DESCRIPTIVE QUESTION-

Q1. What is meant by Model Evaluation?  
Ans: Model evaluation is the process that uses some metrics which help us to analyze the performance of the model. As we all know that model development is a multi-step process and a check should be kept on how well the model generalizes future predictions. Therefore, evaluating a model plays a vital role so that we can judge the performance of our model. The evaluation also helps to analyze a model’s key weaknesses. There are many metrics like Accuracy, Precision, Recall, F1 score, Area under Curve, Confusion Matrix, and Mean Square Error. Cross Validation is one technique that is followed during the training phase and it is a model evaluation technique as well.  
  
Q2. What is meant by ‘Accuracy’ in Performance Metric?  
Ans: Accuracy: Accuracy measures the proportion of correct predictions made by the model out of the total predictions made. It's calculated as the number of correct predictions divided by the total number of predictions. While accuracy is a commonly used metric, it might not be suitable for imbalanced datasets where one class dominates the others.  
  
Q3. What is meant by ‘Precision’ in Performance Metric?  
Ans: Precision: Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It measures the accuracy of positive predictions. Precision is sensitive to false positives; a higher precision indicates fewer false positives.  
  
Q4. What is meant by ‘Recall (sensitivity)’ in Performance Metric?  
Ans: Recall (Sensitivity): Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive instances in the dataset. It's calculated as the ratio of true positives to the sum of true positives and false negatives. Recall is sensitive to false negatives; a higher recall indicates fewer false negatives.  
  
Q5. What is Confusion Matrix?  
Ans: A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model’s predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.  
The matrix displays the number of instances produced by the model on the test data.  
  
True positives (TP): occur when the model accurately predicts a positive data point.

True negatives (TN): occur when the model accurately predicts a negative data point.

False positives (FP): occur when the model predicts a positive data point incorrectly.

False negatives (FN): occur when the model mispredicts a negative data point.

Q6. Why do we need Confusion Matrix?  
Ans: When assessing a classification model’s performance, a confusion matrix is essential. It offers a thorough analysis of true positive, true negative, false positive, and false negative predictions, facilitating a more profound comprehension of a model’s recall, accuracy, precision, and overall effectiveness in class distinction. When there is an uneven class distribution in a dataset, this matrix is especially helpful in evaluating a model’s performance beyond basic accuracy metrics.  
  
Q7. What is Cross-Validation?  
Ans: Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds.  
  
The main purpose of cross validation is to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. By evaluating the model on multiple validation sets, cross validation provides a more realistic estimate of the model’s generalization performance, i.e., its ability to perform well on new, unseen data.  
  
Q8. What is underfitting?

Ans: A statistical model or a machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data. In simple terms, an underfit model is inaccurate, especially when applied to new, unseen examples. It mainly happens when we use very simple model with overly simplified assumptions. To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.  
  
Q9. What is overfitting?  
Ans: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore, they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.  
  
Q10. What is PEFT or Parameter Efficient Fine-Tuning?  
Ans: Parameter Efficient Fine-Tuning (PEFT) is a technique used in machine learning, particularly in the context of natural language processing (NLP), to fine-tune pre-trained language models with a limited number of parameters, while still achieving competitive performance.  
Fine-tuning refers to the process of taking a pre-trained model (often trained on a large dataset for a general task) and further training it on a smaller dataset specific to a particular task or domain. This process allows the model to adapt its parameters to better suit the target task.

Parameter efficiency is crucial, especially in resource-constrained environments or when deploying models to edge devices with limited computational resources. PEFT aims to achieve good performance while minimizing the number of parameters that need to be fine-tuned, thus reducing computational cost and memory requirements.

Q11. What is meant by ‘adapters’ in PEFT?  
Ans: Adapters were one of the first parameter-efficient fine-tuning techniques released. In the paper, they showed that you can add more layers to the pre-existing transformer architecture and only finetune them instead of the whole model. They showed that this technique resulted in similar performance when compared to complete fine-tuning.

Q12. What is LoRA- Low Rank Adaptation?  
Ans: LoRA, or Low Rank Adaptation, is a technique used in the context of fine-tuning pre-trained language models, particularly in natural language processing (NLP). It is aimed at efficiently adapting pre-trained models to specific downstream tasks while minimizing the computational cost and memory requirements.  
  
A much simpler explanation is that during finetuning only a very few weights are updated a lot as most of the learning is done during the pretraining phase of the neural network. LoRA uses this information to reduce the number of trainable parameters.  
  
Q13. What is Prompt Tuning?  
Ans: Prompt tuning was one of the first papers to build upon the idea of finetuning only with soft prompts. The ideas of P-Tuning and Prefix Tuning come from this paper. Prompt tuning is a very simple and easy-to-implement idea. It involves prepending a specific prompt to the input and using virtual tokens or new trainable tokens for that specific prompt. These new virtual tokens can be finetuned during the process to learn a better representation of the prompt. This means that the model is tuned to understand the prompt better.

Q14. What are Hard and Soft prompts?  
Ans: Hard Prompts can be seen as the idea of a defined prompt which is static, or at best a template. A generative AI application can also have multiple prompt templates at its disposal to make use of.  
“Hard prompts are manually handcrafted text prompts with discrete input tokens. ~ HuggingFace”  
Prompt templating allows for prompts to be stored, re-used, shared, and programmed. And generative prompts can be incorporated in programs for programming, storage and re-use.

Soft prompts are created during the process of prompt tuning. Unlike hard prompts, soft prompts cannot be viewed and edited in text. Prompts consist of an embedding, a string of numbers, that derives knowledge from the larger model.

So for sure, a disadvantage is the lack of interpretability of soft prompts. The AI discovers prompts relevant for a specific task but can’t explain why it chose those embeddings. Like deep learning models themselves, soft prompts are opaque.

Soft prompts act as a substitute for additional training data. Researchers recently estimated that a good language classifier prompt is worth hundreds to thousands of extra data points.

Q15. What is meant by Entropy?  
Ans: Entropy is a measure that quantifies uncertainty and is obtained as the inverse of probability of an event occurring. Higher the probability, lesser is the uncertainty. Hence, the goal of the language model is to minimize the entropy of generating a sequence of words that are similar to the training sequences.

MULTIPLE CHOICE QUESTIONS-

1. What is a common performance metric used to evaluate the effectiveness of machine learning models?

- A) Error rate

- B) Precision

- C) F1 score

- D) All of the above

- Correct Answer: D) All of the above

2. Which metric measures the proportion of true positive predictions among all actual positives in the dataset?

- A) Accuracy

- B) Precision

- C) Recall

- D) F1 score

- Correct Answer: C) Recall

3. Which technique involves dividing a dataset into multiple subsets, training the model on some subsets, and evaluating it on others?

- A) K-means clustering

- B) Decision tree

- C) Cross-validation

- D) Gradient boosting

- Correct Answer: C) Cross-validation

4. What does underfitting in machine learning refer to?

- A) Model performs well on training data but poorly on unseen data

- B) Model captures noise in the training data

- C) Model is too simple to capture the underlying structure of the data

- D) Model has too many parameters

- Correct Answer: C) Model is too simple to capture the underlying structure of the data

5. Which technique aims to adapt pre-trained models to new tasks while minimizing computational cost and memory requirements?

- A) Feature engineering

- B) Overfitting

- C) PEFT (Parameter Efficient Fine-Tuning)

- D) Gradient descent

- Correct Answer: C) PEFT (Parameter Efficient Fine-Tuning)

6. What is the term used to describe a situation where a model learns the training data too well but fails to generalize to unseen data?

- A) Overfitting

- B) Underfitting

- C) Bias

- D) Variance

- Correct Answer: A) Overfitting

7. Which metric is used to measure the uncertainty or randomness in a probability distribution?

- A) Precision

- B) Entropy

- C) Cross entropy

- D) Perplexity

- Correct Answer: B) Entropy

8. Which of the following is NOT a component of a confusion matrix?

- A) True positives

- B) False positives

- C) True negatives

- D) True neutrals

- Correct Answer: D) True neutrals

9. What is the metric used to evaluate the performance of language models by measuring the average number of choices they have for the next word?

- A) Accuracy

- B) Precision

- C) Cross entropy

- D) Perplexity

- Correct Answer: D) Perplexity

10. Which technique involves providing task-specific instructions to pre-trained language models to guide their responses?

- A) Feature extraction

- B) Prompt tuning

- C) Dropout

- D) Batch normalization

- Correct Answer: B) Prompt tuning