17. What is disenete attributes?

Ans:

A disenete attributes has a finite on Countably infinite set of values; which may on may not be nepresented as integens.

18. What is Continuous attributes?

Ans:

A Continuous attribute has a numenic on Continuous attribute values.

Disenete: All the values are Categonical

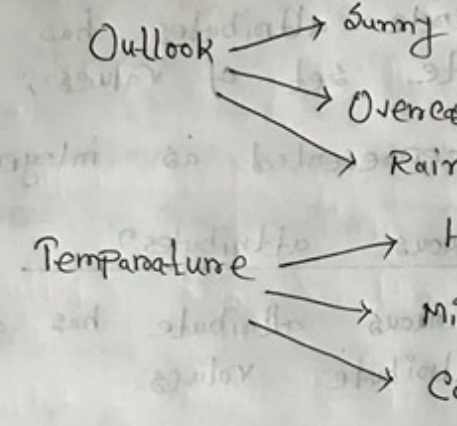
The Weathen Problem

Outlook	Temperature	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No

Table: Weathen Table

- এখানে কয়টা কলাম হলো feature বা attribute.
- উপরের টেবিলে 4 টা attribute আছে
  i. Outlook
  ii. ~~Temp~~ Temperature
  iii. Humidity
  iv. Wind
- Play নামক কলামের কয় class.
- এই table টা হলো historical data

- প্রতিটি feature এর feature value আছে,
যেমন:



- Permutation; Combination বললে

$3 \times 3 \times 2 \times 2 = 36$ টা Combination সম্ভব।

- Decision Tree এর ক্ষেত্রে এই Combination গুলোর rules থাকবে sort out করে বলে দেয়া যায়, অর্থাৎ rules optimal হবে।

Machine Learning Models Rules এর মাধ্যমে learn করে,

1. If Outlook = Sunny & Humidity = High then Play = No
2. If Outlook = Sunny & Humidity = Normal then Play = Yes
3. If Outlook = Overcast then Play = Yes
4. If Outlook = Rain & Wind = Strong then Play = No
5. If Outlook = Rain & Wind = Weak then Play = Yes

=> এই rules generate করার সময় দেখা গেছে Temparature less important feature.

এগুলো table এ এগুলো Numeric value দিয়ে represent করা হয়।

Table: Weather Data with some Numeric Attributes

Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	Weak	No
Sunny	80	90	Strong	No
Overcast	83	86	Weak	Yes
Rain	70	96	Weak	Yes
Rain	68	80	Weak	Yes
Rain	65	70	Strong	No
Overcast	64	65	Strong	Yes
Sunny	72	95	Weak	No
Sunny	69	70	Weak	Yes
Rain	75	80	Weak	Yes
Sunny	75	70	Strong	Yes
Overcast	72	90	Strong	Yes
Overcast	81	75	Weak	Yes
Rain	71	91	Strong	No

পরের Rules গুলা কেমন ⇒

1. If Outlook = Sunny & Humidity > 75 then Play = No
2. If outlook = Sunny & Humidity ≤ 75 then Play = Yes
3. If outlook = OverCast then Play = Yes
4. If Outlook = Rain & Wind = Strong then Play = No
5. If outlook = Rain & Wind = Weak then Play = Yes.

- এখানে 75 কে Threshold Point.
- 75% এর উপর এবং Humidity কে Separate করছে,
- Threshold Point কে CutPoint / Discrete Discreatice Point বলা হয়,

- Define Concept.

Ans: Concept is the thing to be learned.

- Define Concept Description

Ans: The output Produced by a learning classifier.

- Define Instances

Ans: Things that are to be classified or associated or clustered.

Dataset কে matrix আকারে represent করা হয়, যেখানে একপাশে থাকে instances আর অন্যপাশে থাকে attributes

What is Flat File

Ans:

A flat file database is one that only contains a single table of data.

Define Attribute

Ans:

Attribute is a data field, representing a characteristic or feature of a data object.

What are the steps of data Preprocessing

Ans:

Data PreProcessing has four steps

a. Data cleaning: Missing values, noisy data.
b. Data Integration: Redundancy, Connelation, instance duplication.
c. Data Reduction; Dimensionality Reduction.
d. Attribute Subset selection.

Q. Draw block diagram of Pattern mining Process.

Ans

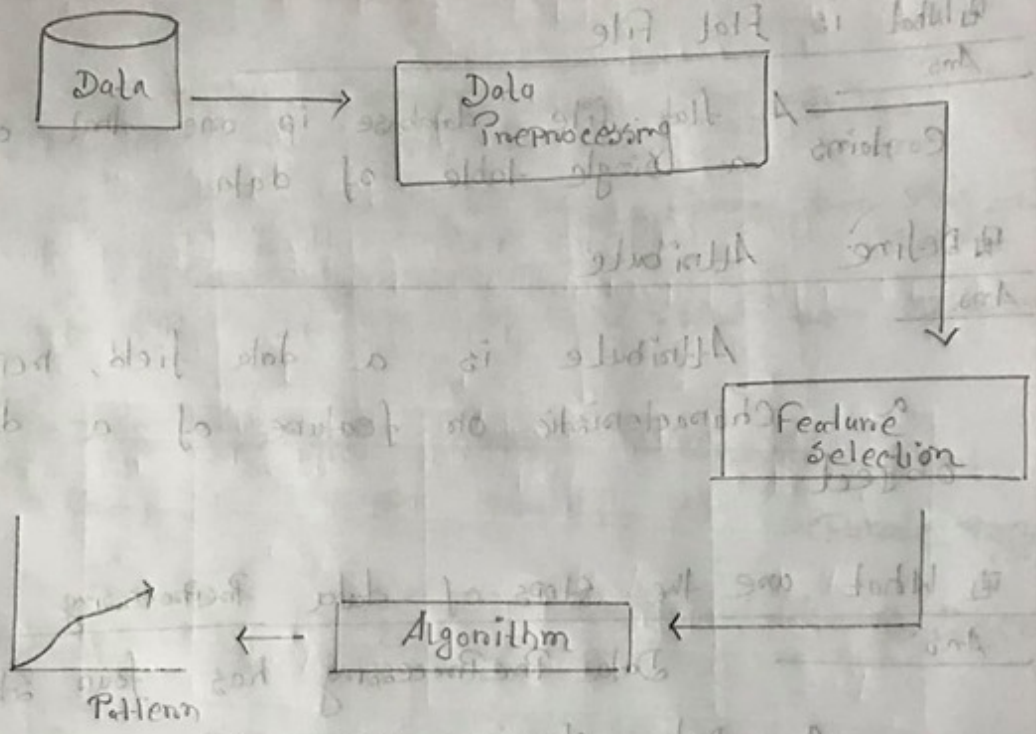


Fig: Data Mining Process.

Q. What is ARFF

Ans:

ARFF stands for "Attribute Relation file Format".

It's mainly an ASCII text file that describes a list of instances sharing a set of attributes.

Write the Work Process of KNN algorithm

Ans:

KNN is a classification algorithm. It is supervised. It takes a bunch of labelled Points & uses them to learn how to label other Points. To label a new Point it looks at the labelled Points closest to the new Point & has those neighbors vote.

Write the KNN algorithm.

Ans:

~~Input : D = { x_1, ~~x~~.~~

Input : $D = \{x_1, \ldots, x_i, \ldots, x_n\}$

Output : KNN Classifier, KNN

Method:

a. find $x \in D$, that identify the k nearest neighbours, Regardless of class label, C_i.

b. Out of these instances, $X = \{x_1, x_2, \ldots, x_k\}$ identify the numbers of instances, k_i that belong to class C_i. where $i = 1, 2, n$; $\Sigma_i k_i = k$

c. Assign X_{test} to the class C_i with the maximum number of k_i of instances.

D হলো training data
x হলো Nearest neighbours of test data

KNN এ dataset কে তিনভাগে ভাগ করা হয়.

i. Training set
ii. Validation set
iii. Test set.

যেকোনো dataset কে কয়েকভাগে ভাগ করা যায়.

	Training set	Validation set	Test set
1.	80 %	10 %	10 %
2.	70 %	15 %	15 %
3.	70 %	x	30 %

Training set

a. Model কে train এবং tune করে
b. Classifier construct করে

Validation set

a. Hyper-parameter tune এর জন্য লাগে
b. Accuracy বাড়ায়,
c. Validation Accuracy maximize করে,
d.

Test Set

a. Test set actual accuracy বলে।

KNN এ distance বের করা হয়। এই distance দুভাবে বের করা যায়.

a. Eudlidean Distance
b. Manhattan Distance

Eudlidean Distance:

(x, y) একটা Point ; (a, b) আর একটা Point, এদের মধ্যকার Eudlidean Distance হল

$$(x,y),(a,b) = \sqrt{(x-a)^2 + (y-b)^2}$$

Manhattan Distance:

(x, y) একটা Point ; (a, b) আর একটা Point, এদের মধ্যকার Manhattan Distance হল

$$(x,y),(a,b) = |x-a| + |y-b|$$

The following dataset (Table 1) with two features (acid durability & strength) to classify weather a special tissue paper is good or not.

Table 1

A_1	A_2	Class
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now, the factory produces a new tissue paper (X_{new}) that pass laboratory test with $A_1 = 3$, $A_2 = 7$. So classify the X_{new} using k-Neanest neighbor (KNN) classifier, where $k=3$ & distance function is Distance [Spring 19]

Ans:

A_1	A_2	Distance
7	7	$\sqrt{(7-3)^2+(7-7)^2} = \sqrt{42} = 4$
7	4	$\sqrt{(7-3)^2+(4-7)^2} = \sqrt{25} = 5$
3	4	$\sqrt{(3-3)^2+(4-7)^2} = \sqrt{9} = 3$
1	4	$\sqrt{(1-3)^2+(4-7)^2} = \sqrt{13} = 3.60$

Number of instances	Distance	Nearest ?	Majority Class	Majority voting
1	4	Yes	Bad	
2	5	No	x	
3	3	Yes	Good	Good
4	3.60	Yes	Good	

So, X_{new} belongs to class "Good"

Explain the difference & similarity between classification & regression.

Ans.

Differences between Classification Vs Regression

Classification	Regression
a. Group the output into a class.	a. Predict the output value using training data.
b. If it is discrete/ categorical variable then it is classification Problem.	b. If it is real number/ continuous then it is regression Problem.
c. Data is labelled into one ~~or~~ of two or more classes.	c. Prediction of a quantity.

Similarities between Classification & Regression

a. Both are Supervised learning

b. Develop Predictive model based on both input & output data

* Classification	→ Supervised Learning
* Regression	→ Supervised Learning
* Clustering	→ Unsupervised Learning

Write the disadvantage of KNN -

Ans: KNN is a Lazy learner. It does not learn anything from training data. It simply uses the training data itself for classification.

Write the drawback of KNN

Ans: It's Complexity is $O(KN)^2$.

Where

N = Total number of neighbours

K = Number of nearest neighbours.

It takes too much time. For new test data it runs from the beginning.