AMA546 Statistical Data Mining
Midterm test

1 For the question below, please judge whether it is true or false (T/F) and explain.

1. Attribute values are numbers or symbols assigned to an attribute for a particular object. **Same attribute are mapped to same attribute value.** For example, height is measured in length units but not in weight units as they are not valid units for measuring height.

   ◯ Yes

   ◯ No

   Solution: No. Page 83. *(The page number refers to the page in i2DM 2nd)*

   Same attribute can be mapped to different attribute values. Example: height can be measured in feet or meters.

2. Values used to represent an attribute can have properties that are not inherent to the attribute itself, and can also lack certain properties that the attribute possesses.

   ◯ Yes

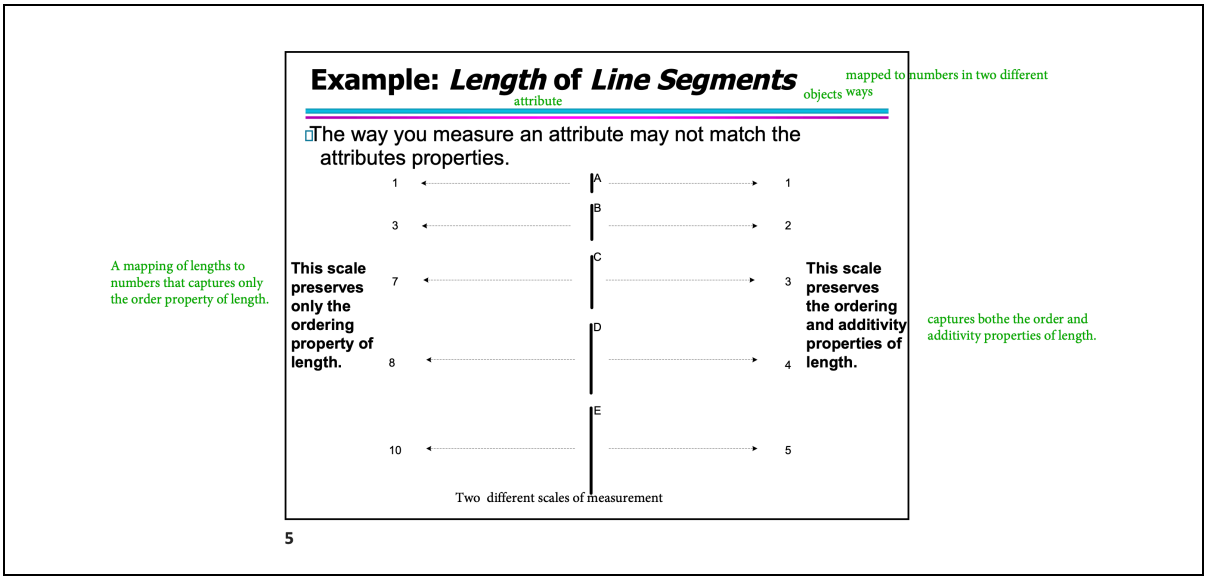
   ◯ No

   Solution: Yes. Page 83.

   Properties of an attribute need NOT be the same as properties of the values used to measure it.

   Alternatively, the values used to represent an attribute can have properties that are NOT properties of the attribute itself, and vice versa.

   In 2, the properties of the scale (attribute value) on the left hand side are different from the the properties of the attribute.



3. The measurement error arises from the measurement process and can include systematic biases from the true value. On the other hand, the data collection error is the error including but not limited to **noise during the data collection**, missing data objects and inappropriately including a data object.

   ◯ Yes

   ◯ No

Solution: No. Page 102.

Measurement error: Any problem resulting from the measurement process, such as **noise**, artifacts, bias, precision, and accuracy.

Data collection error: Errors such as omitting data objects or attribute values, or inappropriately including a data object, including outliers, missing and inconsistent values, and duplicate data.

4. Both precision and bias can be affected by the random error and systematic error.

   ◯ Yes

   ◯ No

Solution: No. Page 105.

Precision can be affected by the random error but not systematic error.

On the other hand, bias can be affected by the systematic error but not random error.

5. Bias is calculated by finding the difference between the mean of the measured values and the true value of the quantity being measured. So bias can only be determined for objects whose quantity is known by external means rather than by the current measurement situation.

   ◯ Yes

   ◯ No

> **Solution:** Yes. Page 105.
>
> For a set of values, the bias can be written as:
>
> $$Bias(\hat{x}) = \mathrm{E}(\hat{x}) - x = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - x$$
>
> where the $\{\hat{x}_i\}_{i=1}^{n}$ is the set of estimated values and x is **the true value of the quantity being measured**.
>
> Thus, bias can be determined only for objects whose **measured quantity is known by means external to the current situation**. Suppose that we have a standard laboratory weight with a mass of 1g and want to assess the precision and bias of our new laboratory scale. We weigh the mass five times, and obtain the following five values: 1.015, 0.990, 1.013, 1.001, 0.9861. The mean of these values is 1.001. and hence, the bias is 0.001.

6. Since the data quality will significantly affect the result of the analysis. During the data cleaning process, it is important for data scientists to identify and remove the noise and outliers from the data to uncover the real patterns and insights in the original data.

   ◯ Yes

   ◯ No

> **Solution:** No. Page 107.
>
> Outliers refer to data objects or values that deviate significantly from the rest of the data points in a dataset. So, **outliers can be legitimate data objects** or values that we are interested in detecting. One **can not simply remove it**.
>
> For instance, in finance, outliers in stock prices or market trends could indicate significant changes in the market or unexpected events that could impact the performance of the stock. Analysts use outlier detection techniques to identify such values and evaluate their impact on the overall market rather than remove them.

7. In the study of precipitation (amount of rainfall) in the UK, for areas without precipitation records, we can use the precipitation records of neighboring areas to estimate.

◯ Yes

◯ No

Solution: Yes. Page 108.

Sometimes missing data can be reliably estimated. In this problem, since the precipitation changes in a reasonably smooth fashion from region to region (spatially correlated). The missing values can be estimated by using the remaining values in their neighborhood.

8. Similarity measure increases when the objects are more alike. Dissimilarity measure decreases when the objects are more alike. It is often the case that similarity measure can be converted to a dissimilarity measure, or vice versa. For example, the cosine similarity can be convert to the euclidean distance as shown in the Tutorial 1. Thus, **similarity measures and dissimilarity measures are interchangeably during data analysis**.

◯ Yes

◯ No

Solution: No. Page 144.

The last sentence is wrong. While similarity and dissimilarity measures may seem similar, they are used for different purposes in data analysis. Therefore, these measures cannot be used interchangeably, and it is important to choose the appropriate measure based on the goals and context of the analysis. For example, in clustering analysis, dissimilarity measures are often used to group similar objects together based on their differences, while in classification analysis, similarity measures may be used to identify the most similar objects to a given target.

9. Distances, such as the euclidean distance, have some well-known properties: symmetry, positivity, and triangle inequality. Measures that satisfy all three properties are referred to as **proximity**. Some individuals restrict the term "distance" to only for dissimilarity measures that satisfy to these properties, but this convention is often disregarded. Some of the similarity or dissimilarity may violate one or more of the properties.

◯ Yes

◯ No

Solution: No. Page 153.

Measures that satisfy all three properties are referred to as **metric**.

10. The Minkowski distance is a generalization of several distance measures, including:

    1. When p = 1, the Minkowski distance is equivalent to the Manhattan distance.
    2. When p = 2, the Minkowski distance is equivalent to the Euclidean distance.
    3. When p = infinity, the Minkowski distance is equivalent to the Supremum distance.

    These special cases of the Minkowski distance are widely used in various applications, including machine learning, pattern recognition, and image processing. All these distances are defined for any numbers of dimensions.

    ◯ Yes

    ◯ No

    Solution: Yes. Page 151.

11. If each binary attribute corresponds to an item in a superstore, where a 1 indicates that the item was purchased, while a 0 indicates that the item was not purchased. The simple matching coefficient can be used to assess the similarity between various transactions.

    ◯ Yes

    ◯ No

    Solution: No. Page 156.

    The number of products not purchased by any customer far out numbers the number of products that were purchased, thus a similarity measure such as $SMC = \frac{TP+TN}{TP+FP+TN+FN}$ would say that all transactions are very similar.

    In this situation, the Jaccard coefficient $= \frac{TP}{TP+FP+FN}$ is frequently used to handle objects consisting of asymmetric binary attributes. Because it only take the items been purchased into consideration.

12. All similarity measures share some common properties, as well as unique properties that are specific to each measure. For example, both the simple matching coefficient (SMC) and **cosine similarity take into account the matches between 0s**, by counting them directly and using inner product, respectively. However, the SMC is restricted to binary vectors like [1, 0, 1], whereas the cosine similarity can handle non-binary vectors such as [3, 2, 0, 5].

    ◯ Yes

    ◯ No

Solution: No. Page 158.

The cosine similarity defined as $\cos(\mathrm{x}, \mathrm{y}) = \frac{\mathrm{x} \cdot \mathrm{y}}{||\mathrm{x}||||\mathrm{y}||}$ ignores the 0-0 match.

One of the reasons why cosine similarity can ignores the 0-0 match is that it is based on the inner product of the vectors x and y, where the product of 0 and 0 equals to 0.

13. When computing the proximity between numerous objects, normalizing the objects to have unit length can decrease the required time without losing accuracy. For example, the cosine similarity and **euclidean distance measure does not consider the length of the two data objects** when assessing similarity. Hence, for vectors with a length of 1, they can be calculated much faster than before.

    ◯ Yes

    ◯ No

Solution: No. Page 159.

The cosine similarity can be rewritten as

$$\begin{aligned} \cos(\mathrm{x}, \mathrm{y}) &= \frac{x \cdot y}{||x||||y||} \\ &= \frac{x}{||x||} \cdot \frac{y}{||y||} \\ &= x' \cdot y' \end{aligned}$$

where $x' = \frac{x}{||x||}$ and $y' = \frac{y}{||y||}$. Dividing x and y by their lengths normalizes them to have a length of 1. This means that cosine similarity does not take the length of the two data objects into account when computing similarity. Actually, it only considers the angle of two vectors.

When length is important, the **Euclidean distance** might be a better choice.

14. The Mahalanobis distance is a generalization of the Euclidean distance that is useful when attributes are correlated, have different ranges of values, and the distribution of the data is approximately Gaussian. However, since we need to calculate the inverse of the covariance matrix when calculating of the Mahalanobis distance. It will be computational intensive to use the Mahalanobis distance in a large dataset.

    ◯ Yes

    ◯ No

Solution: Yes. Page 179.

15. The TF-IDF value for a word increases proportionally to the number of documents in the corpus that contain the word.

    ◯ Yes

    ◯ No

Solution: No. Q15 in Tutorial 1.

Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = \underbrace{tf_{ij}}_{TF} * \underbrace{log\frac{m}{df_i}}_{IDF}$$

where df_i is the number of documents in which the i^{th} term appears and is known as the **document frequency** of the term. This transformation is known as the **inverse document frequency** transformation.

The inverse document frequency helps to adjust for the fact that some words appear more frequently in general. Therefore, the TFIDF value for a word is offset by the number of documents in the corpus that contain the word. So it won't increases proportionally to the number of documents in the corpus that contain the word.

2 Multiple choice questions (there may be more than one correct answer).

1. Which of the following attribute(s) provide(s) enough information to rank the objects?

    A. Hong Kong ID number

    B. Home address

    C. GPA

    D. Phone number

Solution: C. Page 83.

C: GPA is a ordinal attribute, and the others are nominal attribute. As stated in Table 2.2 in page 83, *I2DM, 2nd*, the values of an ordinal attribute provide enough information to order objects using

symbols such as "<" and ">". For example, hardness of minerals, {good, better, best}, GPA and so on.

A, B, D: Nominal attributes provide only enough information to distinguish one object from another, but they do not provide any indication of ranking or hierarchy like the Student ID, Home address and Phone number.

2. Classify the attribute Age in years.
    A. Discrete
    B. Quantitative
    C. Ratio
    D. Interval

Solution: A, B, C. It is the example question of Question 2 in Tutorial 1.

A: Age in years in an integer, so it is discrete.

B, C: Age is a quantitative ratio attribute because it has a true zero point, which is the point at which a person is born.

3. Classify the attribute Education Level with two choices: {High School, College}.
    A. Binary
    B. Interval
    C. Ordinal
    D. Qualitative

Solution: A, C, D. Question 2 in Tutorial 1.

A: It is a binary attribute since there are two possible outcomes.

C, D: The education level College is higher than High School, so it is ordinal and qualitative.

4. Which of the statement(s) is/are true about the attribute employee age
    A. Measurement scale refers to a rule or function that maps the physical value of an attribute of an object to a numerical or symbolic value.
    B. The age values can have properties that are not properties of the age itself.
    C. The age can have properties that are not properties of the age value itself.

D. Attributes are inherent to objects. To facilitate analysis, researchers assign numerical or symbolic values to them for precise and detailed examination of their characteristics.

Solution: A, B, C, D. Page 78.

5. Which of the following statement(s) is/are true regarding measurement errors and data collection errors?

   A. The measurement error refers to any issue that arises from the process of measurement, such as the reading error when measuring the length. So Measurement errors only occur in continuous attributes.

   B. Data collection errors include omitting data objects or attribute values.

   C. Measurement errors and data collection errors can both lead to large bias and low precision in the final model.

   D. Because the error is mixed with the real data, we can't detecting and correcting data entry errors.

Solution: B, C. Page 102.

A: Measurement errors can occur in both continuous and discrete attributes. For example, in a discrete attribute like age, errors can occur due to incorrect data entry or rounding.

B: The term data collection error refers to errors such as **omitting** data objects or attribute values or **inappropriately including** a data object.

C: Both measurement errors and data collection errors can be either systematic or random, thus both of them can lead to large bias and low precision in the final model. Here are some examples of each type of error:

Type of Error	Measurement	Data Collection
Systematic	A scale that is consistently off by 2 pounds for every measurement of weight	A survey question that consistently leads respondents to answer in a certain way
Random	Occasional inaccuracies in measurement due to natural variation or chance	Occasional errors in data entry or sampling due to chance or human error

D: While errors are common in data entry, there are many techniques available to detect and correct these errors. For instance, there is a module called "data validation" in excel to detect the potential invalid input.

6. Suppose a lab is measuring the concentration of a certain chemical in a sample, and the true concentration is 2.5 cm. The lab performs the measurement 10 times and obtains the following values: 2.0, 2.2, 2.8, 2.2, 3.0, 2.8, 2.8, 2.2, 2.4, 2.6. What can we say about the precision (use standard deviation = 0.1 as threshold) and bias of the lab's measurements?

    A. The precision is high, bias is low.

    B. The precision is low, bias is low.

    C. The precision is high, bias is high.

    D. The precision is low, bias is high.

Solution: B. Page 105.

Precision refers to the closeness of repeated measurements to one another. In this case, the measurements range from 2.0 to 3.0, with a standard deviation of 0.31, indicating a low precision.

Bias refers to a systematic variation of measurements from the quantity being measured. In this case, the mean of the measurements is 2.5, which is the same as the true value of the quantity being measured, indicating a low bias.

7. Which of the following technique(s) can be used to detect outliers in a given data set?

    A. Scatter plot

    B. Box plot

    C. Correlation heatmap

    D. None of the above statements are true.

Solution: A, B. Page 106.

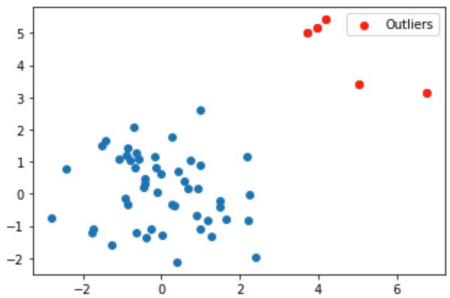
A, B: We can exactly located the outliers by scatter plot and box plot.

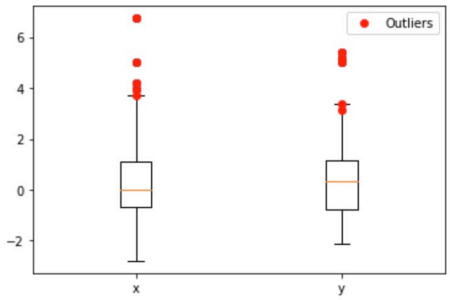
C: The correlation heatmap is a technique used for visualizing the relationship between two or more variables but do not necessarily help in identifying outliers.

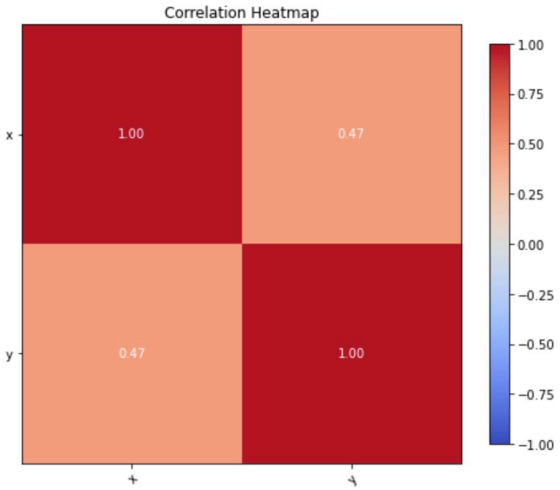
Here we generate 50 normal points and 5 outliers:

```
np.random.seed(123)
x = np.concatenate([np.random.normal(0,1,50), np.random.normal(5,1,5)])
y = np.concatenate([np.random.normal(0,1,50), np.random.normal(5,1,5)])
```

The corresponding Scatter plot, Box plot and Correlation heatmap are illustrated below:







While Scatter plot and Box plot make it easy for us to identify outliers from normal points, it is difficult to do the same in Correlation heatmap.

8. Consider a dataset containing 1000 data objects, where each object has 10 attributes. If 500 data objects have missing values for at least one attribute, which of the following strategies for dealing with missing values is/are applicable in this case?

    A. Eliminate all objects with missing values.

    B. Estimate missing values using the mean, median or mode of the available data for the corre-

sponding attribute.

C. Impute missing values using regression.

D. Using algorithms that can handle missing values.

Solution: B, C, D. Page 107.

A: Eliminating all objects with missing values is not a good strategy in this case because the are 500 out of 1000 objects have missing value. Eliminating 50% of the observations could result in losing valuable information and biasing the model.

B: Replacing missing values with the mean, median or mode of the attribute is a simple and effective imputation method that can be used for both continuous and categorical attributes, respectively. However, it is important to note that this method assumes that the missing values are missing at random and does not take into account any relationships between attributes.

C: Replacing missing values via regression is also acceptable as long as the missing values are missing at random.

D: Using algorithm that can handle missing values is also a viable strategy. Some decision tree algorithms, such as C4.5 and CART, can handle missing values by **treating them as a separate category**. This allows the algorithm to split the data based on the presence or absence of missing values, which can be useful in some cases.

9. Which of the following statement(s) is/are true regarding similarities and dissimilarities?

A. The term "distance" is a synonym for similarity, which will decrease when two objects are alike. Distance may sometimes fall within the range of [0, 1], but it is also common for them to span from 0 to infinity.

B. Different attributes require different proximity measurements. Jaccard and cosine similarity are appropriate for non-sparse (dense) data like time series, as well as correlation and Euclidean distance measures that are suitable for sparse data like documents.

C. Typically, any monotonically decreasing function can be used to convert dissimilarities to similarities, or vice versa. Nevertheless, it is important to consider other factors during the conversion process.

D. The Minkowski distance between $\{0, 1, 0\}$ and $\{5, 2, 0\}$ when $p = \infty$ is 5.

Solution: C, D. Page 144.

A: The term "distance" is a synonym for dissimilarity.

B: This involves various measures, including Jaccard and cosine similarity measures that are appropriate for **sparse** data like documents, as well as correlation and Euclidean distance measures that are suitable for **non-sparse** (dense) data like time series or multi-dimensional points.

D: As p approaches infinity, the Minkowski distance approaches maximum difference of coordinates, which is $5 - 0 = 5$ in this case.

10. Suppose we have a confusion matrix for a set of 2 characters: A and B, where each character is classified 100 times. The diagonal elements of the confusion matrix represent the number of times a character is classified correctly. Given that character A is misclassified as character B 25 times, and character B is misclassified as character A 30 times. We can define a non-symmetric similarity metric based on the misclassification rate

$$s(x, y) = \frac{\text{Number of x that is misclassified as y}}{\text{Total number of x}}$$

and make it symmetric using the formula $s'(x, y) = \frac{s(x,y)+s(y,x)}{2}$. What is the new symmetric similarity measure between character A and B?

A. 0.125

B. 0.225

C. 0.250

D. 0.275

Solution: D. Page 155.

We first calculate the non-symmetric similarity metric:

$$s(A, B) = \frac{25}{100}$$

$$s(B, A) = \frac{30}{100}$$

then the new symmetric similarity measure between character A and B is

$$s'(x, y) = \frac{s(x, y) + s(y, x)}{2} = \frac{\frac{25}{100} + \frac{30}{100}}{2} = 0.275$$

11. Which of the following set(s) has the smallest gap between Simple Matching Coefficient (SMC) and the Jaccard coefficient?

A. {1, 1, 0, 0, 0, 0, 0, 0, 0, 0} and {1, 0, 0, 0, 0, 0, 0, 0, 0, 0}

B. {1, 1, 1, 0, 0, 0, 0, 0, 0, 0} and {1, 0, 0, 0, 0, 0, 0, 0, 0, 0}

C. {1, 1, 1, 0, 0, 0, 0, 0, 0, 0} and {1, 1, 0, 0, 0, 0, 0, 0, 0, 0}

D. {1, 1, 1, 1, 1, 1, 0, 0, 0, 0} and {1, 1, 1, 1, 1, 0, 0, 0, 0, 0}

Solution: D. Page 157.

The gaps of A, B, C and D are: 0.4, 0.467, 0.233 and .067, respectively.

12. Which of the choice(s) has cosine similarity larger than 0.7?

A. {1, 1} and {1, 0}

B. {2, 1} and {1, 0}

C. {1, 2} and {1, 0}

D. {1, 1} and {2, 0}

Solution: A, B, D. Page 158.

The cosine similarity of A, B, C and D are: 0.707, 0.894, 0.447 and 0.707, respectively.

13. Which of the following statements regarding correlation is/are true?

A. Correlation measures the linear relationship between two sets of values.

B. Correlation can only be used to measure the relationship between two variables of the same attribute type.

C. There are many types of correlation.

D. All above.

Solution: A, C. Page 160.

A: The definition of correlation is a measure used to calculate the linear relationship between two sets of values that are observed together. To better demonstrate it, let us consider a special case:

Suppose $x, x \in \mathbb{R}$ and both x and y are scaled (with **mean zero** and **standard deviation one**). Consider the linear model:

$$y = \alpha + \beta * x + \epsilon$$

the least square estimator of β is

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\ &= \frac{S_{xy}}{S_{xx}} \\ &= S_{xy}\end{aligned}$$

where $S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ and $S_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1$. In this situation, the correlation is approximately equal to the slope of the linear model. Therefore, correlation indeed measures the linear relationship between two sets of values.

B: Correlation can be used to measure the relationship between two variables, irrespective of their attribute type. It is often the case that we measure the correlation between two variables of different attribute type in the Correlation matrix or Correlation heatmap. For example, the correlation can measure the relationship between two variables height and weight.

C: There are many types of correlation measures available. Some of the commonly used correlation measures include Pearson correlation, Spearman correlation, and Kendall correlation.

14. What's the Pearson's correlation between x = {-2, -1, 0, 1, 2} and y = {4, 1, 0, 1, 4}, which y is derived from squaring each element in x: $y_k = x_k^2$.

    A. 1

    B. 0.5

    C. 0

    D. -1

Solution: C. Page 161.

The correlation is 0. This case shows the Pearson's correlation only measures the linear relationship between the two sets of values. The nonlinear relationships can exist, but their correlation is 0.

15. Which of the statement best describe the object of following python code?

```
def FunctionName(x, y):
    if len(x) != len(y):
        raise ValueError("Undefined for sequences of unequal length")
    x = np.array(x); y = np.array(y)
    d = np.sqrt(np.sum((x-y)**2))
return d
```

    A. To calculate the cosine similarity

    B. To calculate the euclidean distance

    C. To calculate the simple matching coefficient

    D. To calculate the jaccrad distance

Solution: B. Question 16 in Tutorial 01.

The function takes two arguments x and y, which are the vectors for which the Euclidean distance needs to be calculated.

The first line of the function checks if the length of both vectors is equal. If they are not equal, then it raises a ValueError with the message "Undefined for sequences of unequal length".

Then the function converts both the input vectors into numpy arrays.

The next line calculates the Euclidean distance between the two vectors using the formula for Euclidean distance:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

where n is the number of dimensions of the vectors.

Finally, the function returns the calculated Euclidean distance d.

16. Which of the following choices is/are the data preprocessing techniques?

    A. Cluster sampling

    B. Exploratory data analysis

    C. One-hot encoding

    D. Fourier transformation

Solution: A, C, D. Page 114.

B: Exploratory data analysis (EDA) is not a data preprocessing technique because it does not involve modifying or transforming the original data in any significant way.

EDA is an important step in the data analysis process where the main objective is to gain insights into the data by exploring its distribution, patterns, and relationships. This involves techniques such as data visualization, summary statistics, and data exploration methods like scatter plots, histograms, and box plots. The main goal of EDA is to identify patterns and relationships in the data that can help in making informed decisions about modeling rather than modify the original data to improve its quality, reduce noise, or make it easier to analyze.

17. Which of the following statement is/are true about the data preprocessing?

    A. Data preprocessing involves removing redundant data objects and attributes for analysis or modifying the attributes to prepare the data for further processing.

B. On average, the actual amount of variation of aggregated quantities will be smaller than the individual values being aggregated.

C. Both random error and systematic error can be removed the by data aggregation.

D. Different attributes require different preprocessing techniques. Quantitative attributes can be aggregated by taking a sum or an average. But, data aggregation is not capable for qualitative attributes like gender. We can use one-hot encoder to convert them.

Solution: A. Page 116

B: Aggregated quantities, such as averages or totals, are typically less variable than the individual values being combined. However, in the case of totals, the actual amount of variation can be larger than that of individual objects on average.

C: Data aggregation cannot remove the systematic error during the data collection process. Systematic errors are errors that occur consistently and affect the accuracy of the whole dataset. Aggregating data may reduce the impact of random errors (errors that occur due to chance) but cannot remove systematic errors.

D: When dealing with quantitative attributes such as price, they are usually aggregated by taking a sum or an average. On the other hand, when it comes to qualitative attributes like item type, they can either be omitted or summarized using a higher-level category, such as from different brands of televisions and mobiles to televisions versus electronics.

18. Which of the following statement is/are true about the sampling?

A. The fundamental principle for effective sampling is the sample is representative. To ensure the representativeness, one should keep the probability of selecting any object constant.

B. It is necessary to choose the sample size. The lower the proportion being sampled from the population, the poorer the representativeness of the samples even under different sampling methods.

C. There are many sampling methods. However, in order to ensure the fairness of sampling, sample size should be determined before sampling.

D. Sampling aims to select a smaller, representative group from a larger population. It is commonly used in statistics and data mining, with different motivations.

Solution: D. Page 117.

A: When dealing with rare classes, it is crucial to ensure that they are adequately represented in the sample. As such, a sampling scheme that can accommodate differing frequencies for the object types of interest is needed. Stratified sampling is an approach that starts with pre-specified groups.

B: The sampling method used can also affect the representativeness of the sample. For instance, if a simple random sampling method is used, the sample may not be representative if the population is not homogeneous. In such cases, stratified sampling or cluster sampling may provide more representative samples even the sample size is lower.

C: Page 122. Adaptive or progressive sampling is a sampling technique that is used when the population of interest is heterogeneous, meaning that it contains subgroups with different characteristics. In adaptive sampling, the sample size is adjusted based on the results of previous samples to improve the representativeness of the sample.

19. Which of the following statement is/are true about the dimensionality reduction?

    A. PCA is a technique for dimensionality reduction that captures maximum variation in the data, so it will increase noise as a trade-off for reducing dimensions.

    B. The curse of dimensionality is the difficulty in analyzing high-dimensional data due to sparsity in the space it occupies. This leads to poor classification accuracy and clustering quality, making it harder to create representative models for classification and clustering.

    C. Dimensionality reduction can be achieved by directly deleting some features. Prior knowledge or common sense can be used by the researcher to perform the feature selection.

    D. Different dimensionality reduction techniques can lead to different subsets. For a set of redundant attributes, feature subset selection tends to select a subset of attributes from the original set. While, the PCA may output some linear combinations incorporate attributes.

Solution: B, C, D.

A: Principal Component Analysis (PCA) is used to denoise and reduce dimensionality. The reduction in dimensionality achieved by PCA does not necessarily lead to an increase in noise. In fact, noise in the data can be reduced by retaining only the principal components that capture the most meaningful variation in the data and discarding the components that capture only noise or unimportant variation.

20. Which of the following preprocessings is/are feature creation?

    A. Using convolutional neural networks to separate license plate images from images captured by cameras in the Cross Harbour Tunnel.

    B. Mapping the periodic fluctuated stock yield curve to a series of combinations of sine and cosine functions.

    C. Using attributes distance and duration to construct a new attribute called velocity: velocity = distance / duration.

D. Selecting one attribute randomly from a group of redundant attributes and eliminate the remaining attributes.

Solution: A, B, C, D. Page 129.

There are three types of feature creation:

- Feature extraction
- Mapping data to new space
- Feature construction

A: Feature extraction.

B: Mapping data to new space

C: Feature construction

D: Feature extraction.