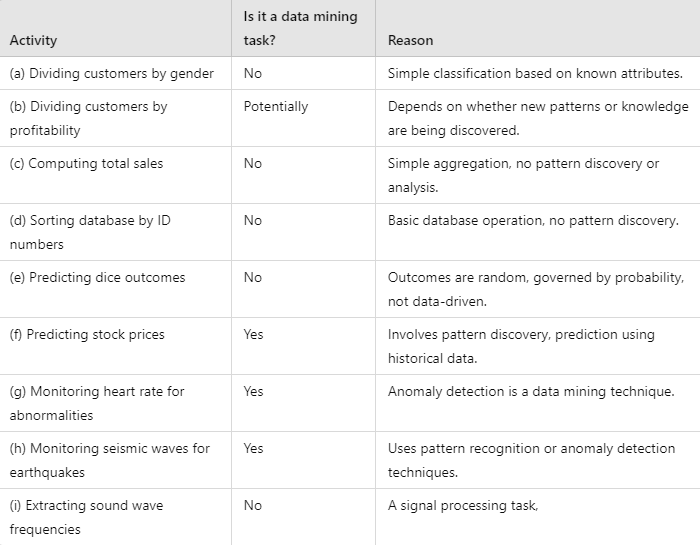
BDM-Tut01

Ex1.7



2.

As a data mining consultant for an Internet search engine company, I would recommend using various data mining techniques to improve the company’s services, enhance user experience, and gain valuable insights. Each of these techniques can be applied in different ways to achieve specific business objectives.

Clustering can be used to group users based on their behaviors, preferences, and interests. For instance, users who frequently search for topics related to technology can be clustered together. This allows the search engine to personalize results and advertisements, ensuring that users are presented with content that is more relevant to their interests. Similarly, clustering can be applied to group similar web pages or documents, which can help the search engine categorize content more efficiently and improve the relevancy of search results. Additionally, this approach can be used to identify emerging topics within certain domains, ensuring that the search engine stays up-to-date with the latest trends.

Classification techniques can be utilized to predict the intent behind a user's search query. For example, queries like “buy laptop” can be classified as transactional intent, while queries like “how to fix laptop keyboard” may be classified as informational intent. This allows the search engine to adjust the ranking of results based on the perceived intent, improving the user experience by presenting the most appropriate content. Moreover, classification algorithms can be employed to detect spam or irrelevant content on websites. By classifying pages as either spam or non-spam, the search engine can filter out low-quality results, ensuring that users are presented with trustworthy and valuable information.

Association rule mining can help identify patterns in user behavior, which can be especially valuable for recommendation systems. For instance, if users who search for "smartphones" also often search for "smartphone cases," the search engine can use this knowledge to recommend related products or content to users, increasing engagement and satisfaction. Association rule mining can also be applied to understand co-occurrence patterns in search queries. For example, if users searching for “healthy recipes” often search for “meal prep ideas,” the search engine can link these topics together, providing users with more relevant content and improving their search experience.

Anomaly detection plays a critical role in identifying unusual patterns in user behavior or website activity. For example, if a user begins performing an unusually high number of searches in a short time, this could indicate automated bot activity or a potential security issue. By detecting such anomalies, the search engine can take appropriate action to prevent fraud and ensure the integrity of its system. Anomaly detection can also be used to identify outlier data, such as web pages with significantly higher or lower click-through rates compared to similar pages. These anomalies could indicate poorly performing content or pages that are unexpectedly popular, prompting the company to investigate and optimize these pages to improve the user experience.

In summary, data mining techniques like clustering, classification, association rule mining, and anomaly detection can significantly enhance a search engine's performance. They help improve the relevancy of search results, personalize content for users, detect and prevent fraud, and ensure that the content indexed by the search engine remains valuable and trustworthy.

3.

**(a) Census Data Collected from 1900–1950:**

Data privacy is not an important issue for census data collected from 1900–1950, as the data is likely to be over 70 years old, and public records like these have generally become open for historical research or public access in many countries. The people involved are no longer alive, so there are no direct privacy concerns regarding these individuals. However, privacy concerns might arise if the data is combined with other information that could potentially re-identify individuals, but as a historical dataset, privacy is generally not a major concern.

**(b) IP Addresses and Visit Times of Web Users Who Visit Your Website:**

Data privacy is a significant issue when dealing with IP addresses and visit times of web users. While IP addresses alone might not always directly identify an individual, they can be used to approximate a person's location or, when combined with other data, could potentially be used to identify or track users across different sessions. Furthermore, collecting visit times and browsing habits could lead to privacy concerns, especially if the data is shared or used for targeted advertising without the users' consent. It's important to protect this data through encryption and implement transparent privacy policies.

**(c) Images from Earth-orbiting Satellites:**

In most cases, data privacy is not a major concern with images from Earth-orbiting satellites. These images typically capture geographical features, weather patterns, or general environmental data. However, privacy concerns might arise if satellites capture images of private property or individuals in a way that could be considered invasive. The key issue would be when satellite imagery is used to monitor specific individuals or private property in an unauthorized manner, which could raise privacy and legal concerns.

**(d) Names and Addresses of People from the Telephone Book:**

While telephone books were once a public resource, with the advent of digital privacy laws and protections, data privacy is an important issue concerning names and addresses. In many countries, personal information such as names and addresses is now protected by data privacy regulations (e.g., GDPR in the EU or CCPA in California). These laws aim to prevent the misuse of personal data, such as identity theft, targeted marketing, or harassment. Therefore, data privacy is a concern when it comes to using, sharing, or making available this information without the individuals' consent.

**(e) Names and Email Addresses Collected from the Web:**

Data privacy is a critical issue when dealing with names and email addresses collected from the web. Individuals' names and email addresses are personally identifiable information (PII), and using or sharing this information without proper consent can lead to privacy violations, spam, and other types of misuse. Collecting email addresses for marketing purposes or without the users' awareness or consent can violate data protection regulations like GDPR. Therefore, it is crucial to ensure that individuals' personal information is handled responsibly, with adequate security measures and transparent consent processes in place.

**1.9**

**1- (a) Conducting surveys and uploading to database:**

**Data Collection**: This step involves gathering raw data from participants via surveys and storing it in a database for further processing. The activity of conducting surveys and uploading the collected responses is part of the data collection phase.

**(b) Correcting missing entries:**

**Data Preprocessing**: This step involves cleaning the data by addressing issues such as missing values, errors, or inconsistencies. Correcting missing entries is a typical task in the data preprocessing phase, where the analyst ensures the data is in a usable form for analysis.

**(c) Designing a recommendation algorithm:**

**Data Analysis**: Designing a recommendation algorithm involves analyzing the preprocessed data and creating a model or system that can make recommendations based on user preferences or behaviors. This is part of the data analysis phase, where insights are extracted from the data to guide decisions or predictions.

**2**

(a) Age: Numeric (Continuous or Discrete)

(b) Salary: Numeric (Continuous)

(c) ZIP code: Categorical (Nominal)

(d) State of residence: Categorical (Nominal)

(e) Height: Numeric (Continuous)

(f) Weight: Numeric (Continuous)

**3**

The original data, which consists of medical notes from a physician, is typically unstructured and in the form of text. These notes are free-form and contain descriptions, symptoms, and other details that are not organized in a standardized format, making it challenging to analyze directly.

After transformation, the data is organized into a structured format, such as a table, where each row represents a patient and each column corresponds to specific attributes, like the medicines prescribed. This tabular format is considered structured data because it follows a clear and predefined organization, making it easier for analysis and processing.

The process of transforming the data from its original unstructured form into a structured format is known as data transformation or data wrangling. This process involves extracting relevant information from the unstructured medical notes and organizing it into a structured format, cleaning, normalizing, and potentially encoding the data to make it suitable for analysis and data mining tasks.

**4.**

The data collected by the sensor network to measure temperature at different locations over a period is **numeric (continuous)**. Temperature is a measurable quantity that can take on a wide range of values, including decimal points (e.g., 21.5°C or 70.3°F). As it represents a continuous variable, it is classified as numeric data. Additionally, because the sensor records measurements over time and across different locations, the data is structured and can be organized in a time series format

**5**

The process of creating a single database by combining the analyst's temperature readings with the pressure readings from a different source is called data integration. Data integration involves combining data from different sources into a unified view or database, making it easier to analyze and work with. In this case, the temperature and pressure readings would be integrated into a single database, potentially aligning them by time or location, depending on the structure of the data.

6

The type of data created from processing web logs to generate records with ordering information for web page accesses from different users is sequential data. Sequential data refers to data that is ordered in a sequence, where the order of events or transactions is important. In this case, the data represents the sequence in which web pages are accessed by users, capturing the flow or order of user interactions with the website. This kind of data is often used for analyzing user behavior, website navigation patterns, or building models for recommendation systems.

**7.**

**Data Type of Nucleotide Sequence:**

The data corresponding to a set of nucleotides arranged in a certain order is **sequential data**. This type of data represents a series of elements (in this case, nucleotides) arranged in a specific sequence, often used in biological contexts like DNA or RNA sequences.

**8.**

**Partitioning Customers into Similar Groups Based on Demographic Profile:**

The best data mining problem for partitioning customers into similar groups based on their demographic profile is **clustering**. Clustering algorithms group data objects (customers) based on their similarities, allowing the analyst to identify patterns or groups within the data, such as customers with similar ages, incomes, or locations.

**9.**

**Identifying Groups of Customers Likely to Buy Widgets in the Future:**

The best data mining problem for identifying groups of customers who might buy widgets in the future, given that some customers' purchasing behavior is already known, is **classification**. Classification techniques can help predict the likelihood of a customer purchasing a widget based on demographic or other available data, allowing for segmentation of customers who may exhibit similar buying behaviors.

**10.**

**Finding Sets of Items Often Bought Together with Widgets:**

The best data mining problem for finding sets of items that are often bought together with widgets is **association rule mining**. Association rule mining identifies relationships between variables in large datasets. In this case, it would uncover patterns that indicate which items are commonly purchased alongside widgets, allowing for insights into product bundling and customer preferences.

**11.**

**Identifying Customers Who Lie About Their Demographic Profile:**

The best data mining problem for finding customers who might lie about their demographic profile, based on discrepancies in buying behavior and demographic data, is **anomaly detection**. Anomaly detection identifies data points that significantly differ from the rest of the data, which in this case could be customers whose demographic profiles do not align with their purchasing behaviors, suggesting potential false or inaccurate information.

**2.6**

(a) Time in terms of AM or PM: Binary and qualitative (nominal), as there are only two categories, AM and PM, with no inherent order.  
(b) Brightness as measured by a light meter: Continuous and quantitative (ratio), because it can be measured precisely with a meaningful zero point.  
(c) Brightness as measured by people’s judgments: Discrete and qualitative (ordinal), since judgments are often categorized with an implied order (e.g., dim, medium, bright) but may lack consistent intervals.  
(d) Angles as measured in degrees between 0 and 360: Continuous and quantitative (interval), as angles can be measured precisely but lack a true zero point, given that 0° is equivalent to 360°.  
(e) Bronze, Silver, and Gold medals as awarded at the Olympics: Discrete and qualitative (ordinal), as the medals have a meaningful order of rank but no measurable distance between levels.  
(f) Height above sea level: Continuous and quantitative (ratio), with a true zero at sea level and the ability to compare heights proportionally.  
(g) Number of patients in a hospital: Discrete and quantitative (ratio), since the count of patients is a whole number with a meaningful zero (no patients).  
(h) ISBN numbers for books: Discrete and qualitative (nominal), as the numbers are identifiers with no inherent numerical meaning.  
(i) Ability to pass light in terms of opaque, translucent, and transparent: Discrete and qualitative (ordinal), because the categories have a meaningful order based on light transmission.  
(j) Military rank: Discrete and qualitative (ordinal), as ranks imply a clear hierarchy but do not have measurable intervals between them.  
(k) Distance from the center of campus: Continuous and quantitative (ratio), with a true zero (at the center) and measurable proportions.  
(l) Density of a substance in grams per cubic centimeter: Continuous and quantitative (ratio), as it can be measured precisely with a meaningful zero point.  
(m) Coat check number: Discrete and qualitative (nominal), as the numbers serve only as identifiers without numerical or ordered significance.

**Aggregation**: Aggregation is the process of summarizing or combining data to reduce its size or complexity while retaining its essential characteristics. It is motivated by the need to simplify analysis, reduce noise, and provide a higher-level understanding of data. When aggregating data, numerical attributes are typically combined using summary statistics (e.g., sum, mean, median), while categorical attributes may be summarized by mode or counts.

**Sampling**: Sampling is the process of selecting a subset of data from a larger dataset. A simple random sample should be used when every item has an equal chance of being selected, ensuring unbiased representation. Stratified sampling is preferred when the data consists of distinct groups, and it is important to ensure representation from each group. The size of the sample is critical, as a larger sample reduces variability and improves the accuracy of the results.

**Feature Extraction vs. Feature Creation**: Feature extraction involves selecting or transforming the most relevant features from existing data, while feature creation involves generating new features based on domain knowledge or by combining existing features.

**Customer Satisfaction Measure**:  
(a) The boss is correct. The number of complaints alone does not accurately measure satisfaction, as it does not account for the total number of customers or sales. A better measure would normalize complaints by the total number of customers or transactions, such as calculating the complaint rate.  
(b) The original satisfaction attribute (number of complaints) is discrete and quantitative (ratio), as it is a count with a meaningful zero but does not directly reflect satisfaction levels without normalization.

**3.7**

**1. Linear Interpolation on a Time Series**  
The given time series is (−3,−1,1,3,5,7,∗)(-3, -1, 1, 3, 5, 7, \*)(−3,−1,1,3,5,7,∗). Using a window size of 3 for linear interpolation, we center the window around the missing value. The relevant window is (5,7,∗)(5, 7, \*)(5,7,∗). Since the difference between consecutive elements is 2, the missing value is estimated as 7+2=97 + 2 = 97+2=9. Therefore, the missing value is 9.

**2. Technologies to Identify Personalities in Text**  
To determine personalities mentioned in text documents, Natural Language Processing (NLP) techniques are applied. Named Entity Recognition (NER), a specific NLP task, is used to extract names of persons, organizations, and locations. This involves preprocessing the text (tokenization and POS tagging) and applying machine learning or deep learning models trained for NER. Libraries like SpaCy, NLTK, or Hugging Face can be used for this purpose.

**3. Normalize and Discretize the Arrhythmia Data Set**  
Normalization involves adjusting each attribute of the dataset such that it has a mean of 0 and a standard deviation of 1. This can be achieved by subtracting the mean and dividing by the standard deviation for each attribute. For discretization into 10 equi-width ranges, divide the range of values for each attribute into 10 intervals of equal size. For 10 equi-depth ranges, divide the data into 10 intervals containing an equal number of data points.

**4. Converting Objects into Multidimensional Data for Clustering**  
When similarity values between pairs of objects are available, these values can be used to create a similarity matrix. To convert the data into a multidimensional form, techniques like Multidimensional Scaling (MDS) or t-SNE can be applied. These methods project the data into a lower-dimensional space while preserving the pairwise similarity relationships.

**5. Converting Sea-Surface Temperature and Associated Text into Multidimensional Data**  
Each data record contains a 10×1010 \times 1010×10 grid of temperature values and associated text. Flatten the 10×1010 \times 1010×10 grid into a 100-dimensional vector to represent the spatial temperature values. For the associated text, apply text vectorization techniques like TF-IDF or word embeddings to represent the text in a multidimensional form. Combine these vectors into a single multidimensional representation for each record.

**6. Creating a Multidimensional Representation of Protein Sequences and Annotations**  
Protein sequences can be encoded using techniques like one-hot encoding or sequence embeddings based on biological properties. The text annotations can be vectorized using TF-IDF, word embeddings, or other NLP techniques. Concatenate the protein sequence representation with the text vector to form a unified multidimensional representation.

**7. Applying PCA to the Musk Data Set**  
Download the Musk dataset and preprocess it to ensure numerical values are mean-centered. Perform Principal Component Analysis (PCA) to find the eigenvalues and eigenvectors. These represent the variance captured by each principal component and the directions of maximum variance, respectively. Report the resulting eigenvalues and eigenvectors after computation.

**8. Applying SVD to the Musk Data Set**  
Perform Singular Value Decomposition (SVD) on the Musk dataset. SVD decomposes the data matrix into three matrices: UUU, Σ\SigmaΣ, and VTV^TVT. The singular values in Σ\SigmaΣ correspond to the square root of the eigenvalues from PCA, and VTV^TVT contains the principal directions. Report the singular values and vectors obtained.

**9. Proof of the Formula**  
To show that the given formula holds for mean-centered data points X1,X2,...,XnX\_1, X\_2, ..., X\_nX1​,X2​,...,Xn​, expand the terms using the definitions of squared Euclidean distance and mean-centered points. Demonstrate algebraically that the formula satisfies the equality by leveraging the properties of summations and symmetry in distances.

**10. Wavelet Decomposition of Time Series**  
The time series 1,1,3,3,3,3,1,11, 1, 3, 3, 3, 3, 1, 11,1,3,3,3,3,1,1 is decomposed using a wavelet transform. The decomposition yields coefficients representing the signal at different resolutions. After applying the wavelet transformation, determine how many coefficients are nonzero by analyzing the output.

**11. Wavelet Transformation on Berkeley Data Set**  
Download the Intel Research Berkeley dataset and extract the temperature values from the first sensor. Apply a wavelet transformation to these values to decompose the signal into coefficients at different frequency levels. Analyze the resulting transformed values.

**12. Wavelet Decomposition of KDD CUP 1999 Data**  
Treat each quantitative variable in the KDD CUP 1999 dataset as a time series. For each time series, apply wavelet decomposition to extract frequency-domain information. This decomposition helps analyze the temporal structure of the dataset.

**13. Sampling and Standard Deviations in KDD CUP 1999 Data**  
Create samples of different sizes (n=1,10,100,1000,10000n = 1, 10, 100, 1000, 10000n=1,10,100,1000,10000) and compute the average value eie\_iei​ for each quantitative attribute. Compare eie\_iei​ with the global mean μi\mu\_iμi​ and compute the deviation zi=∣ei−μi∣σiz\_i = \frac{|e\_i - \mu\_i|}{\sigma\_i}zi​=σi​∣ei​−μi​∣​. Observe how ziz\_izi​ decreases as nnn increases, demonstrating the effect of larger sample sizes on stability.

**14. Right Singular Vector with Zero Singular Value**  
Show that if yyy is a right singular vector of matrix AAA with a singular value of 0, then Ay=0Ay = 0Ay=0. This follows directly from the definition of SVD, where AAA is decomposed into UΣVTU\Sigma V^TUΣVT, and singular values in Σ\SigmaΣ correspond to the magnitude of AyAyAy.

**15. Diagonalization as a Specialized Case of SVD**  
Demonstrate that diagonalization of a square matrix is a special case of SVD. For a diagonalizable matrix, the eigenvectors form the orthogonal matrices UUU and VVV, and the eigenvalues populate the diagonal of Σ\SigmaΣ. This aligns with the structure of SVD for symmetric or Hermitian matrices.

**3.9**

**1. Compute the LpL\_pLp​-norm between (1,2)(1, 2)(1,2) and (3,4)(3, 4)(3,4) for p=1,2,∞p = 1, 2, \inftyp=1,2,∞**

The points are (1,2)(1, 2)(1,2) and (3,4)(3, 4)(3,4). The general formula for the LpL\_pLp​-norm is:

∣∣X−Y∣∣p=(∑i=1n∣xi−yi∣p)1/p||X - Y||\_p = \left( \sum\_{i=1}^n |x\_i - y\_i|^p \right)^{1/p}∣∣X−Y∣∣p​=(i=1∑n​∣xi​−yi​∣p)1/p

For p=1p = 1p=1:

∣∣X−Y∣∣1=∣1−3∣+∣2−4∣=2+2=4||X - Y||\_1 = |1 - 3| + |2 - 4| = 2 + 2 = 4∣∣X−Y∣∣1​=∣1−3∣+∣2−4∣=2+2=4

For p=2p = 2p=2:

∣∣X−Y∣∣2=(1−3)2+(2−4)2=4+4=8≈2.83||X - Y||\_2 = \sqrt{(1 - 3)^2 + (2 - 4)^2} = \sqrt{4 + 4} = \sqrt{8} \approx 2.83∣∣X−Y∣∣2​=(1−3)2+(2−4)2​=4+4​=8​≈2.83

For p=∞p = \inftyp=∞:

∣∣X−Y∣∣∞=max⁡(∣1−3∣,∣2−4∣)=max⁡(2,2)=2||X - Y||\_\infty = \max(|1 - 3|, |2 - 4|) = \max(2, 2) = 2∣∣X−Y∣∣∞​=max(∣1−3∣,∣2−4∣)=max(2,2)=2

The LpL\_pLp​-norm values are:  
For p=1p = 1p=1: 444.  
For p=2p = 2p=2: ≈2.83\approx 2.83≈2.83.  
For p=∞p = \inftyp=∞: 222.

**2. Mahalanobis Distance as Transformed Euclidean Distance**

The Mahalanobis distance between two points XXX and YYY is defined as:

dM(X,Y)=(X−Y)TΣ−1(X−Y)d\_M(X, Y) = \sqrt{(X - Y)^T \Sigma^{-1} (X - Y)}dM​(X,Y)=(X−Y)TΣ−1(X−Y)​

Here, Σ\SigmaΣ is the covariance matrix of the dataset. Let ZZZ represent the transformed data points, where:

Z=WXandW=Σ−1/2Z = W X \quad \text{and} \quad W = \Sigma^{-1/2}Z=WXandW=Σ−1/2

The transformed dataset aligns the axes with the principal components, and normalizes by their standard deviations. After the transformation, the Mahalanobis distance simplifies to:

dM(X,Y)=∣∣ZX−ZY∣∣2d\_M(X, Y) = ||Z\_X - Z\_Y||\_2dM​(X,Y)=∣∣ZX​−ZY​∣∣2​

Thus, the Mahalanobis distance becomes equivalent to the Euclidean distance in the transformed space.

**3. LpL\_pLp​-Distance and Contrast Measure on the Ionosphere Data Set**

For the Ionosphere dataset:

* Compute the LpL\_pLp​-distance (p=1,2,∞p = 1, 2, \inftyp=1,2,∞) for all pairs of data points.
* Contrast measure is defined as the difference between the maximum and minimum pairwise distances divided by their sum:

C=max\_distance−min\_distancemax\_distance+min\_distanceC = \frac{\text{max\\_distance} - \text{min\\_distance}}{\text{max\\_distance} + \text{min\\_distance}}C=max\_distance+min\_distancemax\_distance−min\_distance​

Repeat the above steps for varying rrr, where rrr is the number of dimensions sampled (from 1 to full dimensionality).

**4. Match-Based Similarity, Cosine Similarity, and Jaccard Coefficient**

For the sets {A,B,C}\{A, B, C\}{A,B,C} and {A,C,D,E}\{A, C, D, E\}{A,C,D,E}:

**Match-Based Similarity:** Count the number of common elements. The intersection is {A,C}\{A, C\}{A,C}, so the similarity is 222.

**Cosine Similarity:** Represent the sets as binary vectors X=[1,1,1,0,0]X = [1, 1, 1, 0, 0]X=[1,1,1,0,0] and Y=[1,0,1,1,1]Y = [1, 0, 1, 1, 1]Y=[1,0,1,1,1]. Compute:

cos⁡(X,Y)=X⋅Y∣∣X∣∣ ∣∣Y∣∣=23⋅4=223≈0.577\cos(X, Y) = \frac{X \cdot Y}{||X|| \, ||Y||} = \frac{2}{\sqrt{3} \cdot \sqrt{4}} = \frac{2}{2\sqrt{3}} \approx 0.577cos(X,Y)=∣∣X∣∣∣∣Y∣∣X⋅Y​=3​⋅4​2​=23​2​≈0.577

**Jaccard Coefficient:**

J=∣Intersection∣∣Union∣=25=0.4J = \frac{|\text{Intersection}|}{|\text{Union}|} = \frac{2}{5} = 0.4J=∣Union∣∣Intersection∣​=52​=0.4

**5. Proof of Cosine Similarity Formula**

Given cosine(X,Y)=X⋅Y∣∣X∣∣ ∣∣Y∣∣\text{cosine}(X, Y) = \frac{X \cdot Y}{||X|| \, ||Y||}cosine(X,Y)=∣∣X∣∣∣∣Y∣∣X⋅Y​, expand the dot product and norms:

∣∣X∣∣22+∣∣Y∣∣22−∣∣X−Y∣∣22=2X⋅Y||X||\_2^2 + ||Y||\_2^2 - ||X - Y||\_2^2 = 2 X \cdot Y∣∣X∣∣22​+∣∣Y∣∣22​−∣∣X−Y∣∣22​=2X⋅Y

Divide by 2∣∣X∣∣∣∣Y∣∣2 ||X|| ||Y||2∣∣X∣∣∣∣Y∣∣ to derive the cosine similarity formula.

**6. Nearest Neighbor with Categorical Attributes (KDD Cup)**

Filter the dataset to retain only categorical attributes. Using the match measure, count exact matches between records. Using the inverse occurrence frequency, weight matches inversely proportional to the frequency of attribute values. Compute the nearest neighbor for each data point under both measures and count the cases where the nearest neighbor matches the class label.

**7. Nearest Neighbor with Quantitative Attributes (KDD Cup)**

Filter the dataset to retain only quantitative attributes. Compute pairwise LpL\_pLp​-norms (p=1,2,∞p = 1, 2, \inftyp=1,2,∞) between data points. For each data point, find its nearest neighbor and check if the class labels match. Repeat for all points and analyze the accuracy.

**2.6**

**14. Difference Between Precision of Measurement and Single/Double Precision**

**Precision of Measurement:**  
Precision refers to the consistency or repeatability of a measurement. High precision means that repeated measurements yield similar results, even if they are not accurate (close to the true value).

**Single and Double Precision in Computer Science:**

* **Single Precision:** Uses 32 bits to represent floating-point numbers. It typically provides 23 bits for the fraction (significand), 8 bits for the exponent, and 1 bit for the sign, offering about 7 decimal digits of precision.
* **Double Precision:** Uses 64 bits, with 52 bits for the fraction, 11 bits for the exponent, and 1 bit for the sign. This provides about 15–16 decimal digits of precision.  
  The choice between single and double precision affects memory usage, computational performance, and numerical accuracy.

**15. Advantages of Text Files Over Binary Files**

1. **Human Readability:**  
   Data stored in text files is easily readable and modifiable by humans using a simple text editor, which makes debugging and quick checks convenient.
2. **Platform Independence:**  
   Text files are not tied to a specific machine architecture or operating system, unlike binary files, which may have issues with byte order (endianness) or structure alignment.

**16. Difference Between Noise and Outliers**

**Noise:**  
Random variations in data that do not represent valid information. Noise can distort analysis and model predictions.

**Outliers:**  
Data points significantly different from the majority of the dataset, often due to unusual or extreme conditions. Outliers can provide valuable insights.

**Questions:**

(a) **Is noise ever interesting or desirable? Outliers?**  
Noise is typically undesirable. Outliers, however, can be interesting and represent rare but critical phenomena.

(b) **Can noise objects be outliers?**  
Yes, noisy data points can sometimes appear as outliers.

(c) **Are noise objects always outliers?**  
No, noise can also occur within typical values, depending on the dataset.

(d) **Are outliers always noise objects?**  
No, outliers might represent meaningful information rather than noise.

(e) **Can noise make a typical value into an unusual one, or vice versa?**  
Yes, noise can distort data distributions, making typical values appear unusual and vice versa.

**17. Issues with Algorithm 2.3 and Fix**

(a) **Problems with Duplicate Objects:**  
If there are duplicate objects, the algorithm might return the same data point multiple times for the kkk-nearest neighbors.

(b) **Fix:**  
Keep track of object indices and exclude the current object and duplicates from the kkk-nearest neighbor list.

**18. Proximity Measure for Elephants**

To compare or group elephants based on attributes (weight, height, tusk length, trunk length, ear area), the **Euclidean distance** is most appropriate, as the attributes are continuous and numerical. If attributes have different scales (e.g., weight vs. ear area), normalization is necessary to avoid dominance by larger-scale attributes.

**19. Sampling Schemes**

(a) **Stratified Sampling:**  
Proportionally select n×mi/mn \times m\_i / mn×mi​/m elements from each group, ensuring that each group is represented according to its size.

(b) **Random Sampling:**  
Randomly select nnn elements without considering group sizes, which may result in under- or over-representation of certain groups.

**20. Inverse Document Frequency Transformation**

(a) **Effect of Transformation:**

* If a term appears in one document: tfij=tfij×log⁡mtf\_{ij} = tf\_{ij} \times \log mtfij​=tfij​×logm.
* If a term appears in all documents: log⁡mdfi=0\log \frac{m}{df\_i} = 0logdfi​m​=0, so tfij=0tf\_{ij} = 0tfij​=0.

(b) **Purpose:**  
Reduces the weight of common terms and emphasizes terms that are unique to specific documents.

**21. Square Root Transformation**

(a) **Interval in Terms of xxx:**  
If x∗x^\*x∗ is linear to yyy in the interval (a,b)(a, b)(a,b), then xxx corresponds to (a2,b2)(a^2, b^2)(a2,b2).

(b) **Equation Relating yyy to xxx:**  
If x∗=xx^\* = \sqrt{x}x∗=x​, then y=c1x+c2y = c\_1 \sqrt{x} + c\_2y=c1​x​+c2​, where c1c\_1c1​ and c2c\_2c2​ are constants.

**22. Distances Between Points P1=(2,0)P\_1 = (2, 0)P1​=(2,0) and P2=(3,1)P\_2 = (3, 1)P2​=(3,1)**

(a) **Hamming Distance:**  
Count differing dimensions: 222.

(b) **Euclidean Distance:**

(3−2)2+(1−0)2=1+1=2≈1.41\sqrt{(3 - 2)^2 + (1 - 0)^2} = \sqrt{1 + 1} = \sqrt{2} \approx 1.41(3−2)2+(1−0)2​=1+1​=2​≈1.41

(c) **Supremum Distance:**  
Maximum absolute difference: max⁡(∣3−2∣,∣1−0∣)=1\max(|3 - 2|, |1 - 0|) = 1max(∣3−2∣,∣1−0∣)=1.

**23. Comparing Similarity and Distance Measures**

(a) **Hamming Distance:** 333 differences.  
**Jaccard Similarity:**

∣x∩y∣∣x∪y∣=37≈0.428\frac{|x \cap y|}{|x \cup y|} = \frac{3}{7} \approx 0.428∣x∪y∣∣x∩y∣​=73​≈0.428

(b) **Similarities:**

* Jaccard is closer to cosine similarity.
* Hamming is more akin to Simple Matching Coefficient.

(c) **For Genetic Makeup (Different Species):**  
Jaccard is more appropriate, as it measures shared features relative to the union of features.

(d) **For Genetic Makeup (Same Species):**  
Cosine similarity or correlation, as differences are minimal, and normalization helps.

**24. Similarity/Distance Measures**

Provide calculations for each distance/similarity metric for the specified vectors.

**25. Cosine and Correlation Observations**

(a) **Range of Cosine:** [0, 1] (for non-negative data).

(b) **Cosine Measure of 1:** Vectors are proportional but not necessarily identical.

(c) **Cosine and Correlation Relationship:** Both measure similarity, but correlation considers mean centering.

(d/e) **Graph Observations:** Analyze relationships between metrics using normalized/standardized vectors.

(f/g) **Derivations:** Relate cosine similarity and correlation to Euclidean distance under specific transformations.

**26. Proving that the Set Difference Metric d(A,B)d(A, B)d(A,B) Satisfies Metric Axioms**

The set difference metric is defined as:

d(A,B)=∣A−B∣+∣B−A∣d(A, B) = |A - B| + |B - A|d(A,B)=∣A−B∣+∣B−A∣

where ∣A−B∣|A - B|∣A−B∣ is the size of elements in AAA but not in BBB, and ∣B−A∣|B - A|∣B−A∣ is the size of elements in BBB but not in AAA. Let us verify the metric axioms:

1. **Non-negativity:**  
   By definition, ∣A−B∣≥0|A - B| \geq 0∣A−B∣≥0 and ∣B−A∣≥0|B - A| \geq 0∣B−A∣≥0. Therefore:

d(A,B)=∣A−B∣+∣B−A∣≥0d(A, B) = |A - B| + |B - A| \geq 0d(A,B)=∣A−B∣+∣B−A∣≥0

1. **Identity of Indiscernibles:**  
   If A=BA = BA=B, then A−B=∅A - B = \emptysetA−B=∅ and B−A=∅B - A = \emptysetB−A=∅, so:

d(A,B)=∣∅∣+∣∅∣=0d(A, B) = |\emptyset| + |\emptyset| = 0d(A,B)=∣∅∣+∣∅∣=0

Conversely, if d(A,B)=0d(A, B) = 0d(A,B)=0, then ∣A−B∣=0|A - B| = 0∣A−B∣=0 and ∣B−A∣=0|B - A| = 0∣B−A∣=0, implying A=BA = BA=B. Thus, the axiom holds.

1. **Symmetry:**  
   The set difference metric is symmetric because:

d(A,B)=∣A−B∣+∣B−A∣=∣B−A∣+∣A−B∣=d(B,A)d(A, B) = |A - B| + |B - A| = |B - A| + |A - B| = d(B, A)d(A,B)=∣A−B∣+∣B−A∣=∣B−A∣+∣A−B∣=d(B,A)

1. **Triangle Inequality:**  
   For any sets AAA, BBB, and CCC:

d(A,B)+d(B,C)=(∣A−B∣+∣B−A∣)+(∣B−C∣+∣C−B∣)d(A, B) + d(B, C) = (|A - B| + |B - A|) + (|B - C| + |C - B|)d(A,B)+d(B,C)=(∣A−B∣+∣B−A∣)+(∣B−C∣+∣C−B∣)

Each component of the set difference satisfies the triangle inequality individually due to the inclusion-exclusion principle, so:

d(A,B)+d(B,C)≥d(A,C)d(A, B) + d(B, C) \geq d(A, C)d(A,B)+d(B,C)≥d(A,C)

Thus, d(A,B)d(A, B)d(A,B) satisfies all the metric axioms.

**27. Mapping Correlation Values [−1,1][-1, 1][−1,1] to [0,1][0, 1][0,1]**

Correlation values can be mapped to [0,1][0, 1][0,1] using transformations depending on the application:

1. **Clustering Time Series:**  
   In clustering, distances or dissimilarities are often required. A simple linear transformation is:

Mapped value=1+Correlation2\text{Mapped value} = \frac{1 + \text{Correlation}}{2}Mapped value=21+Correlation​

This maps −1-1−1 to 000 and 111 to 111.

1. **Predicting Behavior of One Time Series Given Another:**  
   Here, the absolute value of correlation is often more meaningful, as it measures the strength of the relationship regardless of direction. Use:

Mapped value=∣Correlation∣\text{Mapped value} = |\text{Correlation}|Mapped value=∣Correlation∣

This transformation maps both −1-1−1 and 111 to 111, emphasizing the strength of the relationship.

**28. Transforming Similarity [0,1][0, 1][0,1] to Dissimilarity [0,∞)[0, \infty)[0,∞)**

Two possible transformations are:

1. **Inverse Transformation:**

Dissimilarity=1Similarity−1\text{Dissimilarity} = \frac{1}{\text{Similarity}} - 1Dissimilarity=Similarity1​−1

* + When similarity is close to 111, dissimilarity approaches 000.
  + When similarity is 000, dissimilarity becomes infinite.

1. **Logarithmic Transformation:**

Dissimilarity=−log⁡(Similarity)\text{Dissimilarity} = -\log(\text{Similarity})Dissimilarity=−log(Similarity)

* + This approach ensures dissimilarity is 000 when similarity is 111 and grows as similarity decreases.
  + The logarithmic scale can better handle small similarity values, preventing them from dominating the dissimilarity measure.

Let’s break these exercises into smaller steps and address them systematically.

**Exercise 3.5.1: Verifying Distance Measures on Nonnegative Integers**

To determine if the given functions are distance measures, we must check the four metric axioms:

1. Non-negativity: d(x,y)≥0d(x, y) \geq 0d(x,y)≥0.
2. Identity of Indiscernibles: d(x,y)=0  ⟺  x=yd(x, y) = 0 \iff x = yd(x,y)=0⟺x=y.
3. Symmetry: d(x,y)=d(y,x)d(x, y) = d(y, x)d(x,y)=d(y,x).
4. Triangle inequality: d(x,z)≤d(x,y)+d(y,z)d(x, z) \leq d(x, y) + d(y, z)d(x,z)≤d(x,y)+d(y,z).

**(a) max(x,y)\text{max}(x, y)max(x,y)**

* **Fails Identity of Indiscernibles:** max(x,y)=0\text{max}(x, y) = 0max(x,y)=0 is not satisfied unless x=y=0x = y = 0x=y=0. For any x≠yx \neq yx=y, the condition fails.
* Hence, this is **not a distance measure**.

**(b) diff(x,y)=∣x−y∣\text{diff}(x, y) = |x - y|diff(x,y)=∣x−y∣**

* **Non-negativity:** ∣x−y∣≥0|x - y| \geq 0∣x−y∣≥0.
* **Identity of Indiscernibles:** ∣x−y∣=0  ⟺  x=y|x - y| = 0 \iff x = y∣x−y∣=0⟺x=y.
* **Symmetry:** ∣x−y∣=∣y−x∣|x - y| = |y - x|∣x−y∣=∣y−x∣.
* **Triangle Inequality:** ∣x−z∣≤∣x−y∣+∣y−z∣|x - z| \leq |x - y| + |y - z|∣x−z∣≤∣x−y∣+∣y−z∣.
* Hence, diff(x,y)\text{diff}(x, y)diff(x,y) **is a distance measure**.

**(c) sum(x,y)=x+y\text{sum}(x, y) = x + ysum(x,y)=x+y**

* **Fails Identity of Indiscernibles:** sum(x,y)=0\text{sum}(x, y) = 0sum(x,y)=0 only when x=y=0x = y = 0x=y=0, but for x≠yx \neq yx=y, the condition fails.
* Hence, this is **not a distance measure**.

**Exercise 3.5.2: L1L\_1L1​ and L2L\_2L2​ Distances**

Given points (5,6,7)(5, 6, 7)(5,6,7) and (8,2,4)(8, 2, 4)(8,2,4):

1. **L1L\_1L1​ Distance (Manhattan Distance):**

L1=∣5−8∣+∣6−2∣+∣7−4∣=3+4+3=10L\_1 = |5 - 8| + |6 - 2| + |7 - 4| = 3 + 4 + 3 = 10L1​=∣5−8∣+∣6−2∣+∣7−4∣=3+4+3=10

1. **L2L\_2L2​ Distance (Euclidean Distance):**

L2=(5−8)2+(6−2)2+(7−4)2=9+16+9=34L\_2 = \sqrt{(5 - 8)^2 + (6 - 2)^2 + (7 - 4)^2} = \sqrt{9 + 16 + 9} = \sqrt{34}L2​=(5−8)2+(6−2)2+(7−4)2​=9+16+9​=34​

**Exercise 3.5.3: Proving Li>LjL\_i > L\_jLi​>Lj​ for i<ji < ji<j**

Let ppp and qqq be two points in nnn-dimensional space. The LiL\_iLi​ norm for i<ji < ji<j satisfies:

(∑k=1n∣pk−qk∣i)1/i≥(∑k=1n∣pk−qk∣j)1/j\left( \sum\_{k=1}^n |p\_k - q\_k|^i \right)^{1/i} \geq \left( \sum\_{k=1}^n |p\_k - q\_k|^j \right)^{1/j}(k=1∑n​∣pk​−qk​∣i)1/i≥(k=1∑n​∣pk​−qk​∣j)1/j

The proof involves Hölder’s inequality, which shows that the LiL\_iLi​-norm grows faster than the LjL\_jLj​-norm as i→ji \to ji→j.

**Exercise 3.5.4: Jaccard Distances**

For sets AAA and BBB, the Jaccard distance is defined as:

J(A,B)=1−∣A∩B∣∣A∪B∣J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}J(A,B)=1−∣A∪B∣∣A∩B∣​

1. **(a) A={1,2,3,4},B={2,3,4,5}A = \{1, 2, 3, 4\}, B = \{2, 3, 4, 5\}A={1,2,3,4},B={2,3,4,5}:**

J(A,B)=1−∣{2,3,4}∣∣{1,2,3,4,5}∣=1−35=0.4J(A, B) = 1 - \frac{|\{2, 3, 4\}|}{|\{1, 2, 3, 4, 5\}|} = 1 - \frac{3}{5} = 0.4J(A,B)=1−∣{1,2,3,4,5}∣∣{2,3,4}∣​=1−53​=0.4

1. **(b) A={1,2,3},B={4,5,6}A = \{1, 2, 3\}, B = \{4, 5, 6\}A={1,2,3},B={4,5,6}:**

J(A,B)=1−∣∅∣∣{1,2,3,4,5,6}∣=1−0=1J(A, B) = 1 - \frac{|\emptyset|}{|\{1, 2, 3, 4, 5, 6\}|} = 1 - 0 = 1J(A,B)=1−∣{1,2,3,4,5,6}∣∣∅∣​=1−0=1

**Exercise 3.5.5: Cosine of Angles Between Vectors**

Cosine similarity is defined as:

cos⁡(θ)=u⋅v∥u∥∥v∥\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}cos(θ)=∥u∥∥v∥u⋅v​

For each pair of vectors:

**(a) (3,−1,2)(3, -1, 2)(3,−1,2) and (−2,3,1)(-2, 3, 1)(−2,3,1):**

u⋅v=3(−2)+(−1)(3)+2(1)=−6−3+2=−7\mathbf{u} \cdot \mathbf{v} = 3(-2) + (-1)(3) + 2(1) = -6 - 3 + 2 = -7u⋅v=3(−2)+(−1)(3)+2(1)=−6−3+2=−7 ∥u∥=32+(−1)2+22=14,∥v∥=(−2)2+32+12=14\|\mathbf{u}\| = \sqrt{3^2 + (-1)^2 + 2^2} = \sqrt{14}, \quad \|\mathbf{v}\| = \sqrt{(-2)^2 + 3^2 + 1^2} = \sqrt{14}∥u∥=32+(−1)2+22​=14​,∥v∥=(−2)2+32+12​=14​ cos⁡(θ)=−714=−0.5\cos(\theta) = \frac{-7}{14} = -0.5cos(θ)=14−7​=−0.5

**(b), (c), and (d) can follow similarly.**

**Exercise 3.5.6: Proving Cosine Distance for Binary Vectors ≤ 90°**

For binary vectors u\mathbf{u}u and v\mathbf{v}v:

cos⁡(θ)=u⋅v∥u∥∥v∥\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}cos(θ)=∥u∥∥v∥u⋅v​

Since all components are 000 or 111, the dot product is nonnegative, and the norms are positive. Thus, cos⁡(θ)≥0\cos(\theta) \geq 0cos(θ)≥0, meaning θ≤90∘\theta \leq 90^\circθ≤90∘.

**Exercise 3.5.7: Edit Distances (Insertions & Deletions)**

**(a) "abcdef" and "bdaefc":**

Transform "abcdef"→"bdaefc"\text{"abcdef"} \to \text{"bdaefc"}"abcdef"→"bdaefc" using insertions and deletions. Steps can be calculated by aligning the characters.