Course

Unit 1 : Introduction and Ethical theories. 4L

Unit 2 : Ethics in IT societies, Intellectual right and Privacy 6L

Unit 3 : Professional Relationship Responsibilities, Ethics in Computing 6L

Unit 4 : Online crime hacking, legal aspects of professional ethics 4L

Ethical Theories.

Virtue Ethics.

Deontological Ethics : doing actions.

Consequentalism

Philosophy.

Introduction to Data Mining.

Big data : It has three characteristics :
- If data is coming at very high speed (velocity)
- quantity is large (volume).
- not a single type of data (varacity)
  Eg. twitter data, facebook data

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
Selection, normalisation, Preprocessing (like cleaning and feature extraction), transformation, data mining, interpretation / evaluation
transformation : converting data in a particular format so that ML algorithms can be applied

If interpretation is wrong, then we will go back and evaluate at each step : Postprocessing.

PCA, SVD, correlation analysis - preprocessing techniques.

Higher dimension analysis is not possible so dimensi
reduction. Eg. gene data.

Data Mining Tasks.
Prediction Methods (like supervised)
Description Methods (like unsupervised).

Data Mining Tasks.

Classification (Predictive).
Decision trees, Naive Bayes

Clustering (descriptive) dividing into groups but final result can't be said
We will use similarity measures like cosine similarity
KNN, hierarchial clustering.

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)

Sequential Pattern Discovery. (Descriptive)
There is one timeline which is always there
Eg. stock market data, fire alarm ringing, due to election results market value goes high or low.

Regression (Predictive) Whenever data set is continuous
(Linear, Logistic, Quantum).
Used in stock market prediction.

Deviation Detection (Predictive)
Eg. credit card fraud detection.
Normal behaviour is stored and changes are checked.

Training set is also called record data test.

Association Rule Discovery :
Support and confidence : use and define rules

Challenges of Data Mining .

1) Scalability : Data size is increasing day by day.
2) Dimensionality : Data has many dimensions (many attributes) High dimension data needs to be reduced in less dimensions.
3) Complex and heterogeneous data :

4) Data Quality : is degrading due to anamoly and noise
5) Data ownership and distribution : Distribution from where no one knows, big issue. Origin of whatsapp message can't be detected immediately.
6) Privacy preservation : This is a major issue, so no google in China.
7) Streaming data : Data that is coming and going continuously called tunnel.
Eg. Twitter analysis : In 5 mins whatever tweets go through the tunnel can be analysed.

04.09.19

Eucledian distance is used in K means.

imp. Minkowski distance

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_0 - \bar{x}_j)(x_n - \bar{x}_i)$$

Asignment Question :

Calculate the Euclidian, Minkowski and Mahalanobis distance with following parameters :

P1 (2, 4), P2 (4, 2), P3 (5, 5), P4 (4, 2).

Also, draw the relationship between all three.

Jaccard Coefficient (J coefficient) Only for binary

SMC = no. of matches / no. of attributes.

J = no. of 11 matches / no. of not both zero attribute values.

Cosine Similarity.

$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$

Extended Jaccard Coefficient. For continuous or count attributes.

Correlation : measures the linear relationship between objects.

For correlation, first find mean, then S.D.

General Approach for Combining Similarities.

Slide 64.

DBSCAN : Clustering based algorithm

Density Based SCAN.

It is based on three measures :

- Euclidean density
- Probability density
- Graph - based density.

Write down the classification of measures in terms of classification algorithms, clustering algorithms or association algorithm. (Make a table for it)
Eg. Euclidean will be in classification, clustering or association.

Data Exploration.

Measures of Location : Mean and Median.
Mean is the average value

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Range and Variance
Range = diff between max and min.

$$\text{Variance or S.D }(x) = S_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$$

Visualization.
Two parts : Either before starting or after completing
Representation is important for visualization.

Box Plot : Above portion shows the outliers

Classification :

Techniques :

Decision Trees based Methods.
- There can be more than 1 decision tree for same data

Decision Tree Induction : (Induction means training part)
- Hunt's Algorithm
- CART
- ID3 , C4.5
- SLIQ , SPRINT.

How to specify test condition?
- Depends on attribute types:
Nominal , Ordinal , Continuous.
Eg. Person's eyes.
Like height - short, medium, tall.
Continuous values.
- Depends on no. of ways to split.
2-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
Gini Index, Entropy , Misclassification Error.

We want homogeneous case, minimum impurity.

Gain = M0 – M12 vs M0 – M34.

$GINI(t) = 1 - \sum_j [p(j|t)]^2$

Gini = 0.000, Gini = 0.278, Gini = 0.444, Gini : 0.500

Gini = 0.000 → first choice most interesting because it is homogeneous

Gini = 0.500 should be avoided as it is non homogeneo

Gini index calculation for continuous is tough as compared to nominal and ordinal.

Error = $1 - \max(P_i|t)$.

Out of all the three curves, Gini is the best.
Gini can be improved by better splitting.