8/10/24

ASSIGNMENT 1

Unit-1, 2 & 3(1) Introduction to Machine Learning, Modelling and evaluation & Basics of Feature engineering

Q.1 What is Machine learning? Describe several types of learning each with a necessary example.

→ Machine Learning (ML) is a branch of AI that enables systems to learn from data and improve performance without being explicitly programmed. The core idea is to build algorithms that can recognize patterns in data, make decisions and predict outcomes by training on large datasets.

→ Types of Machine Learning are :-

i) Supervised Learning:
The Algorithm is trained on a labeled dataset, meaning that each input is paired with a corresponding correct output. The goal is ~~to~~ for the model to learn the mapping between input and output and make accurate ~~problem~~ predictions for new, unseen data.

eg: Predicting house prices based on features such as size, number of rooms and location.

- Algorithm : linear Regression, Decision Trees.
- Data : Labeled (input features and corresponding target prices).

Also Spam Email Detection, image

classification, sentiment analysis & more.

ii) Unsupervised Learning.
The algorithm is given an unlabeled dataset and must find hidden patterns or structures in the data. The goal is to discover interesting patterns without having predefined outputs.

eg: Customer segmentation for a retail business, where customer are grouped based on buying behaviour like frequent vs. one-time buyers.

- Algorithm : K-means Clustering, Principal Component Analysis (PCA)
- Data : Unlabeled (only input features, no target labels).

Also Market basket analysis, anomaly decision, recommendation systems & more.

iii) Reinforcement Learning
It involves an agent that interacts with an environment to maximize some notion of cumulative reward. The agent learns by receiving feedback from ~~the~~ its actions in the form of rewards or punishments, making decisions to optimize long-term performance.

eg: A robot learning to navigate a maze by receiving positive rewards for reaching the goal and negative rewards for hitting obstacles.

- Algorithm : Q-learning, Deep Q Networks (DQN)
- Feedback : Rewards (positive or negative based on actions)

Also Robotics, autonomous vehicles, game-playing tools. & more.

Q.2 Write a short note on each of the following topics :

a. Natural Language Processing.

→ NLP is a field of AI that enables machines to understand and generate human language. It combines computational linguistics, ML and deep learning to process and analyze large amount of natural language data. Key NLP tasks include sentiment analysis, language translation, text summarization and named entity recognition.

Applications are Chatbots, Google Translate, speech recognition, automated customer support.

b. Computer Vision

→ Computer Vision is a subfield of AI that enables computers to interpret and understand visual information from the

world, such as images and videos. It uses ML, DL and pattern recognition to analyze and extract meaningful insights from visual data. For example, identifying whether an image contains a cat or a dog.
Applications are facial recognition, autonomous driving, medical image analysis, surveillance.

c. Conversational AI
—4
Conversational AI refers to technologies like chatbots and virtual assistants, that enable human-like interactions between computers and humans. It leverages NLP, speech recognition and ML to process input, understand ~~ing~~ context and generate human-like responses. For example, virtual assistants like Amazon Alexa, Apple Siri or customer service chatbots.
Applications are Customer Support, virtual personal assistants, interactive eductional tools.

d. Generative AI.
—4
Generative AI involves algorithms that can create new content such as text, images, music or even software code. It leverages generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to generate

realistic outputs by learning from a given dataset. For example, ChatGPT for text generation or DALL·E for image creation.
Applications are art generation, synthetic data creation, content creation.

e. Explainable AI (XAI)

→ XAI is a set of methods and frameworks designed to make the decision-making processes of ML models understanable and transparent to humans. It aims to address the "black-box" nature of many AI models, especially DL models, by providing explanation for how models reach their conclusions, which ensures trust and fairness in AI systems. For example, in healthcare, XAI can explain why AI model recommends a certain treatment, helping doctors validate the decision.
Applications are regulatory compliance in finance, healthcare, model debugging.

Q.3 Discuss Model Selection procedure and Model training.

→ Model selection refers to the process of choosing the best ML model from a set of options for a particular problem. The goal is to find a model that performs well on unseen data, which involves balancing model complexity and generalization ability.

(i) Cross-Validation : Divides the dataset into multiple folds, training the model on one fold and validating it on others, providing a robust performance estimate and reducing overfitting.

(ii) Holdout Method : The dataset is split ~~f~~ into training and test sets. While quick and simple, this method may yield unreliable performance estimates if the split ~~if~~ is not representative.

(iii) Bootstrap Sampling : This method involves repeated sampling from the data with replacement ~~it~~ to create multiple training sets.

-4 Model training refers to the process of teaching a ML model to recognize patterns in data. The model is fed training data and it adjusts its internal parameters to minimize the error or maximize the accuracy of predictions.

(i) Training Data : Quality and representativeness of the training data are crucial for effective model performance.

(ii) Optimization : During training, optimization techniques (like gradient descent) are used to adjust the model parameters to minimize a loss function. Effective optimization helps achieve better predictive performance

(iii) Underfitting and Overfitting.

→ Underfitting, occurs when a model is too simple to capture patterns, resulting in poor performance on both training and validation sets.

→ Overfitting, on the other hand, happens when the model learns noise and details from the training data too well, leading to excellent performance on the training set but poor generalization to unseen data. Regularization techniques and careful model selection can help mitigate these issues

★ Also others are hyperparameter tuning, monitor training

Q.4 Explain Validation and Cross-Validation.
→ Validation refers to the process of evaluating a ML model's performance on a validation dataset during training. The purpose is to assess how well the model generalizes to unseen data before testing it on the final test set. Also, ~~valdi~~ validation helps in :-

i) Tuning Hyperparameters : By monitoring model performance on the validation set, we can adjust hyperparameters (learning rate, number of layers or depth of decision trees) to improve the model.

ii) Preventing Overfitting : Provides an estimate of how the model will perform on new, unseen data. If the model performs well on training data but poorly on validation set, it indicates overfitting.

iii) Early Stopping : During training, if model's performance on the validation set starts to decline, training can be stopped to avoid overfitting.

eg: Suppose, you a train a model to predict house prices. You split the data:
> 80% - Training data (used to fit the model)
> 10% - Validation data (used to fine-tune model)
> 10% - Test data (used to check final performance).

→ Cross-Validation is a more robust technique for model evaluation, especially with limited data. Instead of using one fixed validation set, the data is split into multiple folds to ensure every data point gets a chance to be in both training and validation sets.

Most common form is K-Fold Cross Validation, here the dataset is randomly divided into K equal-sized folds. The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times. The performance metrics are averaged across all K iterations to get an overall estimate of model's performance.

eg: 5-Fold Cross-Validation
- Divide the dataset into 5 folds.
- Train the model on 4 folds and test it on the remaining fold. Repeat this 5 times, using a different fold as the test set each time.
- Compute the average of the evaluation metrices across all 5 runs.

Q.5 Describe Lazy Learner and Eager Learner.
—4
These are two types of ML algorithms based on when and how they generalize from the training data.

1. Lazy Learner.
Lazy learners delay learning from the data until a query or prediction is required. These algorithms store the training data and perform minimal computation upfront. When asked to make predictions, they ~~are~~ use the stored data to compute results on the fly.

- No immediate model is built during training
- Higher Prediction Time, as most computation happens at query time.
- Works well with small datasets.
- Sensitive to irrelevant features

- Use case: Predicting a patient's disease by comparing their symptoms with stored records of previous patients.

eg: K-Nearest Neighbors (K-NN), Case-Based Reasoning.

2. Eager Learner.

Eager learners build a model from the training data immediately and store the learned patterns. After training, they use this pre-built model to make predictions quickly, without consulting the full dataset.

- Model is trained upfront before predictions are made.
- Faster Prediction Time, as the model is ready beforehand.
- More suitable for large datasets.
- Risk of overfitting / underfitting if not trained properly.

- Use case: Predicting house prices based on features like location, size and age of the property, using a pre-trained linear regression model.

eg: Decision Trees, linear Regression, Neural Networks.

Q.6 Enlist and write a note on various parameters used for evaluating performance the model.

→ Evaluating the performance of a ML model is crucial to ensure that it can make accurate predictions on unseen data.

1. Accuracy

The ratio of correctly predicted instances to the total instances in the dataset. Suitable for balanced datasets ~~are~~ where classes are equally represented.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

2. Precision.

It measures the accuracy of positive predictions. It is the ratio of true positive predictions to the total predicted positives. Important when the cost of false positives is high (eg, Spam detection)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. Recall

It measures the ability of a model to identify all relevant instances (True positives). It is the ratio of true positives to the actual classmate

Date

Page 14

positives. Crucial in scenarios where it's important to identify as many positive instances as possible (eg, disease detection).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4. Mean ~~Absolute~~ squared Error (MSE)

MSE measures the average of the squares of the errors i.e the average squared difference between estimated values and actual ~~the~~ value. MSE is sensitive to outliers and is used in regression tasks to evaluate model performance.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

5. Sum of Squared Errors (SSE).

SSE measures the total squared difference between actual values and predicted values. A lower SSE indicates a better fit. It's sensitive to large errors since it squares the difference.

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Q.7 Discuss feature engineering and describe feature transformation.

→ Feature Engineering is the process of

selecting, modifying or creating new features from raw data to improve the performance of a ML model. Well-engineered features help algorithms make more accurate predictions and generalize better to unseen data. It involves technique like encoding categorical ~~values~~, variables, handling missing values, scaling and feature extraction.

→ Feature Transformation is a specific subset of feature engineering that involves modifying features to make them more suitable for modeling. The goal is to enhance the predictive power of the features by altering their scale, distribution or format. Two approaches:-

(i) Feature Construction: Combines or derives new features from existing ones to better capture the underlying pattern in the data.

eg: Creating an interaction between two variable (multiplying features), calculating ratios or differences, generating polynomial features.

(ii) Feature Extraction: Reduces the dimensionality of the feature space by extracting the most important information from the original features.

eg: PCA, image feature extraction, word embeddings, t-SNE.

Q.8 Discuss the issues associated with higher dimensional data and explain dimensionality reduction along with its any two algorithms.

→ Higher-dimensional data, often referred to as the curse of dimensionality, introduces several challenges in ML. As the number of features increases, it becomes harder for models to learn efficiently, leading to several issues:

i) Overfitting: with too many features, the model may fit noise in the data, leading to poor generalization on unseen data.

ii) Increased computational cost: More features require more computation and memory, slowing down the model's training and prediction.

iii) Sparsity of Data: In high-dimensional spaces, data points become sparse, making it difficult to identify patterns or similarities.

iv) Interpretability Issues: Models with many features are harder to interpret, making it challenging to understand the relationships between variables.

→ Dimensionality Reduction involves transforming high-dimensional data into a lower dimensional ~~data~~ space while retaining as much information as possible which improves model performance and interpretability

i) Principal Component Analysis (PCA).
PCA is a linear technique that transforms the data by projecting it onto a new coordinate system, where each axis (principal component) captures the maximum variance. (eg; image data)

- Center the data by subtracting the mean.
- Compute the covariance matrix.
- Find the eigenvalues and eigenvectors of the covariance matrix.
- Select the top 'K' eigenvectors to represent the data.

ii) linear Discriminant Analysis (LDA).
LDA is a dimensionality reduction and classification technique that maximizes class separation by finding linear combinations of features that distinguish between labeled classes. LDA improves classification performance by enhancing between class differences.

- Measures variation within the same class.
- Measures how class mean differ from overall mean.
- Finds discrimants that maximize class separability.
- Selects top discriminants to reduce dimensionality.

eg: Face recognition; bioinformatics.

~~~ X ~~~ X ~~~
~~~