2018 Scheme

1	N	T							

NITTE EDUCATION TRUST | NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(An Autonomous Institution | Accredited with 'A+' Grade by NAAC | NBA | Approved By UGC | AICTE | Govt. of Karnataka | Affiliated to VTU, Belagavi)
Yelahanka, Bangalore – 560 064

KNOWLEDGE • CHARACTER • UNITY

Department of Computer Science and Engineering

Course Title with code	Data Mining, 18CS54	**Maximum Marks**	**30 Marks**
Date and Time	19/11/2021, 9.30am to 10.30am	**No. of Hours**	1.0
Course Instructor(s)	Dr. Vijaya Shetty S, Dr. Sujata Joshi, Dr. Vani V		
SCHEME & SOLUTION			

Q. No	Question	MAX MARKS
1.a	Solution The following are **examples of possible answers**. • 1. Clustering can be used to generate movie thumbnails based on the watch pattern of an individual users. 2. Clustering can be used to group user-user movie ratings • Classification can assign results to pre-defined categories such as genre, language, cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, • Association rule mining can recommend best (high probability) movies based on the watch list of an individual user (search and recommend the content that interests an individual)	2x 3 data mining tasks = 6 marks
1.b	Solution (1) Order the Universities based on the student enrollment rate. – **Not a data mining task.** This is a **database query** where using order by clause , universities can be ordered based the student enrollment rate. (2) Grouping of similar traffic frames from a video. **Yes,** This is a data mining task. Precisely, **unsupervised learning – clustering** is the data mining task that would be exploited to group similar traffic frames (patterns) from a video (3) Determine the maximum salary of academic staff in every department at NMIT. **Not a data mining task.** This is a simple **database query** where group by clause is used with an aggregate funciton max(salary) (4) Detect the key frames from a rugby sport video footage. **Yes,** This is a data mining task. In rugby sport video footage key frames (goal, foul, shoot etc) can be detected with the help of **unsupervised learning - clustering** . One of the simple approaches is to detect the key frames based on the color variants by eliminating off the field views.	1.5 x 4 activities = 6 marks

1.c	Any 3 Motivating Challenges **Scalability:** Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive data sets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms. **High Dimensionality**: It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high dimensional data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases. **Heterogeneous and Complex Data:** Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyperlinks; DNA data with sequential and three-dimensional structure; and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on the Earth's surface. Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structured text and XML documents **Data Ownership and Distribution:** Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. Among the key challenges faced by distributed data mining algorithms include (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security issues. **Non-traditional Analysis:** The traditional statistical approach is based on a hypothesize-and-test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labour intensive. Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically not the result of a carefully designed experiment and often represent opportunistic samples of the data, rather than random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.	For Each challenge: 1 mark 3 x 1 = 3 marks

2. a	***Solution*** A **quantitative variable** can be either continuous or discrete. A continuous variable is one that in theory could take any value in an interval. Examples of continuous variables are body mass, height, blood pressure and cholesterol. A discrete quantitative variable is one that can only take specific numeric values (rather than any value in an interval), but those numeric values have a clear quantitative interpretation. Examples of discrete quantitative variables are number of needle punctures, number of pregnancies and number of hospitalizations. **Qualitative variables** Qualitative or categorical variables describe a quality or attribute of the individual. Categorical data can be either nominal or ordinal. Sex is an example of a nominal variable, and histologic stage is an example of an ordinal variable. Ordinal variables have a specific order; for the other variable, they do not. If one patient has histologic stage 4 and another patient has histologic stage 1, you know that the stage 4 patient has more severe disease. Although the histologic stages are categories, the categories have an inherent order. The same cannot be said for the variable sex. Qualitative data with unordered categories is referred to as nominal; qualitative data with ordered categories is referred to as ordinal. i) Military rank -Discrete, qualitative, ordinal ii) Number of patients in a hospital- Discrete, quantitative, ratio iii) Temperature in Fahrenheit-Continuous, quantitative, Interval iv) Eye color- Discrete, qualitative, Nominal	1.5m 1.5m 0.5×4=2m
2. b	**Solution** Given attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70 (a) min_A= 13, max_A=70,new-min_A=0.0, new-max_A=1.0,v_i=45 $v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$ $=\frac{(45-13)(1-0)}{(70-13)} + 0$ =0.56 (b) z-score normalization to transform the value 45 for age, where the standard deviation of age is 12.94 years. A value, v_i , of A is normalized to v_i' by computing $v_i' = \frac{v_i - \bar{A}}{\sigma_A},$ where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. $\bar{A}$ = 1/ n (v1 + v2 + ⋯ + vn) =809/27=29.96 $=\frac{(45-29.96)}{12.94}$ =1.16 (c) Use normalization by decimal scaling to transform the value 45 for age.	2m 1.5m 1.5m

	A value, vi , of A is normalized to v_i' by computing $v_i' = \frac{v_i}{10^j}$, where j is the smallest integer such that $max(\lvert v_i' \rvert) < 1$. $=\frac{45}{100}$ =0.45	
2. c	Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed. Data miners sample because it is too expensive or time consuming to process all the data. In some cases, using a sampling algorithm can reduce the data size to the point where a better, but more expensive algorithm can be used. The key principle for effective sampling is the following: Using a sample will work almost as well as using the entire data set if the sample is representative. In turn, a sample is **representative** if it has approximately the same property (of interest) as the original set of data. If the mean (average) of the data objects is the property of interest, then a sample is representative if it has a mean that is close to that of the original data. Because sampling is a statistical process, the representativeness of any particular sample will vary, and the best that we can do is **choose a sampling scheme that guarantees a high probability of getting a representative sample.** **Sampling Approaches** 1. R**andom sampling**. For this type of sampling, there is an equal probability of selecting any particular item. There are two variations on random sampling: (a) **sampling without replacement**-as each item is selected, it is removed from the set of all objects that together constitute the population. (b) **sampling with replacement**-objects are not removed from the population as they are selected for the sample. In sampling with replacement, the same object can be picked more than once. The samples produced by the two methods are not much different when samples are relatively small compared to the data set size, but **sampling with replacement is simpler to analyze since the probability of selecting any object remains constant during the sampling process.** When the population consists of different types of objects, with widely different numbers of objects, **simple random sampling can fail** to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all object types. For example, when building **classification models for rare classes**, it is critical that the rare classes be adequately represented in the sample. Hence, a sampling scheme that can accommodate differing frequencies for the items of interest is needed. **2. Stratified sampling**, which starts with prespecified groups of objects, is such an approach. In the simplest version, equal numbers of objects are drawn from each group even though the groups are of different sizes. In another variation, the number of objects drawn from each group is proportional to the size of that group. **3. Progressive Sampling** The proper sample size can be difficult to determine, so **adaptive or progressive** sampling schemes are sometimes used. These approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained. While this technique eliminates the need to determine the correct sample size initially, it requires that there be a way to evaluate the sample to judge if it is large enough.	2M 3m
3. a	Various methods for handling the problem of missing values (any 5 can be written) **1. Ignore the tuple**: This is usually done when the class label is missing. This method is not very effective, unless the tuple contains several	1 x 5 = 5m

	attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably. By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand. **2. Fill in the missing value manually**: In general, this approach is time consuming and may not be feasible given a large data set with many missing values. **3. Use a global constant to fill in the missing value**: Replace all missing attribute values by the same constant such as a label like *"Unknown"* or -999. If missing values are replaced by, say, *"Unknown*," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of *"Unknown."* Hence, although this method is simple, it is not foolproof. **4. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value**: For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median. For example, suppose that the data distribution regarding the income of *AllElectronics* customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for *income*. **5. Use the attribute mean or median for all samples belonging to the same class as the given tuple**: For example, if classifying customers according to *credit risk*, we may replace the missing value with the mean *income* value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice. **6. Use the most probable value to fill in the missing value**: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for *income*	
3. b	Compute the following for the vectors X=(1,1,1,0,1,0) and Y=(1,0,1,0,1,0) i. Jaccard similarity ii. Cosine similarity Euclidean distance f00=2, f01 = 0, f10=1 f11=3 Jaccard similarity $$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}.$$ J=3/(0+1+3)= 0.75 Cosine similarity $$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|},$$	2m 2m 1m

	where · indicates the vector dot product, $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$, and $\|\mathbf{x}\|$ is the length of vector $\mathbf{x}$, $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$. X.Y= 1 x1 + 1 x 0 + 1 x1+ 0 x0+ 1 x1+ 0 x 0 =3 ||X||= 2 ||Y||=1.732 Cos(X,Y)= 3/(2 x 1.732) = 0.866 Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2},$ D(x,y) = $\sqrt{(1-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2} = 1$	
3. c	(a) This may conflict with one's privacy to a certain extent. Suppose we have installed a new security system in our house but do not want this to be known by others. If the bank were to contact us regarding offers of special loans for the purchase of additional security systems, this could be seen as an infringement on personal privacy. Another example is if we had bought a safe to store your own collection of treasures. This infringement from the bank may invoke danger if this information falls into the hands of potential thieves. (b)Another situation where we feel that data mining can infringe on our privacy is Suppose we have a Supermarket Club Card. Having access to our card, Supermarket has the potential, without our knowledge, to study our shopping patterns based on our transactions made with the Club Card. The drugs for medical use purchased durng specific periods of time is personal information that we may not wish to be used or revealed. (c) Since the primary task in data mining is to develop models about aggregated data, we can develop accurate models without accessing precise information in individual data records. For example, when building a decision-tree classifier, we can use training data in which the values of individual records have been masked. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. The original distribution of a collection of distorted data values can be approximated using a reconstruction algorithm. By using this it is possible to estimate the distribution of original data values. We can then build a classifier using the approximated values, thereby protecting the privacy of individuals. By using these reconstructed distributions, the resulting classifier will be the same as or very similar to that built with the original data.	2m 1m 2m

Faculty Signature	Course Coordinator/Mentor Signature	HoD Signature Dr. Thippeswamy M N