Unit – 4

Probability and Classification

Syllabus

- Introduction to Probability
- Introduction to Statistics
- Need of probability and statistics in Data Mining
- What Is Classification?
- General Approach to Classification
- Bayes Classification Methods, Bayes' Theorem
- Naıve Bayesian Classification
- Rule-Based Classification Using IF-THEN Rules for Classification
- Rule Extraction from a Decision Tree

Introduction to Probability

- Probability refers to the likelihood or chance of a particular event occurring based on available data.
- Probability is used to understand how likely an event is to happen based on the data available.
- It is a key concept in statistical analysis, machine learning, and data mining, and can be used to gain valuable insights into patterns and trends in large datasets.
- Probability is a key concept in data warehouse that helps businesses to better understand the behaviour of their data and make more informed decisions based on that information.

Probability Contd…

- For example, a data warehouse may be used to analyse customer behaviour and determine the likelihood of a customer making a purchase based on their past behaviour. By calculating the likelihood ratio of a customer making a purchase, the data warehouse can provide valuable insights into customer behaviour and help businesses to make informed decisions about marketing and sales strategies.

- A social media platform can use probability to predict which posts or ads a user is likely to engage with. By analyzing a user's past behaviour, such as which posts they have liked or shared, the platform can predict the likelihood of the user engaging with a particular post or ad in the future.

Introduction to Statistics

Statistics is a branch of mathematics that deals with collecting, analysing, interpreting, and presenting data.

Statistics is used to summarize and analyse large amounts of data to provide insights into the organization's operations and performance.

Data mining involves the use of statistical methods to discover patterns and relationships in large datasets.

Some common statistical techniques used in data warehouse and data mining include descriptive statistics, inferential statistics, correlation analysis, regression analysis, and hypothesis testing.

Statistics Contd..

Following are the certain use-cases of Statistics in the field of Data Warehousing & Data Mining:

Statistics can be used to summarize large amounts of data into meaningful information by performing various operations such has mean, median, mode, standard deviation, and variance.

Statistics can provide insights into the distribution of data and identify potential outliers or anomalies.

For example, a retailer might use statistics to estimate the average purchase amount of a particular customer segment based on a sample of customer purchase data.

Financial Institution might use regression analysis(A Statistics Method) to predict the likelihood of a loan default based on historical customer data.

Classification

Classification is a popular technique in data mining that involves the process of predicting the class or category of a target variable based on the values of one or more input variables.

The goal of classification is to build a predictive model that can accurately classify new instances or observations based on the training data.

In classification, the target variable is typically a categorical variable, such as "yes" or "no", "true" or "false", or a label such as "spam" or "not spam".

The input variables are often numeric or categorical, and can be discrete(Data which can be divided into categories) or continuous(Data which can be represented in Decimal Format).

For example, a classification model might use the age, income, and occupation (input variables) of a customer to predict whether they are likely(likeliness – target variable) to purchase a particular product.

General Approach to Classification

The classification process involves several steps, including:

Data preparation: This involves collecting, cleaning, and preparing the data for analysis. This may involve selecting relevant input variables, dealing with missing or invalid data, and transforming the data into a format that can be used for analysis.

Model building: This involves selecting an appropriate classification algorithm and training the model using the prepared data. The model is typically evaluated using a performance metric such as accuracy, precision, recall, or F1-score.

Model evaluation: This involves evaluating the performance of the model on a separate test dataset to determine how well it can predict new instances or observations.

Model deployment: Once the model has been evaluated and found to be accurate and reliable, it can be deployed in a production environment to classify new instances or observations in real-time.

Algorithms used for Classification

Various Algorithms that can be used for Classification are as follows:-

Decision Trees(Seen in Unit 3)
Bayesian Classifiers(We will learn this algorithm in this unit)
Neural Networks
K-Nearest Neighbour(Seen in Practical)
Support Vector Machines
Linear Regression
Logistic Regression

Bayes Classification Methods

Bayesian classification is a statistical classification technique that uses Bayes' theorem to classify new instances or observations into one of several predefined classes.

Bayesian classification can be used to classify data with both discrete and continuous input variables.

Bayesian classification can handle missing data, noisy data, and imbalanced data sets.

They are powerful and flexible tools for classification that can be used in a wide range of applications. Bayesian classification can be used in a wide range of applications, such as image recognition, spam filtering, medical diagnosis, and credit risk assessment.

There are several different methods for implementing Bayesian classification, including:

Naive Bayes
Bayesian Networks
Bayesian Belief Networks
Hierarchical Bayesian Models

Naïve Bayes Classification

This Method is called "naive" because it assumes that the input variables are independent of each other, which makes the calculations easier and faster.

It works by calculating the prior probability of each class based on the training data, and then using Bayes' theorem to calculate the conditional probability of each class given the input variables.

The class with the highest probability is then assigned to the new observation.

Naive Bayes classification can be used for both discrete and continuous input variables, and it is particularly effective for text classification and spam filtering.

One of the key advantages of Naive Bayes classification is that it requires relatively little training data and can be trained quickly.

It also works well with high-dimensional data and is not sensitive to irrelevant input variables.

Naïve Bayesian Equation/Rule

This procedure is based on Bayes Rule, which says: if you ha v e a hypothesis h and data D which bears on the hypothesis, then:

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

(1) P(h): independent probability of h: prior probability
(2) P(D): independent probability of D
(3) P(D|h): conditional probability of D given h: likelihood
(4) P(h|D): cond. probability of h given D: posterior probability

Rule Based Classification

Rule-based classification is a type of approach used in machine learning and artificial intelligence to make predictions or decisions based on a set of predefined rules.

These rules are created by experts in the specific domain and are typically based on prior knowledge and experience in that field.

These rules are then used to categorize new data points or instances into different classes or categories.

For example, in a spam filter, the rules might specify that if an email contains certain words or phrases such as "discounts" or "free trial," then it is likely to be classified as spam. If an email does not contain these words, then it is more likely to be classified as legitimate.

Rule Based Classification – If Then Rules

If-then rules are a key component of rule-based classification, which is a type of approach used in machine learning and artificial intelligence to make predictions or decisions based on a set of predefined rules.

An if-then rule is a simple statement that consists of two parts: an "if" condition and a "then" conclusion.

The "if" part specifies a condition that must be satisfied for the rule to be applied, while the "then" part specifies an action or decision to be taken if the condition is met.

Here's an example of an if-then rule in the context of a classification problem:

If a patient has a high temperature and a cough, then they are likely to have the flu.
In this rule, the "if" part specifies the conditions that must be met (i.e., high temperature and cough), while the "then" part specifies the conclusion (i.e., likely to have the flu).

Rule Based Classification – If Then Rules

Rule-based classification is a method of assigning a label or category to a new data point based on a set of predefined rules. These rules can be expressed using if-then statements.

Suppose we have a dataset of customer transactions at a retail store, and we want to predict which transactions are likely to be fraudulent.

We might come up with the following set of if-then rules based on our domain knowledge and experience:

If a transaction is larger than $1000 and is made by a first-time customer, then it is likely to be fraudulent.
If a transaction is made with a credit card that has been reported stolen, then it is likely to be fraudulent.
If a transaction is made from a country that is different from the customer's billing address, then it is likely to be fraudulent.

Rule Based Classification – If Then Rules

When a new transaction is presented for classification, we can apply each of these rules in turn to determine whether it is likely to be fraudulent.

For example, if we encounter a transaction for $1500 made by a first-time customer, we can apply the first rule and classify it as likely to be fraudulent.

By using a set of if-then rules, we can create a straightforward and interpretable classification system that can be easily understood and updated as needed.

However, as with any classification system, the quality and completeness of the rules are critical to its accuracy and effectiveness.

If – Then Rules (Another Example)

Let's say we want to classify animals as either "mammals" or "non-mammals" based on their characteristics.

We might come up with the following set of if-then rules:

If an animal has hair and produces milk, then it is a mammal.
If an animal lays eggs, then it is a non-mammal.
If an animal has scales and breathes underwater, then it is a non-mammal.

When a new animal is presented for classification, we can apply each of these rules in turn to determine its label. For example, if we encounter an animal that has hair and produces milk, we can apply the first rule and classify it as a mammal.

Rule-Based Classifier: Rule Coverage and Accuracy - Example

IF **(Status=Single)** THEN **Class=No**

Coverage = 4/10 = 40%

Accuracy = 2/4 = 50%

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	**Single**	125K	**No**
2	No	Married	100K	**No**
3	No	**Single**	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	**Single**	85K	**Yes**
9	No	Married	75K	**No**
10	No	**Single**	90K	**Yes**

Rule-Based Classifier: Rule Coverage and Accuracy

- A rule r **covers** an instance x if the attributes of the instance satisfy the condition of the rule.

Coverage of a rule:

- Fraction of tuples that satisfy the condition of a rule.

Accuracy of a rule:

- Fraction of tuples that satisfy the condition that also satisfy the consequent of a rule.

$\mathbf{n_{covers}}$: # of tuples covered by rule R

$\mathbf{n_{correct}}$: # of tuples correctly classified by rule R

$\mathbf{coverage(R) = n_{covers} / |D|}$ /* D: training data set */

$\mathbf{accuracy(R) = n_{correct} / n_{covers}}$

Rule-Based Classifier - Example

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: IF (Give Birth = no) AND (Can Fly = yes) THEN Class=Birds
R2: IF (Give Birth = no) AND (Live in Water = yes) THEN Class=Fishes
R3: IF (Give Birth = yes) AND (Blood Type = warm) THEN Class=Mammals
R4: IF (Give Birth = no) AND (Can Fly = no) THEN Class=Reptiles
R5: IF (Live in Water = sometimes) THEN Class=Amphibians

How a Rule-Based Classifier Works?

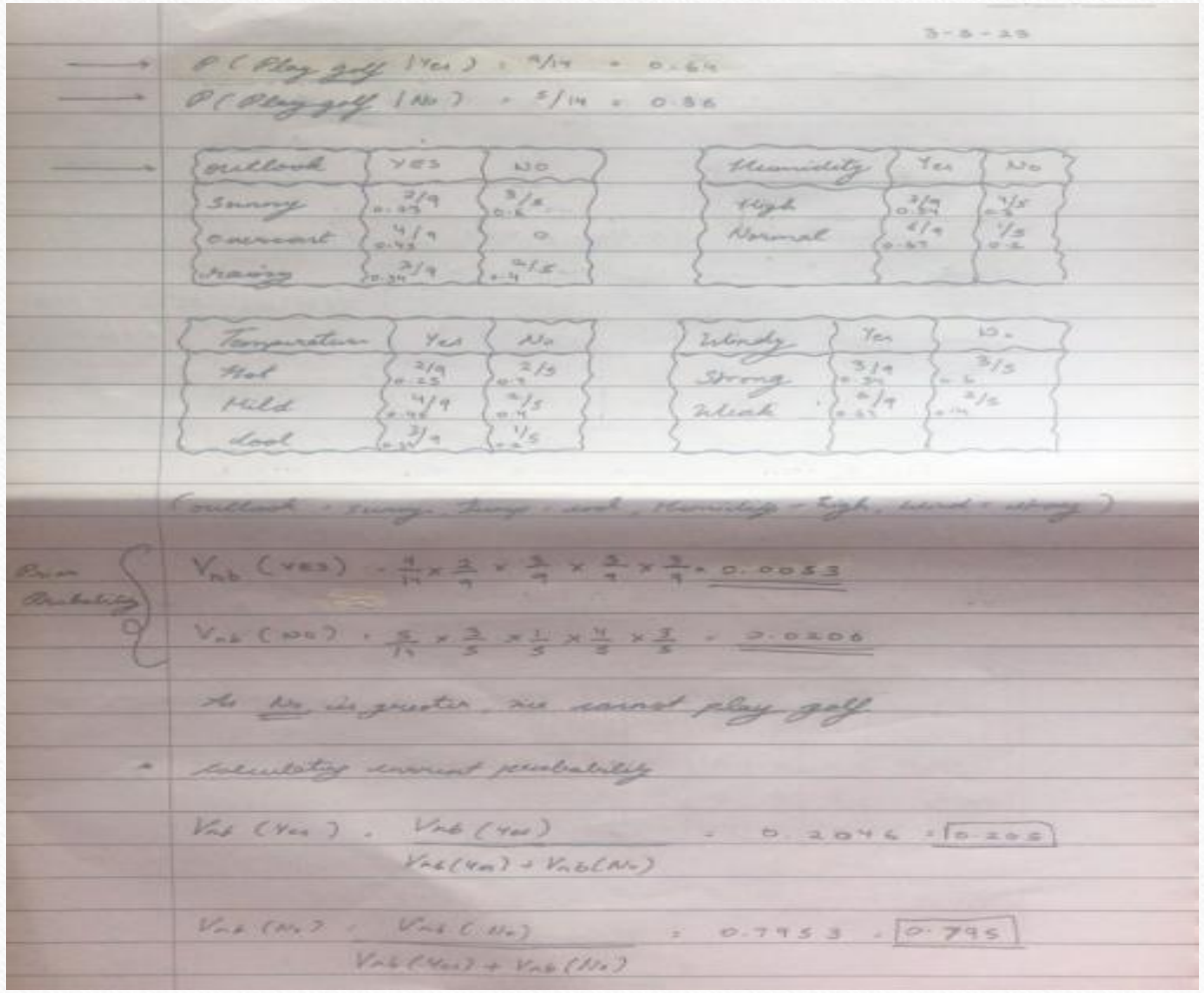
- **A rule-based classifier classifies a tuple based on the rule triggered by the tuple.**

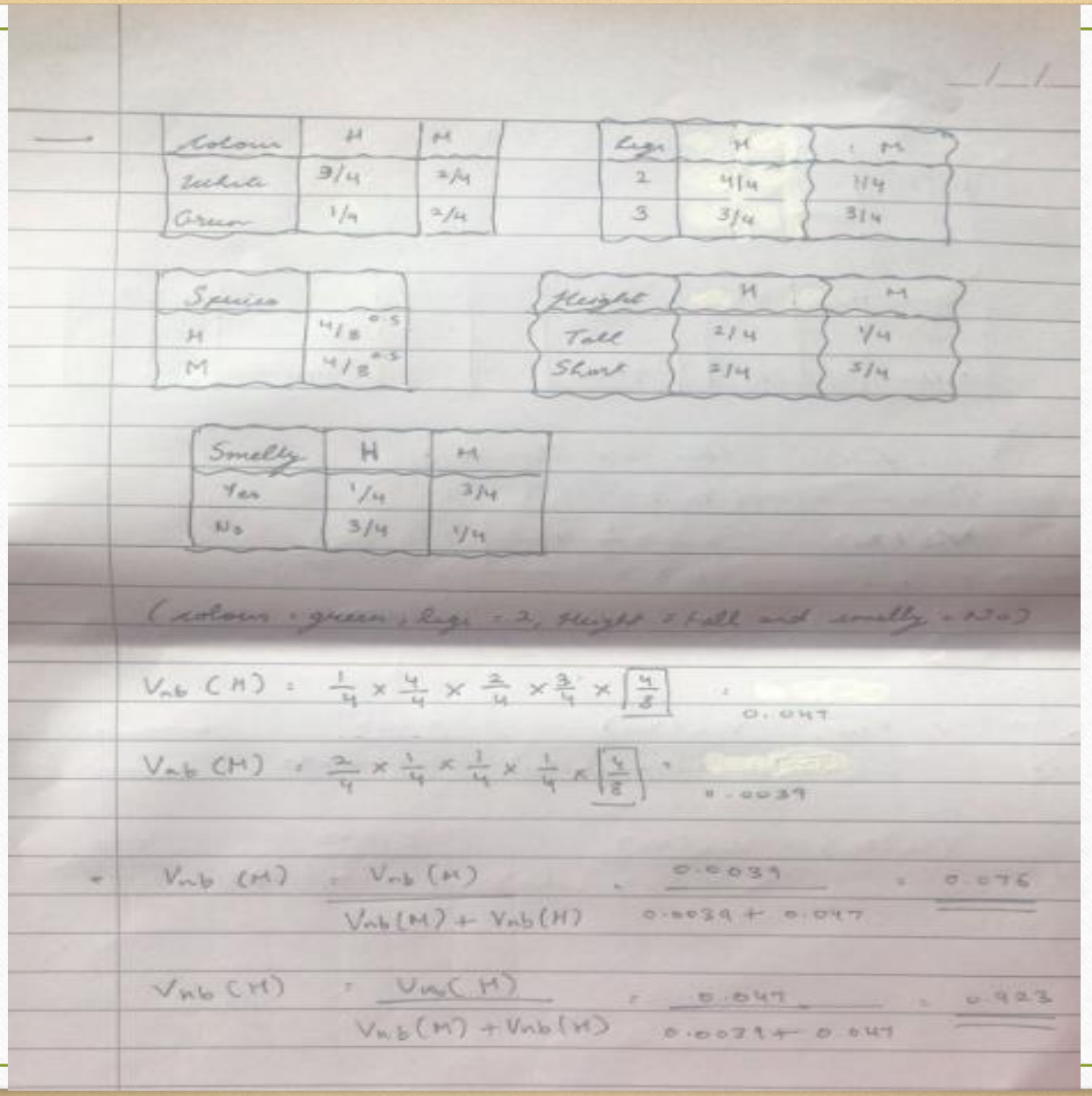
Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

R1: IF (Give Birth = no) AND (Can Fly = yes) THEN Class=Birds
R2: IF (Give Birth = no) AND (Live in Water = yes) THEN Class=Fishes
R3: IF (Give Birth = yes) AND (Blood Type = warm) THEN Class=Mammals
R4: IF (Give Birth = no) AND (Can Fly = no) THEN Class=Reptiles
R5: IF (Live in Water = sometimes) THEN Class=Amphibians

- A **lemur** triggers rule R3, so it is classified as a **mammal**
- A **turtle** triggers both R4 and R5
  - Since the classes predicted by the rules are contradictory (**reptiles** versus **amphibians**), their conflicting classes must be resolved.
- A **dogfish shark** triggers none of the rules
  - we need to ensure that the classifier can still make a reliable prediction even though a tuple is not covered by any rule.

Naïve Bayes Example





Colour	H	M
White	3/4	2/4
Green	1/4	2/4

Legs	H	M
2	4/4	1/4
3	3/4	3/4

Species	
H	$4/8^{0.5}$
M	$4/8^{0.5}$

Height	H	M
Tall	2/4	4/4
Short	2/4	3/4

Smelly	H	M
Yes	1/4	3/4
No	3/4	1/4

(colour = green, legs = 2, Height = tall and smelly = No)

$V_{nb}(M) = \frac{1}{4} \times \frac{4}{4} \times \frac{2}{4} \times \frac{3}{4} \times \boxed{\frac{4}{8}} = 0.047$

$V_{nb}(H) = \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \boxed{\frac{4}{8}} = 0.0039$

$V_{nb}(M) = \frac{V_{nb}(M)}{V_{nb}(M) + V_{nb}(H)} = \frac{0.0039}{0.0039 + 0.047} = 0.076$

$V_{nb}(H) = \frac{V_{nb}(H)}{V_{nb}(M) + V_{nb}(H)} = \frac{0.047}{0.0039 + 0.047} = 0.923$

__/__/____

Yes	No
5/10	5/10
= 0.5	= 0.5

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
SUV	1/5	3/5
Sports	4/5	2/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

New Instance : (Red, SUV, Domestic)

→ P(New Instance | Yes)

$= \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times \boxed{\frac{5}{10}} = 0.024$

→ P(New Instance | No)

$= \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \boxed{\frac{5}{10}} = 0.072$

Reference Questions

1. What is probability and statistics elaborate with example

2. What Is Classification ? what is Bayes Classification Methods?

3. Explain Naıve Bayesian Classification with example

4. Explain Rule-Based Classification using IF-THEN Rules for Classification

5. Explain Naïve Bayesian equation to calculate the posterior probability for each class, For given Data Problem.