approachable.
   - *Data storytelling is effective*. Data stories typically outperform the metrics that content marketing teams care about, such as time on page, bounce rates, click-through rates, and conversions. They are also more likely to receive backlinks from other websites, which is crucial for your content's performance in search engine rankings.
   - *Data storytelling has scalability*. The development of no-code platforms means that you can generate stunning data stories without hiring engineers.

5. Best Practices for data storytelling:
    - Always label your axes and give your plot a title.
    - When legends are required, use them.
    - Colors that are lighter on the eye and in proportion should be used.
    - Avoid adding superfluous detail to your visualisation, such as backdrops or themes that make it difficult to read.
    - Only a point can encode two quantitative values based on a horizontal and vertical location at the same time.

Ai In Life

Can have multiple solutions

Ai Deep Thinking

Can have multiple solutions

Ai Ready

I. Can have multiple solutions

II. 1. order 2. contrast, patterns 3. focal-point 4. line