Lecture 1

Question 1: What is the primary goal of a recommendation system?

A. Classification
B. Clustering
C. Prediction
D. Personalized suggestion

Question 2: In collaborative filtering, what does the "user-item matrix" represent?

A. Users' demographic information
B. Items' characteristics
C. Users' preferences for items
D. Ratings given by users to items

Question 3: Which evaluation metric measures the accuracy of predicted ratings in recommendation systems?

A. Precision
B. Recall
C. Mean Absolute Error (MAE)
D. F1 score

Question 4: In the context of collaborative filtering, what is the purpose of the user-based approach?

A. To identify similar items
B. To find similar users
C. To recommend popular items
D. To address the cold start problem

Question 5: Which algorithm is commonly used for generating recommendations in an item-item collaborative filtering system?

A. k-Nearest Neighbors (k-NN)
B. Decision Trees
C. Apriori algorithm
D. Naive Bayes

Lecture 2

Question 1: What is the primary objective of clustering algorithms in data mining?

A. Classification
B. Prediction
C. Grouping similar data points
D. Outlier detection

Question 2: What is the main drawback of the k-means clustering algorithm?

A. Sensitivity to initial cluster centroids
B. Inability to handle non-linear clusters
C. Lack of scalability
D. Dependency on data distribution

Question 3: How does the DBSCAN algorithm handle outliers in the dataset?

A. Assigns outliers to the nearest cluster
B. Removes outliers from the dataset
C. Labels outliers as noise
D. Forms separate clusters for outliers

Question 4: Which clustering algorithm is particularly effective for identifying clusters with varying shapes and densities?

A. K-means
B. DBSCAN
C. Hierarchical clustering
D. None of above

Question 5: Given the 1D data points {1, 2, 3, 4, 6, 7, 8, 9, 10}, if k-means clustering is run with k=2 and initial cluster centers at C1=2 and C2=8 using Manhattan distance. Which points will be in the cluster with center 8 in the first iteration?

A. {4,6,7,8,9,10}
B. {6,7,8,9,10}
C. {7,8,9,10}
D. {8,9,10}

Lecture 3

Question 1: What does covariance measure in the context of data mining?

A. Strength of linear relationship between two variables
B. Independence between two variables
C. Magnitude of individual variables
D. Mean absolute difference between two variables

Question 2: In eigenvalue decomposition of a matrix, what is the significance of the eigenvectors?

A. They represent the scaling factors
B. They represent the rotation angles
C. They define the transformation matrix
D. They define the principal components

Question 3: In PCA, what is the primary goal of transforming the original variables into principal components?

A. To reduce the dimensionality of the data.
B. To increase the interpretability of the data.
C. To increase the correlation between variables.
D. To make the data non-linear.

Question 4: In the context of SVD, what is the significance of singular values?

A. They represent the magnitude of eigenvalues
B. They indicate the strength of covariance between variables
C. They define the principal components
D. They quantify the importance of individual dimensions

Question 5: What is a key advantage of SVD over eigenvalue decomposition?

A. SVD is faster
B. SVD is more accurate
C. SVD works only for square matrices
D. SVD can be applied to non-square matrices

Lecture 4

Question 1: In a self-organizing map (SOM), which neuron's weights are adjusted during training?

A. Only the winning neuron
B. The winning neuron and its neighbors
C. All neurons in the map
D. A randomly selected subset of neurons

Question 2: Given a trained SOM and a new input vector [0.2, 0.7], which of the following neurons would be selected as the winner using Euclidean distance?

A. Neuron 1: [0.1, 0.5]
B. Neuron 2: [0.3, 0.8]
C. Neuron 3: [0.6, 0.4]
D. Neuron 4: [0.8, 0.2]

Question 3: Multidimensional scaling (MDS) aims to ensure that points close together in the embedded space are also close in the original space. What term quantifies the reconstruction error?

A. Accuracy
B. Stress
C. Entropy
D. Inertia: the distance between each data point and its centroid

Question 4: What is the main advantage of using t-Distributed Stochastic Neighbor Embedding (t-SNE) over traditional methods like Principal Component Analysis (PCA)?

A. Faster computation
B. Better preservation of local structures in the data
C. Reduced risk of overfitting
D. Improved interpretability

Lecture 5

Question 1: What is the primary purpose of encoding in the context of searching and ranking in data mining?

A. Compression of data
B. Conversion of data into a format
C. Creation of an index for efficient retrieval
D. Encryption of data

Question 2: In the context of searching and ranking in data mining, what is indexing used for?

A. Sorting data for presentation
B. Efficient retrieval of relevant information
C. Encoding data for storage
D. Encrypting sensitive information

Question 3: What does the term "inverted index" refer to in the context of searching and ranking?

A. An index that is sorted in reverse order
B. An index where the positions of terms in documents are recorded
C. A type of encoding technique
D. A ranking algorithm

Question 4: Which of the following is an essential component of the matching process in searching and ranking?

A. Clustering
B. Indexing
C. Ranking
D. Similarity measure

Question 5: How does tf-idf (Term Frequency-Inverse Document Frequency) contribute to the ranking of documents in information retrieval?

A. It measures the frequency of a term in a document
B. It penalizes common terms and emphasizes those with a high frequency in a given document
C. It encrypts the document content
D. It compresses the document for efficient storage

Question 6: How is TF-IDF calculated for a term w in a document d? f, F are the frequency of w, the frequency of all the words appeared in d. N, n are the total number of documents, and the number of documents where the term w appears.

A. F/f * log (N/n)
B. f/F * log (N/n)
C. F/f * log (n/N)
D. f/F * log (n/N)

Lecture 6

Question 1: What is the key assumption made by the Naive Bayes classifier?

A. Features are correlated
B. Features are independent given the class
C. Features follow a Gaussian distribution
D. Features have equal variance

Question 2: What is the purpose of the "assumed" parameter in the Naive Bayes classifier for text data?

A. It is the assumed value for the prior probability
B. It is the assumed value for the posterior probability
C. It is the assumed value for the likelihood probability when a word is not present
D. It is the assumed value for the feature independence assumption

Question 3: In Fisher's method, the combined p-value is calculated by:

A. Adding the individual p-values
B. Multiplying the individual p-values
C. Taking the average of the individual p-values
D. Using a chi-square statistic on the individual p-values

Question 4: Suppose we have a binary classification problem with two classes: spam (c1) and not spam (c2). The prior probabilities are P(c1) = 0.4 and P(c2) = 0.6. For a given document d, the likelihood probabilities are P(d|c1) = 0.7 and P(d|c2) = 0.3. If we use the Naive Bayes method, is this document d a spam or not?

A. Yes
B. No

Lecture 7

Question 1: Which of the following is NOT a characteristic of decision trees?

A. Tree-like structure with nodes representing feature tests
B. Branches represent outcomes of feature tests
C. Leaf nodes represent class labels
D. Nodes must split data into equal-sized partitions

Question 2: What is the primary goal when constructing a decision tree for classification?

A. To maximize the depth of the tree
B. To create a balanced tree
C. To minimize the number of leaf nodes
D. To maximize the purity of leaf nodes

Question 3: If a node contains 10 examples, 6 of class A and 4 of class B, what is the Gini impurity of this node?

A. 0.48
B. 0.52
C. 0.64
D. 0.36

Question 4: In the context of decision trees, what does the term "greedy" refer to?

A. The algorithm's appetite for more data
B. The algorithm's tendency to overfit
C. The algorithm's ability to find the globally optimal solution
D. The algorithm's strategy of making the locally optimal choice at each step

Question 5: Which ensemble method builds multiple decision trees on random subsets of features?

A. Bagging
B. Boosting
C. Random forests
D. None of the above

Lecture 8

Question 1: What is the key idea behind the k-NN algorithm?

A. Use training records directly to predict the class label of test cases by considering their neighbor correlations
B. Train an explicit model using the training data
C. Use majority voting to determine the class label
D. a and c

Question 2: Given the following data points and their class labels: (1, 1, Class A), (2, 2, Class B), (3, 3, Class A), (4, 4, Class B), what would be the class label assigned to the unknown point (2, 3) using the k-NN algorithm with k=3 and Euclidean distance?

A. Class A
B. Class B
C. Tie between Class A and Class B
D. Not enough information to determine the class label

Question 3: What is a potential issue when using the k-NN algorithm with different attribute scales?

A. The distance measures may be dominated by one of the attributes
B. The algorithm may not converge
C. The algorithm may overfit the training data
D. The algorithm may be biased towards the minority class

Question 4: What is the purpose of using Locality-Sensitive Hashing (LSH) in the context of k-NN?

A. To improve the accuracy of the algorithm
B. To reduce the computational complexity of the algorithm
C. To handle missing data in the dataset
D. To perform dimensionality reduction

Question 5: Which of the following is NOT a disadvantage of the k-NN algorithm?

A. It is sensitive to the choice of k
B. It requires storing all the training data
C. It is not suitable for large datasets due to computational complexity
D. It cannot handle non-linear decision boundaries

Lecture 9

Question 1: Which of the following statements best describes a transaction in market basket analysis?

A. A set of all items sold in a store
B. An individual item or article in a basket
C. A set of items purchased together in a single shopping basket
D. The process of generating association rules

Question 2: What is the purpose of the Apriori algorithm in market basket analysis?

A. To find all possible association rules
B. To find all frequent itemsets that satisfy the minimum support threshold
C. To generate high-confidence rules from frequent itemsets
D. Both b) and c)

Question 3: The anti-monotone property of support in the Apriori algorithm states that:

A. If an itemset is frequent, then all of its supersets must also be frequent
B. If an itemset is infrequent, then all of its subsets must also be infrequent
C. If an itemset is frequent, then all of its subsets must also be frequent
D. If an itemset is infrequent, then some of its supersets might be frequent

Question 4: In a dataset with 5000 transactions, the itemset {X, Y} appears in 1200 transactions, and the itemset {X} appears in 2000 transactions. What is the confidence of the association rule {X} → {Y}?

A. 0.24
B. 0.4
C. 0.6
D. 2.4

Lecture 10

Question 1: What is the underlying principle behind Latent Semantic Analysis (LSA)?

A. Identifying co-occurrence patterns of words in text data
B. Applying supervised learning on labeled text data
C. Extracting topics using clustering techniques
D. Finding semantic features with dimensionality reduction on the term-document matrix

Question 2: In LSA, what does the Bag of Words (BoW) model provide?

A. Semantic word embeddings
B. A sparse term-document matrix
C. A clustering of documents
D. A probabilistic topic distribution

Question 3: Why might some LSA weights be difficult to interpret semantically?

A. They are constrained to be binary (0 or 1)
B. They can take unconstrained, including negative values
C. They ignore co-occurrence information
D. They require labeled topic data

Question 4: What is the effect of applying truncated SVD in LSA?

A. It increases the size of the term-document matrix
B. It removes the most important dimensions
C. It reduces noise and captures the most important patterns
D. It converts words into probability distributions

Lecture 11

Question 1: Topic modeling can be conceptualized as:

A. A supervised classification method for text data
B. A process of uncovering the underlying themes or topics in a collection of documents
C. A method for extracting keywords from documents
D. A dimensionality reduction technique for text data

Question 2: What is the main advantage of probabilistic topic models like PLSA and LDA over deterministic models like NMF?

A. They can handle uncertainties and complexities in textual data more effectively
B. They are more computationally efficient
C. They can automatically determine the number of topics
D. They produce more interpretable topics

Question 3: In the context of Latent Dirichlet Allocation (LDA), what is the role of the Dirichlet prior distribution?

A. It is used to model the distribution of words given a topic
B. It is used to model the distribution of topics given a document
C. It is used to model the distribution of documents given a corpus
D. Both (a) and (b)

Question 4: Which of the following inference techniques is commonly used for parameter estimation in PLSA?

A. Expectation-Maximization (EM) algorithm
B. random sampling
C. k-means clustering
D. Both (a) and (b)

Lecture 12

Question 1: What is the main idea behind using a Gaussian Mixture Model (GMM) for outlier detection?

A. Data points with low probability density under the fitted GMM are considered outliers
B.) Data points that belong to low-weight Gaussian components are marked as outliers
C. Data points far away from the mean of any Gaussian component are outliers
D. Both a) and c)

Question 2: Which of the following is NOT a parameter that needs to be learned when fitting a GMM?

A. Mean of each Gaussian component
B. Covariance matrix of each Gaussian component
C. Weight/Mixing proportion of each Gaussian component
D. Number of Gaussian components in the mixture

Question 3: Which of the following is a potential advantage of using DBSCAN over Gaussian mixture models for outlier detection?

A. DBSCAN can handle data with varying densities better
B. DBSCAN does not require setting any parameters like neighborhood radius
C. DBSCAN is more computationally efficient for high-dimensional data
D. DBSCAN can automatically determine the optimal number of clusters

Question 4: Which statement best describes a key difference between GMMs and DBSCAN for outlier detection?

A. GMMs are probabilistic models, while DBSCAN relies on density estimation
B. GMMs can only identify outliers, while DBSCAN can find clusters and outliers
C. GMMs assume Gaussian distributions, while DBSCAN makes no distribution assumptions
D. GMMs can handle outliers automatically, while DBSCAN requires pruning outliers

Reference/Answer:

Note: The following answers are provided solely for reference purposes, and their accuracy cannot be guaranteed:

- Lecture 1: D, D, C, B, A
- Lecture 2: C, A, C, B, B
- Lecture 3: A, D, A, D, D
- Lecture 4: B, B, B, B
- Lecture 5: B, B, B, D, B, B
- Lecture 6: B, C, D, A
- Lecture 7: D, D, A, D, C
- Lecture 8: D, A, A, B, D
- Lecture 9: C, D, C, C
- Lecture 10: D, B, B, C
- Lecture 11: B, A, D, A
- Lecture 12: D, D, A, A