

**CS 5990 (Advanced Data Mining) - Assignment #1**  
**Maximum Points: 100 pts.**

**Bronco ID: 017314921**

**Last Name: Panda**

**First Name: Subham**

**Answers:**

**1.**

- a) Dividing the customers of a company according to their gender:** Not a data mining task. Because there is no pre-processing implementation done or pattern analysis done.
- b) Monitoring seismic waves for earthquake activities:** Yes, this can be considered a data mining task if it involves pattern recognition and predictive analysis in seismic data to forecast earthquakes.
- c) Computing the total sales of a company:** Not a data mining task. It's a basic summation/aggregation operation without the need for in-depth data analysis or pattern discovery.
- d) Predicting the outcomes of tossing a (fair) pair of dice:** Not a data mining task. It's based on probability theory and does not involve analyzing large datasets to extract patterns or insights.
- e) Predicting the future stock price of a company using historical records:** Yes, this is a data mining task. It involves complex analysis and the use of predictive models to forecast future values based on historical data.
- f) Monitoring the heart rate of a patient for abnormalities:** Yes, this is a data mining task if it involves analyzing data for anomaly detection or pattern recognition to identify health issues.

2.

- a) **Brightness as measured by a light meter:** Continuous, Ratio
- b) **Brightness as measured by people's judgments:** Discrete, Ordinal
- c) **Density of a substance in grams per cubic meter:** Continuous, Ratio
- d) **Time of each day in the meaning of a 12-hour clock:** Discrete, Interval
- e) **CPP bronco IDs:** Discrete, Nominal

3.



### Data Preprocessing:

- **Visualization:** Exploratory data analysis (EDA) at this point requires visualization. In order to better guide subsequent preprocessing operations like cleaning and transformation, visualization tools can offer instantaneous visual feedback regarding the data.
- **Dimensionality Reduction:** This is another preprocessing approach that's used, especially when working with high-dimensional data. Reduce the number of features by using techniques like Principal Component Analysis (PCA) or feature selection methods. This will improve computing performance and simplify the models that need to be developed during the data mining phase.
- **Machine Learning Techniques:** A few machine learning algorithms can be used to preprocess data in order to find its structure. For example, they can be used to estimate missing values or find clusters of related data points (an unsupervised learning technique).

### Data Mining:

- **Machine Learning Techniques:** This is the core phase where machine learning algorithms are applied to extract patterns and build predictive models. Techniques range from clustering, classification, regression, association rules to more complex methods like deep learning. The choice of algorithm depends on the problem at hand and the nature of the data.

### Postprocessing:

- **Visualization:** After mining the data, visualization is used again to represent the results in an understandable manner. It helps in interpreting the patterns or models discovered by the data mining algorithms and in evaluating their quality and validity.
- **Dimensionality Reduction:** In some cases, dimensionality reduction might be used post hoc to visualize high-dimensional results in a 2D or 3D space, helping in the interpretation and understanding of the outcomes.
- **Machine Learning Techniques:** Postprocessing might involve applying additional machine learning techniques to refine the models further. For example, ensemble methods might be used to combine the results of multiple models, or decision threshold moving might be applied in classification tasks to adjust the trade-off between precision and recall.

4.

- a) **Classification** - (a) for categorizing queries and content.
- b) **Clustering** - (b) for grouping related queries or content, which improves the search engine's relevance and suggestions.
- c) **Association Rule Mining** - (c) for uncovering correlations between user actions or queries to enhance recommendations and targeted content.
- d) **Anomaly Detection** - (d) now understood as identifying major themes and the prominence of certain topics, which could help in trend detection and in prioritizing certain types of content or ads.

By integrating these data mining techniques, consulting company can craft a search engine that not only responds to user queries with accuracy but also anticipates user needs, offers personalized content, and keeps users informed about the latest trends. This story of consulting company is a simplified illustration of how modern web search engines continuously refine their data mining processes to improve the user experience in an ever-evolving digital world.

5.

- a) **Most likely task:** The data scientists are most likely trying to accomplish a classification task of scam or fraud transaction. The goal appears to be to predict whether an individual is likely to "Cheat" on their taxes based on attributes like "Refund," "Marital Status," and "Taxable Income."
- b) **What is a feature:** A feature is an attribute which is a property or characteristic of an object. In the context of this data, a feature would be one of the columns that might be used to predict the target column, such as "Refund," "Marital Status," or "Taxable Income."
- c) **What is a feature value:** A feature value is the actual data point for a particular feature/attribute. For instance, looking at the "Marital Status" feature, a feature value could be "Single," "Married," or "Divorced," depending on the individual's marital status.
- d) **What is dimensionality:** Dimensionality refers to the number of features that are present in the dataset. In this dataset, the dimensionality is three, corresponding to the three features: "Refund," "Marital Status," and "Taxable Income."
- e) **What is an instance:** An instance is an individual entry or row or entire record in the dataset, representing all the feature values and the class label for a particular example. For instance, the first row is an instance where the individual has a "Refund" of "Yes," a "Marital Status" of "Single," a "Taxable Income" of "125K," and a "Cheat" class of "No."
- f) **What is a class:** A class is the outcome or label from the target column that the model is trying to predict. In this dataset, the class is represented by the "Cheat" column, which has two class labels: "Yes" or "No," indicating whether the individual is likely to cheat or not.

6.

The statistician's conclusion that fields 2 and 3 are identical holds true, as it appears that scaling has been applied to the data in column 3.

Given the values presented:

- *Field 2*: 233.8, 119.7, 168.0
- *Field 3*: 33.4, 17.1, 24.0

We can check for direct proportionality by dividing the numbers in field 3 by the corresponding numbers in field 2:

- $33.4 / 233.8 \approx 0.143$
- $17.1 / 119.7 \approx 0.143$
- $24.0 / 168.0 \approx 0.143$

Each of these divisions yields approximately the same quotient, suggesting that field 3 is indeed a scaled version of field 2. Therefore, the statistician's conclusion that fields 2 and 3 are basically the same is likely based on this observation of direct proportionality.

7.

Let's formulate the similarities:

**Jaccard Similarity:**

$$J(X, Y) = \frac{a_{11}}{a_{11} + b_{10} + c_{01}}$$

$a_{11}$  = number of labels for X & Y both being 1

$b_{10}$  = number of labels for X & Y both being 1

$c_{01}$  = number of labels for X & Y both being 1

**Cosine Similarity:**

$$\text{Cos}(X, Y) = \frac{X \cdot Y}{\|X\| * \|Y\|}$$

$X \cdot Y$  = dot product between X & Y

$\|X\| * \|Y\|$  = Corresponding magnitudes of vector X & Y

**Euclidean Distance:**

$$E(X, Y) = \sqrt{\sum (X_i - y_i)^2}$$

**Correlation:**

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{s_x * s_y} = \frac{s_{xy}}{s_x * s_y}$$

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

**a)**

$$\mathbf{X} = (1 \ 1 \ 0 \ 0 \ 0)$$

$$\mathbf{Y} = (0 \ 0 \ 0 \ 1 \ 1)$$

$$J(\mathbf{X}, \mathbf{Y}) = \frac{a_{11}}{a_{11} + b_{10} + c_{01}}$$

$$\Rightarrow \frac{0}{0 + 2 + 2} = \mathbf{0}$$

$$\text{Cos}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| * \|\mathbf{Y}\|}$$

$$\Rightarrow \frac{(1.0 + 1.0 + 0.0 + 0.1 + 0.1)}{\sqrt{(1^2 + 1^2 + 0^2 + 0^2 + 0^2)} * \sqrt{(0^2 + 0^2 + 0^2 + 1^2 + 1^2)}} = \mathbf{0}$$

$$E(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum (X_i - y_i)^2}$$

$$\Rightarrow \sqrt{((1 - 0)^2 + (1 - 1)^2 + (0 - 1)^2 + (0 - 1)^2 + (0 - 1)^2)}$$

$$\Rightarrow \sqrt{4} = \mathbf{2}$$

$$\text{Corr}(\mathbf{X}, \mathbf{Y}) = \frac{s_{xy}}{s_x + s_y}$$

$$\bar{x} = 0.4$$

$$\bar{y} = 0.4$$

$$s_{xy} = \frac{1}{4}[(1 - 0.4)(0 - 0.4) + (1 - 0.4)(0 - 0.4) + (0 - 0.4)(0 - 0.4) \\ + (0 - 0.4)(1 - 0.4) + (0 - 0.4)(1 - 0.4)]$$

$$\Rightarrow \frac{1}{4}[-0.24 - 0.24 + 0.16 - 0.24 - 0.24] = -0.2$$

$$s_x = \sqrt{\frac{(1 - 0.4)^2 + (1 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2}{4}}$$

$$\Rightarrow 0.547 = s_y \text{ as well}$$

$$\text{Therefore, Corr}(\mathbf{X}, \mathbf{Y}) = \frac{-0.2}{(0.547)^2} = \mathbf{-0.66}$$



b)

$$\mathbf{X} = (0 \ 1 \ 0 \ 1 \ 1)$$

$$\mathbf{Y} = (1 \ 0 \ 1 \ 0 \ 0)$$

$$J(\mathbf{X}, \mathbf{Y}) = \frac{a_{11}}{a_{11} + b_{10} + c_{01}}$$

$$\Rightarrow \frac{0}{0 + 3 + 2} = 0$$

$$\cos(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| * \|\mathbf{Y}\|}$$

$$\Rightarrow \frac{(0.1 + 1.0 + 0.1 + 1.0 + 1.0)}{\sqrt{(0^2 + 1^2 + 0^2 + 1^2 + 1^2)} * \sqrt{(1^2 + 0^2 + 1^2 + 0^2 + 0^2)}} = 0$$

$$E(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum (X_i - y_i)^2}$$

$$\Rightarrow \sqrt{((0 - 1)^2 + (1 - 0)^2 + (0 - 1)^2 + (1 - 0)^2 + (1 - 0)^2)}$$

$$\Rightarrow \sqrt{5} = 2.23$$

$$\text{Corr}(\mathbf{X}, \mathbf{Y}) = \frac{s_{xy}}{s_x + s_y}$$

$$\bar{x} = 0.6$$

$$\bar{y} = 0.4$$

$$s_{xy} = \frac{1}{4} [(0 - 0.6)(1 - 0.4) + (1 - 0.6)(0 - 0.4) + (0 - 0.6)(1 - 0.4) + (1 - 0.6)(0 - 0.4) + (1 - 0.6)(0 - 0.4)]$$

$$\Rightarrow \frac{1}{4} [-0.36 - 0.16 - 0.36 - 0.16 - 0.16] = -0.3$$

$$s_x = \sqrt{\frac{(0 - 0.6)^2 + (1 - 0.6)^2 + (0 - 0.6)^2 + (1 - 0.6)^2 + (1 - 0.6)^2}{4}} = 0.547$$

$$s_y = \sqrt{\frac{(1 - 0.4)^2 + (0 - 0.4)^2 + (1 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2}{4}} = 0.547$$

$$\text{Therefore, } \text{Corr}(\mathbf{X}, \mathbf{Y}) = \frac{-0.3}{(0.547)^2} = -1.003 \approx -1$$

**8.**

$$u = (2, k)$$

$$v = (3, -2)$$

**a)**

$$u \cdot v = 0$$

$$6 - 2k = 0$$

$$\mathbf{k = 3}$$

**b)**

$$(2, k) = x(3, -2)$$

$$2 = 3x$$

$$x = \frac{2}{3}$$

$$k = -2x$$

$$k = -2 * \left(\frac{2}{3}\right) = \frac{-4}{3}$$

9.

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	\$10
2	Shoes	0	09/10/22	\$15
3	TV	1	09/09/22	\$20

a)

ID	Item	Downtown	Date	Price
1	Electronics	2	09/09/22	\$30
2	Shoes	0	09/10/22	\$15

b)

Item	Date	Price
Watch	09/09/22	\$10
Shoes	09/10/22	\$15
TV	09/09/22	\$20

c)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	0.2
2	Shoes	0	09/10/22	0.5
3	TV	1	09/09/22	0.04

d)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	\$10
2	Shoes	0	09/10/22	\$15

I have systematically addressed the dataset mentioned above, outlining the corresponding techniques used, as detailed in the bullet points:

**a) Aggregation:**

- The "Item" category "Watch" and "TV" in the raw dataset has been aggregated into "Electronics" in the processed dataset. Additionally, the "Price" for ID 1 has increased from \$10 to \$30, which suggests that the prices of the "Watch" and "TV" items might have been combined.
- The "Downtown" feature has also changed from 1 to 2, which could indicate an aggregation of some sort as well.

**b) Feature Selection:**

- In the processed dataset, the "ID" and "Downtown" columns have been removed. This indicates that feature selection has been performed, where only a subset of the original features ("Item," "Date," and "Price") is retained for further analysis.

**c) Normalization:**

- The "Price" column values have been changed from dollar amounts to what appears to be normalized values between 0 and 1. This is a typical normalization process that often involves dividing by the maximum value, min-max scaling, or some other method to standardize the range of feature values.

**d) Sampling:**

- If we had to consider that the raw dataset had more entries and only these were selected, it could be considered sampling.

**10.**

**[Click to get directed to GitHub repo with code !](#)**