TODAY - SESSION - 8 – AGENDA – JUNE 13 2020

1. POSITIVE POINT / STORY-1
2. A QUICK RECAP OF WHAT'S BEING COVERED……exam is approaching
3. FOCUS POINTS/PROBLEMS FOR MID-TERM EXAMINATION - SOME SAMPLE PROBLEMS..../QUESTIONS PATTERN - 1215 pm… (1115) – 5 mts for Q
4. **ASHUTOSH** –

ONE APRIORI ALGORITHM – SOLVED PROBLEM (midterm)

MEASURES: SUPPORT, CONFIDENCE, LIFT AND INTEREST - UPTO SLIDE 16

5. SUBJECTIVE VS OBJECTIVE MEASURES
6. MAJOR TASKS IN DATA MINING
7. POSITIVE POINT / STORY-2**
8. Q & A SESSION
9. FP GROWTH ALGORITHM - SOLVED PROBLEM & TIME PERMITS SHOW HOW TO CALCULATE GINI INDEX – **if time permits – OTHERWISE WE WILL DISCUSS IN THE NEXT CLASS**

MID-TERM EXAM COVERAGE OF PORTIONS – MARKS DISTRIBUTION

21st June - 2:00 PM to 3:30 PM

Written Assignment type but with time limit of **1 hour 30 minutes**. Question paper would be available on Canvas at 2:00 pm on the scheduled date, which the students can download and solve it offline on A4 sheets, and then upload a scanned copy of their solution on canvas within 1 hour 30 minutes.

Since the exam is online, it's more like an open book exam where the students have access to online resources, books, and whatsapp to discuss the question.

Syllabus/portions for the mid-term exam (30 MARKS)

The syllabus is the first 4 modules till Classification module (including decision tree, rule-based classifiers, evaluation measures)**, No topic from Association rule mining, apriori algorithm, FP growth, etc – comprehensive exam…..**

Question on Data Cleaning

(eg, for the given dataset, to identify problems associated with the data and how to clean it) - **5 marks (Q)** **..CLEARLY USE THE TERMS……. ANSWER KEYS..**

QUALITATIVE… NOMINAL….. .COLUMN A HAS INTERVAL TYPE.. COLUMN B HAS BINARY TYPES… use the notations./DM key words /terms as much as possible..

Any topic from data cleaning module - DATA QUALITY .. .DATA PREPROCESSING – googling…

(eg. Binning, Normalisation, **z-score**, Pearson coefficient, covariance, etc,) - **4+4 marks (Q2, Q3)**

Questions from Data exploration

(eg, Box plot, Similarity/dissimilarity on mixed attributes types) - **5 marks (Q4)**

Draw neat.. use pencil and scale…….mostly same… there may be some numbers/data attributes… would change.. binning… /binning… …

Question from Classification

will have 2 parts

Decision tree & Rule based Classifiers - **5 + 3 marks (Q5, Q6)**

Table… DT… rules…. Strong rules… ? support coverage…. Order the rules…!?

Evaluation Measures- **4 marks (Q7) TWO METHODS precision, recall, accuracy….**

There will be 5 questions (some questions will have 2 parts) **7 questions in the QP * 30 marks**

SESSION-1

1. Data explosion problem;
2. Brief overview of Data Warehouse
3. Datamart - diff between DWH and Datamart; Brief overview of Datamarts
4. OLTP Vs OLAP
5. Brief overview of DWH and ETL architecture
6. What is Data mining?. Why Data mining?
7. Functionalities of Data mining! Data Mining Process
8. Issues/Challenges with Data Mining
9. Applications of Data Mining

some demo of python codes

SESSION-2

1. Introduction to Data Preprocessing - Why Preprocessing ? DATA IS WITH HIGHLY QUALITY
2. Major Tasks in Data Preprocessing

Data cleaning, Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies; Data integration, Integration of multiple databases, data cubes, or files

Data reduction, Dimensionality reduction, Numerosity reduction, Data compression

Data transformation and data discretization, Normalization, Concept hierarchy generation

3. Data Quality - handling missing data and noisy data/errors **QUESTION… MIDTERM
4. Outliers – PROBLEMS – 5 point summary ; box plot analysis
5. Handling Redundancy/duplicates in the data
6. Correlation Analysis

Pearson's product moment coefficient Coefficient, Chi-Square, Co-Variance – PROBLEMS

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$Cov(A,B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

7. Data Discretization divide the range of continuous attributes into intervals - BINNING PROBLEMS

Equal-width (distance) partitioning - 2 to 3 marks

Equal-depth (frequency) partitioning

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

8. Data Reduction – Strategies/ methods – dimensionality, Numerosity reduction, **Principal Component Analysis(PCA) – CONFIRMATORY FACTOR ANALYSIS 20 VARIABLES –COMPONENTS, EIGEN VALUE, DISCRETE WAVELET TRANSFORMATION(DWT) – frequency/distribution… of the data**

9. DATA REDUCTION – NUMEROSITY REDUCTION

Reduce data volume by choosing alternative, *smaller forms* of data representation

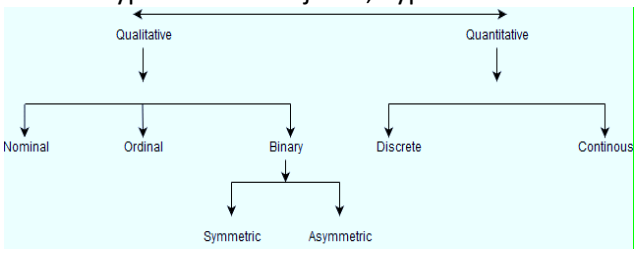
Parametric Methods - regression

Non-Parametric Methods – Min-Max Normalization and Z-Score Normalization – PROBLEMS

(0 to 1)… hard … MINI 10 MAX 20

SESSION-3 DATA EXPLORATION - MORE ANALYTICAL PROBLEMS……

1. Types of Data Sets
2. Types of Data Objects ; Types of Data Attributes



NOMIAL, BINARY, NUMERIC – INTERVAL, RATIO

Discrete Vs Continuous Attributes

3. Basic Statistical Data Descriptions - PROBLEMS

MEAN, MEDIAN, MODE

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

Value that occurs most frequently in the data(MODE)

Symmetric Vs Skewed Data

4. Measuring the dispersion of data - PROBLEMS

BOX PLOT ANALYSIS

Quartiles (Q1 25th percentile, Q3 – 75TH percentile, IQR = Q3 – Q1

FIVE-NUMBER SUMMARY

min, Q1, median, Q3, max

Outlier - a value higher/lower than 1.5 x IQR

VARIANCE $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

Standard deviation s (or σ) is the square root of variance s2 (or σ2)

PARTITIONING USING 3-4-5 RULE

5. Data Standardisation / Normalization – PROBLEMS

MIN-MAX NORMALIZATION
Normalized value => VALUE – MINIMUM / MAXIMUM – MINIUM

$$U_i = \frac{\mathbf{V}_i - X1}{X2 - X1} \ (Y2 - Y1) + Y1$$

Z-score normalization of the given value is given value – μ / σ
μ is the mean value of the feature and σ is the standard deviation of the feature

6. DATA SIMILARITY AND DISSIMILARITY
   DISTANCE MEASURES ----------PROBLEMS
   EUCLIDIAN DISTANCE
   MANHATTAN DISTANCE

Euclidean: Take the square root of the sum of the squares of the differences of the coordinates.

For example, if $x = (a, b)$ and $y = (c, d)$, the Euclidean distance between x and y is

$\sqrt{(a - c)^2 + (b - d)^2}$.

Manhattan: Take the sum of the absolute values of the differences of the coordinates.

For example, if $x = (a, b)$ and $y = (c, d)$, the Manhattan distance between x and y is

$|a - c| + |b - d|$.

MINKOWSKI distance

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

COSINE SIMIARLITY

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \ ||d_2||$,
If d_1 and d_2 are two vectors (e.g., term-frequency vectors), $\bullet$ indicates vector dot product, $||d||$: the length of vector d
Find the **similarity** between documents 1 and 2.
d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
$d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25
$||d_1||$ = $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5}$ = 6.481
$||d_2||$ = $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5}$ = 4.12
$\cos(d_1, d_2)$ = 0.94
50 to 60% of the MIDTERM QUESTIONS ARE FROM SESSION2 AND SESSION3 **** -- folder 8…
SESSION-4

1. CLASSIFICATION AND PREDICTION
2. SUPERVISED VS UNSUPERVISED LEARNING
3. DECISION TREE .. MODEL CONSTRUCTION
4. CLASSIFICATION TASKS
5. LAZY VS EAGER LEARNING

6. DECISION TREE INDUCTION
7. OVERFITTING AND UNDERFITTING – tree pruning *Ashutosh
8. HUNT's algorithm
9. **Decision Tree – Splitting Attributes (Multi-way split, binary split ; Ordinal and continuous attributes) … Decision tree… what is the best node… root node… what is your branch node..**

SESSION-5

1. MEASURES OF NODE IMPURITY

GINI INDEX ……sample problem…. Of calculating index

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

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

MISCLASSIFICATION ERROR

$$Error(t) = 1 - \max_i P(i|t)$$

- Measures misclassification error made by a node.
  - Maximum (1 - 1/n_c) when records are equally distributed among all classes, implying least interesting information
  - Minimum (0.0) when all records belong to one class, implying most interesting information

A	B	C	Number of Instances	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

Answer: The error rate for the data without partitioning on any attribute is

$$E_{orig} = 1 - \max(\frac{50}{100}, \frac{50}{100}) = \frac{50}{100}.$$

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- This attribute minimizes the expected number of tests needed to classify a given tuple.
- Let p_i be the probability that a tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$, m is the number of distinct classes, v is the number of distinct values in an attribute.
- Expected information (entropy) needed to classify a tuple in D:

$$Info\ (D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info\ (D_j)$$

  - The smaller the expected information required, greater the purity of the partitions.

SESSION-5
MORE COVERAGE OF DECISION TREE – ADDITIONAL PROBLEMS

GINI INDEX

The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. **It favors larger partitions**.

Information Gain multiplies the probability of the class times the log (base=2) of that class probability. Information Gain favors smaller partitions with many distinct values. Ultimately, you have to experiment with your data and the splitting criterion.

Algo / Split Criterion	Description	Tree Type
Gini Split / Gini Index	Favors larger partitions. Very simple to implement.	CART
Information Gain / Entropy	Favors partitions that have small counts but many distinct values.	ID3 / C4.5

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Favors larger partitions.
- Uses squared proportion of classes.
- Perfectly classified, Gini Index would be zero.
- Evenly distributed would be 1 – (1/# Classes).
- You want a variable split that has a low Gini Index.
- The algorithm works as 1 – (P(class1)^2 + P(class2)^2 + … + P(classN)^2)

9

ENTROPY & INFORMATION GAIN

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

- Favors splits with small counts but many unique values.
- Weights probability of class by log(base=2) of the class probability
- **A smaller value of Entropy is better.**
- That makes the difference between the parent node's entropy larger.
- Information Gain is the Entropy of the parent node minus the entropy of the child nodes.
- Entropy is calculated [P(class1)*log(P(class1),2) + P(class2)*log(P(class2),2) + … + P(classN)*log(P(classN),2)]

When you use Information Gain, which uses Entropy as the base calculation, you have a wider range of results. The Gini Index caps at one.
The maximum value for Entropy depends on the number of classes. It's based on base-2, so if you have…

- Two classes: Max entropy is 1.
- Four Classes: Max entropy is 2.
- Eight Classes: Max entropy is 3.
- 16 classes: Max entropy is 4.

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (**entropy**) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

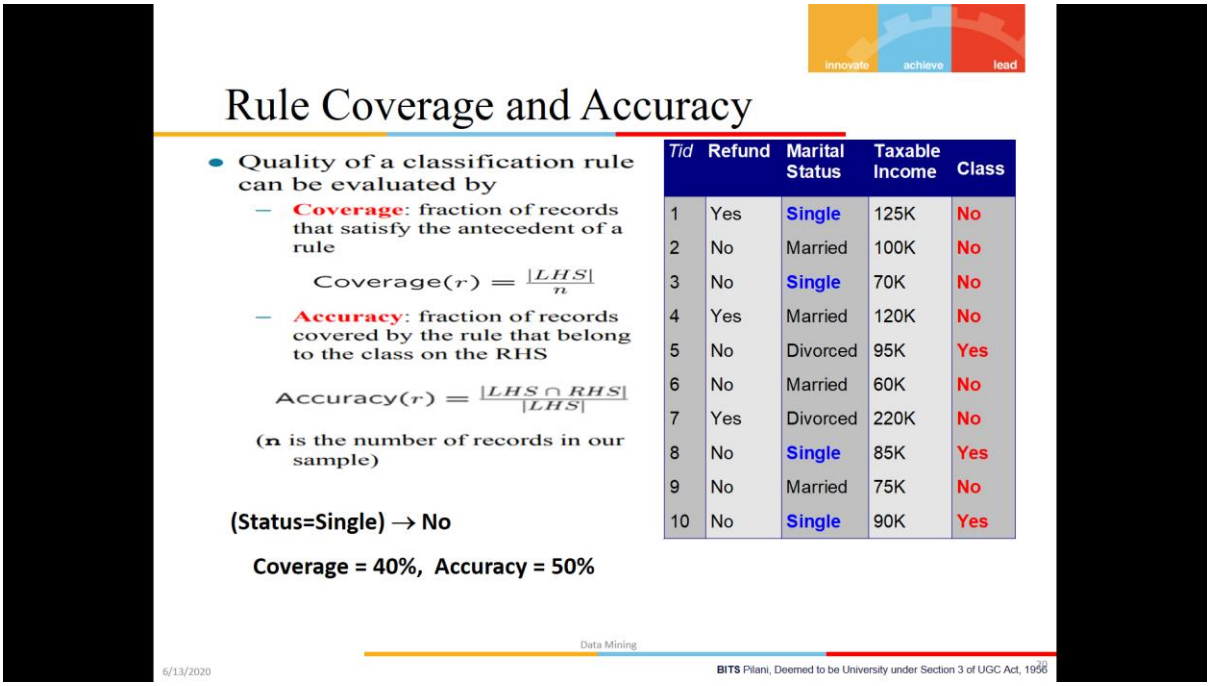
- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

RULE BASED CLASSIFICATION



ORDERED RULES
RULE BASED VS CLASS BASED ORDERING
EVALUATION OF RULES
FOIL GAIN
FOIL PRUNE
LIKELIHOOD RATIO ; BUILDING CLASSIFICATION RULES – DIRECT VS INDIRECT methods; EXAM.. .DATA SET…. DECISION TREE…. ASK YOU TO WRITE THE CLASSIFICATION RULES – slide.. examples..
SESSION-6

Regression
Types of regression Linear Vs non-linear regression

Model evaluation
Confusion Matrix

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

innovate achieve lead

Classifier Accuracy, or recognition rate: percentage of test set tuples that are correctly classified

Accuracy = (TP + TN)/All

Error rate: *1 – accuracy*, or

Error rate = (FP + FN)/All

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

- **Class Imbalance Problem**:
  - One class may be *rare*, e.g. fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class
  - **Sensitivity**: True Positive recognition rate
    - **Sensitivity = TP/P**
  - **Specificity**: True Negative recognition rate
    - **Specificity = TN/N**

Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

Recall: completeness – what % of positive tuples did the classifier label as positive?

Perfect score is 1.0

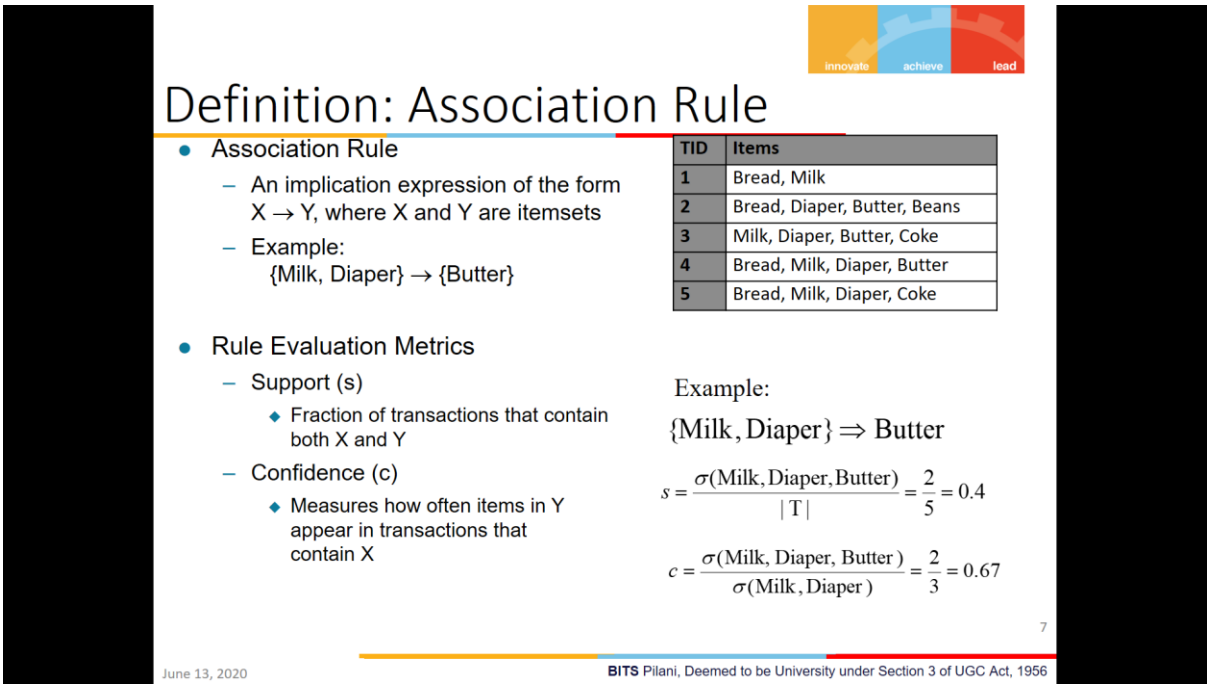
Inverse relationship between precision & recall

$$recall = \frac{TP}{TP + FN}$$

***F* measure (*F_1* or *F*-score)**: harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

SESSION-7 : ASSOCIATION ANALYSIS
ASSOCIATION RULE MINING



Frequent Itemsets
Apriori Algorithm….. Ashutosh…
FP GROWTH ALGORITHM… **Venkat**
Apriori algorithm code

SESSION-8

ASSOCIATION RULES

STATISTICAL MEASURES**

RECAP