8/10/24

ASSIGNMENT 2

Unit-3(2,3) - Probability in Machine Learning & Supervised and Unsupervised Learning

Q.1 Discuss the need of probability in machine learning.

-> Probability plays a fundamental role in ML by enabling models to handle uncertainity and make predictions based on incomplete or noisy data.

i) Uncertainity Quantification:

Probability provides a framework to model this uncertainity (real world data), allowing algorithms to make informed decisions even when data is incomplete or noisy.

ii) Interference and Prediction

Probability is the foundation of statistical interference, which is essential for making predictions from data. For example, Bayesian Interference.

iii) Decision Making

Probability aids in evaluating risk and reliability of decisions made by ML models. For example, in classification tasks, regression and clustering.

iv) Evaluation Metrics

Many performance metrics, such as precision, recall and F1-score are derived from the probabilities associated with true positive, false positive, etc.

Q.2 Differentiate between discrete and continuous random variable with a suitable example.

—y

Discrete Random Variable	Continuous Random Variable
Takes on a countable number of distinct values	Takes on an infinite no. of values within a range
Described by a probability mass function (PMF)	Described by a probability density function (PDF)
Often represented with a bar graph	Represented with a continuous curve (line graph)
Probabilities of all outcomes sum up to 1.	Area under the curve equals 1.
Measured in whole numbers (integers)	Measured in real numbers (also fractions).
eg: No. of students in a classroom (0, 1, 2, ...)	Height of students (5.5 ft, 5.10 ft, 5.11 ft, ...)

Q.3 Explain Central limit theorem.

—y The Central limit Theorem (CLT) is a fundamental concept in statistics and probability. It states that, for a sufficiently large sample size, the sampling distribution of the sample mean (or sum) of independent, identically distributed random

variables will approximate a normal distribution, regardless of the original distribution of the population.

> The larger the sample size, the closer the sampling distribution is to a normal distribution. A common thumb rule is that a sample size of 30 or more is generally sufficient.

> The data points in the sample must be independent of each other.

> The mean of the sampling distribution will be equal to the population mean (μ)

> The variance of the sampling distribution will be equal to the population variance divided by the sample size (σ^2/n).

— If $X_1, X_2, \cdots, X_3$ are independent, identically distributed random variables with mean μ and variance σ^2, then the sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

follows a normal distribution as $n \to \infty$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Q.4 What do you mean by hypothesis testing?

Explain in content of machine learning example.

-> Hypothesis Testing is a statistical method used to make decisions or interferences about population parameters based on sample data. It involves formulating two competing hypotheses and determining which hypothesis is supported by the data.

(i) Null Hypothesis (H_0) : A statement that there is no effect or no difference. It represents a baseline or default position.

(ii) Alternative Hypothesis (H1 or Ha) : This represents the effect or difference we are testing for. It is what we hope to support with our data.

-> In ML, hypothesis testing can be applied to evaluate the performance of models, compare different models.

eg: Comparing Two Models.
Let's say you are developing two different ML models for predicting house prices : Model A (linear Regression) and Model B (decision trees).

(i) Define Hypothesis
* Null Hypothesis : Mean prediction error of Model A equals that of Model B (no difference).

- Alternative Hypothesis : The Mean prediction error of Model A does not equal Model B.

(ii) Select significance level : set $\alpha = 0.05$.

(iii) Collect Data : Evaluate both models using a validation dataset and calculate the mean prediction error for each model.

(iv) Calculate the Test Statistic : Compare the performance metrics of the two models.

(v) Determine the P-Value. :

(vi) Make a Decision
- If the p-value is less than α, reject H_0, indicating a significant difference.
- If the p-value is greater than α, fail to reject H_0, indicating no significant difference.

Q.5 Explain Bayesian method for classification along with Bayes' theorem.

—> The Bayesian method for classification is a probabilistic approach that applies Bayes' theorem to predict the class of a given instance based on its features.

—> Bayes' Theorem

It provides a way to update

the probability of a hypothesis based on new evidence. The theorem is mathematically expressed as:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} ;$$

- $P(H|E)$: Posterior Probability - the $P(H)$ given evidence E.
- $P(E|H)$: Likelihood - the $P(E)$ given that hypothesis H is true.
- $P(H)$: Prior Probability - the initial $P(H)$ before observing evidence E.
- $P(E)$: Marginal Probability - the total probability of evidence E across all hypotheses.

→ Steps in Bayesian Classification :-

1. Identify the classes and features relevant to the classification task.
2. Calculate Prior Probabilities ($P(H)$)
3. Calculate Likelihoods ($P(E|H)$)
4. Compute Posterior Probabilities.
5. Assign the instance to the class with the highest posterior probability :

$$\text{class} = \arg\max_{H} P(H|E).$$

eg: Classifying email as either spam or not spam based on the presence of certain words.

* Also, Naive Bayes Classifier is a specific implementation of Bayesian Classification.

Q.6 What is classification ? Explain in detail taking an appropriate real-world example.

-4 Classification is a supervised learning technique in ML where the goal is to assign a class label to new instances based on a training dataset. The process involves learning a function or model that can accurately predict the category of new data points based on the patterns observed in the training data. Key concepts includes classes (categories), features, training phase and prediction phase.

eg: Email Spam Detection

(i) Problem : Classify incoming emails as "Spam" or "Not Spam" (also called ham).

(ii) Features
- Presence of specific words ("free", "win")
- Number of links or attachment
- Sender's email address.
- Frequency of capital letters.

(iii) Classes (Categories)
- Spam : Unwanted or malicious emails
- Not Spam (Ham) : Legitimate emails

(iv) Training Process : A labeled dataset is used to train the model, allowing it to learn patterns typical of spam.

(v) Prediction : For new emails, the model extracts features and assigns probabilities to each class. An email with a 0.85 probability of being "Spam" is classified as such.

(vi) Performance Evaluation : Metrics like accuracy, precision, recall and F1-score assess the model's effectiveness on both categories.

* Also, types of classification are Binary, Multi-class and Multi-label classification.

Q.7 Discuss decision tree classifier taking the reference of real-world dataset. (Mention the link and feature details of dataset explicitly).

→ The Decision tree classifier is a supervised ML algorithm used for both classification and regression tasks. It is used to predict outcomes based on a set of input features. It operates by splitting the data at each node based on feature values, aiming to maximize the separation between different classes.

eg: Bank Marketing Dataset (UCI Machine learning Repository).

It aims to predict whether a

client will subscribe to a term deposit ('yes' or 'no'). The dataset includes 21 features, such as:

- Demographics: Age, job type, marital status and education level.
- Financial Status: Balance, housing loan, and personal loan.
- Campaign Information: No. of previous contacts, last contact duration, communication type and more.
- Economic Indicators: Employment rate, consumer confidence index and Euribor rates.

→ The Process of Decision Tree :-

(i) Start at the root node: The algorithm analyzes features to find the best split (eg, assessing housing loan or marital status).

(ii) Recursive Splitting: Data is divided recursively until a stopping ~~criterias~~ criterion is met, such as maximum depth or minimum samples per leaf.

(iii) Prediction: leaf nodes assign the most probable outcome ('yes' or 'no').

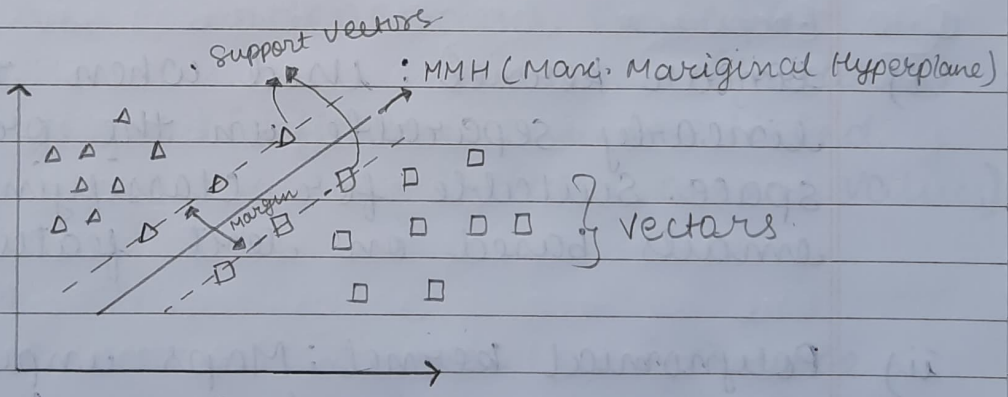
Example Insights:

- Longer call durations often correlate with higher subscription rates.

→ Previous successful campaigns can boost likelihood of subscription.

Q.8 Explain Support Vector Machines for classification. Also discuss kernel methods in SVM.

→ Support vector Machines are supervised learning algorithm used primarily for classification tasks but can also be adapted for regression. The primary objective of SVM is to find the optimal hyperplane that separates data points of different classes in a high-dimensional space.



→ The Process of SVM for classification :-

i) Separating Hyperplane
For a two-class problem, SVM identifies the hyperplane that best separates the data points of both classes. (For 2-D, the hyperplane is a line, for 3-D it is a plane).

ii) Maximizing the Margin.

SVM selects the hyperplane that has the maximum distance (margin) from the closest data points of each class (called support vectors). This helps in reducing generalization errors, ensuring the model performs well on unseen data.

iii) Support Vectors.

Only the closet points to the hyperplane influence its position and orientation. These points are called support vectors.

—↳ Kernel Methods in SVM are :-

i) linear kernel : used when the data is linearly separable in the original feature space. Suitable for classifying spam emails based on text features.

ii) Polynomial kernel : Maps input data into a higher polynomial space. (pattern recognition)

iii) Sigmoid kernel : Similar to the activation function in NN. Suitable for speech recognition tasks and performs well in NN.

Q.9 Discuss multiple linear regression with a numerical example.

—↳ Multiple linear Regression (MLR) is a statistical method that models the

relationship between a dependent variable and multiple independent variables. It extends simple linear regression by incorporating more than one feature to make better predictions.

- Mathematical Model of MLR:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon \;;$$

- y = dependent variable (target)
- β_0 = intercept
- $\beta_1, \beta_2, \dots, \beta_n$ = coefficients of independent variables
- $x_1, x_2, \dots, x_n$ = independent variables (features)
- ε = error term (difference in predicted and actual values).

eg: Predicting House Prices
Suppose we are trying to predict the price of a house based on two features: square footage and number of bedrooms.

Data:

Square footage (x_1)	Bedroom (x_2)	Price (y) in $
1400	3	300,000
1600	3	320,000
1700	4	350,000
1875	4	370,000
1100	2	220,000

$\therefore \hat{y} = 50{,}000 + 120.(x_1) + 15{,}000.(x_2)$;
↳ after fitting the data using Python or excel.

> Intercept, $\beta_0 = 50,000$
> Co-efficient for Square footage, $\beta_1 = 120$
> Co-efficient for number of bedrooms, $\beta_2 = 15,000$

If we want to predict the price of a house with 1800 square feet and 3 bedrooms, we substitute these values in the eqn:

$$\hat{y} = 50,000 + 120 \cdot (1800) + 15,000 \cdot (3)$$
$$\therefore \hat{y} = 50000 + 216000 + 45000 = 311000$$

⇒ Thus, the predicted price of house is $311,000.

Q.10 Explain Logistic Regression.

—↳ Logistic Regression is a supervised ML algorithm used for binary classification tasks, where the goal is to predict the probability of an instance belonging to one of two classes. (eg; 0 or 1, spam or not spam). Despite its name, it is used for classification, not regression.

—↳ Logistic Regression Model :-

It predicts the probability of the dependent variable 'Y', belongs to a specific class (eg, $y=1$) as a function ~~to~~ of the input features. This is done using the logistic function (sigmoid function) which maps any real-valued input to a range between 0 and 1.

$$\therefore P(y=1|X) = \hat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}};$$

- $P(y=1|X)$ = predicted probability that $y=1$
- β_0 = intercept
- $\beta_1, \beta_2, \cdots, \beta_n$ = coefficients for independent variables
- $x_1, x_2, \cdots, x_n$ = independent variables.
- e = Euler's number.

- The model classifies a data point into one of the two classes based on a decision boundary. Typically, if the predicted probability is:

$$P(y=1|X) \geq 0.5 \Rightarrow \text{Class 1.}$$
$$P(y=1|X) < 0.5 \Rightarrow \text{Class 0.}$$

It works well on linearly separable data and provides the probability of belonging to a particular class.

But it performs poorly when the classes are not linearly separable and it can only handle two ~~set~~ classes without modification (one-vs-rest for multi-class).