	Cowrse
	Unit 1 : Introduction and Ethical theories 4
	Unit 2: Ethics in IT societies, Intellectual right
	The state of the s
	_ Unit 3: Professional Relationship Responsibilities,
<u> </u>	
-	Unit 4: Online crime hacking, legal aspects
-	professional ethics
	proposition
	Ethical Theories.
	Virtue Ethics.
	lonsequentalism
	De : local la
	Philosophy

Introduction to Data Mining.	al sent
8	
Big data: It has three characteristics:	,
- If data is coming at very high speed (velocity)	Augustin .
- quantity is large (volume)	_
- not a single type of data (varacity)	_
Eg. twitter data, facebook data.	sid
What is data science?	-
Difference between data analytics, data science,	-
data mining?	_
Human Computer Interaction: If machine is not	
able to interact with human beings, then it is waste	_
KDD: Knowledge Data Discovery. 60-70% effort	0
Selection Preprocessing (like cleaning and feature normalisation), transformation, data mining,	-
extraction), transformation, acta mining,	5
interpretation/evaluation	LOT
transformation: converting data in a particular	
format so that algorithms can be applied	
Il interpretation is wrong then we will go back	
If interpretation is wrong, then we will go back and evaluate at each step: Postprocessing.	
	ed.
PCA, SVD, correlation analysis - preprocessing techniques	

Higher dimension analysis is not possible so dimension reduction. Eg. gene data. Data Mining Jasks.
Prediction Methods (like supervised)
Description Methods (like unsupervised)

Data Mining Yasks.
Classification (Predictive).
Classification (Predictive). Decision trees, Naive Bayes
constering (descriptive) diving into groups but final result can't be said
result can't be said
result can't be said we will use similarity measures like cosine similarity KNN, hierarchial clustering
KNN, hierarchial chrotering
J. C.
Association Rule Discovery (Descriptive)
We see association of different item sets.
like if we go to purchase bread. there is high probability to purchase butter. (No timeline)
probability to purchase butter. (No timeline)
and are I is a second for the second
Sequential Pattern Discovery. (Descriptive) There is one timeline which is always there
There is one timeline which is always there
Eg, stock market data, fire alarm runging, due
Eg. stock market data, fire alarm runging, due to election results market value goes high or low.
pring has princes i tady and the commentalities
Regression (Predictive) whenever data set is continuous
(Linear Logistic Quantum).
(Linear, Logistic, Buantum). Used in stock market prediction.
Vsca it stock with provided
Deviation Detection (Predictive)
Eg. credit card frand detection
Normal behaviour is stored and changes are checked

Fraining set is also called record data test.
Training set is also called selection
Association Rule Discovery
Support and confidence : use and aigure sules
Association Rule Discovery: Support and confidence: use and define rules
The state of the s
Challenges of Data Mining.
a day by day
1) Scalability: Data size is attitudence of
a) Dimensionality: Data has many authorise many
1) Scalability: Data size is increasing day by day a) Dimensionality: Data has many dimensions (many attributes) High dimension data needs to be reduced
in les dimensions
3) complex and heterogeneous data:
- John Janes Land washing to the first
4) Data Guality: is degrading due to anamoly and noise
5) Data ownership and distribution: Distribution from
where no one knows, big issue Drigin of whatsapp
message can't be detected immediately
6) brivary preservation This is a major inne , so no
grage winds some village and the second of
Thaming data: Data that is coming and going
continuously called tunnel
Eg. Juritter analysis and
1
nough the turnel can be analysed.
through the turnel can be analysed
- Eutledian distance is a live of
I unation distance
Mintowski distance $\frac{1}{2}(x-\bar{x})(x-\bar{x})$
$\frac{1}{n-1} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}_i \right) \left(x_i - \overline{x}_i \right)}{\left(x_i - \overline{x}_i \right)}$

Assignment Buestion:
Calculate the Euclidean, Minkowski and Mahalanobis
listance with following parameters.
fistance with following parameters. P1 (2,4), P2 (4,2), P3 (5,5), P4 (4,2).
Also, draw the relationship between all three.
Tarcard Coefficient (I coefficient) Only for binary
Jaccard Coefficient (I coefficient) Only for binary SMC: no. of matches / no. of attributes.
I = no. of 16 matches/ no. of not both zero attribute
values.
Cosine Similarity. cos (d, d2) = (d1.d2)/ 1 d111 d2
peningly had an ad
Extended Jacord Coefficient For continuous or count attributes.
attributes
To the first
Correlation: measures the linear relationship between
objects were alle as another to go led will a shoot out
for correlation, first find mean, then S.D.
General Abbroad for Combining Similarities.
General Approach for Combining Similarities. Clide 64.
DBSCAN: Clustering based algorithm
Density Based SCAN.
- Euclidean densite
- Probabilit. descit
- Euclidean density - Probability density - Graph - based density.
The survival devising.

Write down the classification of measures in terms of
Write down the dassy
classification and
association algorithm (the a thing of alustoning
Eg. Endidean will be en dansfication, motor
Eg Endidean will be in darsification, clustering or association:
Dorta Exploration
the state of the s
Measures of Location: Mean and Median.
Man i the average value
Mean is the average value mean $(x) = \bar{x} = 1 \sum_{i=1}^{m} x_i$
1 611 11 6 11 1 (chan biz)
Range and Variance
hange = diff between max and min
Variance or S.D (n) = $S_n^2 = \frac{1}{m-1} \left(\frac{\pi}{2} \left(\frac{\pi}{2} - \frac{\pi}{2} \right)^2 \right)$
m-1 121
Visualization Juro parts: Either before starting or after completing Representation is important for visualization
Two parts: Either before starting or after completing
Representation is important for visualization
, voultation
Box Plot: Above partion shows the outliers
partion shows the outliers
think the state of
er - a i Chasterina board algorithms
MASS L. I
theory hand stand algorithms theory of short was and the stand it is board on these encountered.
The state of the s
middless dinition of the second
And the state of t

Massification:			
Jechniques:			
Decision Trees based Methods. - There can be more than I decision tree for same data			
Decision Iru Induction (Induction means training part) - Hunt's Algorithm - CART - 103, c4.5 - SLIB, SPRINT.			
How to specify test condition? - Depends on attribute types: Nominal, Ordinal, Continuous. Eq. Person's eyes. Like height - short, medium, fall. Continuous values			
Continuous values. Depends en no. of ways to split. a-way split or multi-way split.			
How to determine the best split? We need to consider that where error is minimum.			
Measures of node impurity Jini Index, Entropy, Misclassification Error.			
We want homogeneous case, minimum impurity.			

Jain - MO-M12 vs MO-M34.
1 - 2 / 0 / 1 / 2
GINICE
Gini = 0.000, Gini = 0.278, Gini = 0.444, Gini : 0.500
Gini = 0.000 - first choice most interesting
beauxe it is homogeneous.
Gini = 0.500 should be avoided as it is non-homogeneous
Gini index calculation for continuous is tough as compared to nominal and ordinal.
compared to nominal and ordinal.
P. 1 (P. 1 t.)
Error = 1 - max (P; 1t)
n. L. of all the three ward limit is the least
Out of all the three curves, lyini is the best. Gini can be improved by better splitting
money parished lawing
Lie Personis 1472
Hat were been took - the water
Cantinuous values
- Depends on no of ways to offil
Lity pour iduce so lite you be
3
" Light Last will me work to be made !
continued to make the delican of bear off.
Marian of mail may only
girai Index Entropy His classification inos