cap5771sp25
April 9, 2025

Exam 2 version b

Question 1. Based on the lecture slides on Pattern Mining (patternMining-part1.pdf), slide 16, consider the following scenario: A retail store is using pattern mining to analyze customer transactions. They find that the itemset {toothpaste, chocolate, orange juice} is infrequent. Based on the Apriori Principle and the Downward Closure Property, which of the following is NOT a correct conclusion to draw from this finding?

(a) `Some subsets of \{toothpaste, chocolate, orange juice\} may still be frequent.`

(b) `If \{toothpaste, chocolate\} and \{chocolate, orange juice\} are both frequent, it does not necessarily mean \{toothpaste, chocolate, orange juice\} will be frequent.`

(c) `The itemset \{toothpaste, chocolate, orange juice, gum\} might still be frequent.`

(d) `The individual items 'toothpaste', 'chocolate', and 'orange juice' could still be frequent on their own.`

(e) `Any supersets of \{toothpaste, chocolate, orange juice\} must also be infrequent.`

Question 2. According to the Lecture on Pattern Mining slide - 19 and 20, Consider a dataset where the Apriori algorithm is applied with a minimum support threshold of 50 percent. After two iterations, the algorithm fails to generate any $(k+1)$ (consider $K \geq 3$)-itemsets despite multiple 2-itemsets being frequent. Which of the following is the most likely reason for this failure?

(a) `The dataset contains only pairwise associations, meaning no three-item combinations appear frequently enough to meet the threshold.`

(b) `The Downward Closure Property has mistakenly eliminated valid itemsets due to an error in candidate generation.`

(c) `The dataset contains many high-confidence association rules, but those rules do not necessarily correspond to frequent itemsets.`

(d) `The support threshold is too low, causing Apriori to stop early before discovering larger patterns.`

(e) `The dataset is too sparse, meaning that even frequent itemsets in early iterations fail to combine into larger frequent patterns.`

Question 3. According to the lecture on clustering (slide 5), which of the following is a key requirement for a good clustering method?

(a) `High inter-class similarity and low intra-class similarity.`

(b) `Equal similarity within and between clusters.`

(c) `Maximizing the number of clusters regardless of data structure.`

(d) `Using only Euclidean distance for all data types.`

(e) `High intra-class similarity and low inter-class similarity.`

Question 4. According to the lecture on Classification II (Slide: "Feature Selection & Feature Engineering"), which of the following statements accurately describes the differences between feature selection and feature engineering?

(a) `Feature selection and feature engineering are interchangeable terms that both refer to the automatic selection of the best features using deep learning models.`

(b) `Feature selection only works with structured data, while feature engineering is only applicable to unstructured data such as text or images.`

(c) `Feature selection is the process of manually constructing new features using domain knowledge, while feature engineering involves automatically selecting the most relevant features based on statistical techniques.`

(d) `Feature selection involves identifying the most effective subset of features from a set of initial features, while feature engineering involves creating new, more informative features from the existing ones, often using domain knowledge or deep learning methods.`

(e) `Feature selection involves removing redundant and irrelevant features, while feature engineering focuses on identifying and eliminating missing or incomplete data.`

Question 5. According to the lecture on Clustering (slide 14), what is one key advantage of the K-Medoids clustering method over the K-Means clustering method?

(a) `It requires less computational complexity.`

(b) `It is less sensitive to outliers and noise.`

(c) `It is faster for large datasets.`

(d) `It does not require specifying the number of clusters in advance.`

(e) `It works only for numerical data.`

Question 6. According to the lecture on text retrieval and extraction, what is the key difference between IO and IOB encoding schemes in sequence labeling?

(a) `IO encoding is used exclusively in POS tagging, and IOB in sentiment analysis.`

(b) `IO requires an additional End tag which is not present in IOB.`

(c) `IO encoding can capture nested entities, while IOB cannot.`

(d) `IO and IOB are the same, the only difference is in the implementation.`

(e) `IOB adds a Beginning tag to differentiate the start of entities, which IO does not.`

Question 7. According to the lecture on Clustering (Slide: "Assessing the Suitability of Clustering"), when evaluating whether data has inherent grouping structure, what is the challenge in determining clusterability?

(a) `The challenge is that clustering methods always assume data has inherent groupings, making it unnecessary to evaluate clustering tendency.`

(b) `The challenge lies in the many different definitions of clusters (e.g., partitioning, hierarchical, density-based, and graph-based) and the difficulty in defining an appropriate null model for the data.`

(c) `The difficulty arises from the fact that clustering methods require pre-defined cluster labels, which are often unavailable in real-world datasets.`

(d) `Assessing clusterability is straightforward because there are standardized methods that always give clear results, regardless of data type.`

(e) `Clusterability is easy to assess since all clustering methods, regardless of their approach, produce the same results when applied to any dataset.`

Question 8. From slide 16 of pattern-mining-part1, what is the key purpose of pruning in the Apriori algorithm?

(a) `To remove transactions from the database before mining frequent itemsets.`

(b) `To filter out itemsets with low confidence before rule generation.`

(c) `To remove itemsets that contain infrequent subsets and reduce the number of candidate itemsets.`

(d) `To remove redundant association rules after rule mining.`

(e) To reduce the dataset size by eliminating all one-item itemsets.

Question 9. According to the lecture notes on text analysis: When preparing text data for a machine learning classification model, which of the following steps specifically weights terms according to their frequency in a document relative to their frequency across all documents, ensuring that commonly used words are given less influence than rare but important words?

(a) Converting all text to lowercase

(b) Frequency counts and computing TF-IDF term weights

(c) Word (term) extraction: easy

(d) Stopwords removal

(e) Stemming

Question 10. According to the lecture on Clustering (Slide 48, page 48), what is the primary advantage of the Silhouette Coefficient over the Dunn Index as an intrinsic clustering evaluation metric?

(a) The Silhouette Coefficient has a fixed range of values between 0 and 1, unlike the unbounded Dunn Index.

(b) The Silhouette Coefficient requires less computational complexity than calculating the Dunn Index for large datasets.

(c) The Silhouette Coefficient works only with categorical data while the Dunn Index requires numerical data.

(d) The Silhouette Coefficient evaluates individual data points rather than using extreme distances, making it less sensitive to outliers than the Dunn Index.

(e) The Silhouette Coefficient requires ground truth labels while the Dunn Index is an unsupervised metric.

Question 11. What is a key difference in how the Boolean model and the Vector Space Model represent documents and queries in Information Retrieval?

(a) Both models represent documents as "bags of words," but the Boolean model incorporates term order, while the Vector Space Model does not.

(b) The Vector Space Model considers the frequency of terms within documents and across the collection, whereas the Boolean model primarily focuses on the presence or absence of terms.

(c) The Boolean model uses cosine similarity to rank retrieved documents, while the Vector Space Model relies on exact logical matching of query terms.

(d) The Boolean model represents documents as vectors with weighted terms, while the Vector Space Model treats them as sets of words.

(e) The Boolean model is well-suited for handling natural language queries, while the Vector Space Model requires strict Boolean operators.

Question 12. According to the lecture on Text Retrieval and Extraction (Slide: "TF-IDF Term Weighting Scheme"), which of the following formulas correctly represents the calculation of the TF-IDF term weight?

(a) The TF-IDF term weight is the term frequency (TF) divided by the inverse document frequency (IDF), representing the importance of a term relative to a document's context.

(b) The TF-IDF term weight is calculated by multiplying the term frequency (TF) by the number of documents containing the term, and then adding the total number of documents in the dataset.

(c) The TF-IDF term weight is only influenced by the term frequency (TF) and does not take into account the inverse document frequency (IDF).

(d) The TF-IDF term weight is calculated as the product of the term frequency (TF) and inverse document frequency (IDF), where TF is the frequency of the term in a document, and IDF is the logarithm of the total number of documents divided by the number of documents containing the term.

(e) The TF-IDF term weight is calculated as the sum of term frequency (TF) and inverse document frequency (IDF), where TF is the frequency of the term in a document, and IDF is the number of documents containing the term.

Question 13. According to the Pattern Mining Part 1 pdf (slide 16), the Apriori Algorithm is a key method in frequent pattern mining that leverages the Apriori Principle to improve efficiency. In what way does the Apriori Principle help reduce computational complexity during the mining process?

(a) It ensures that all itemsets generated during the mining process are frequent, reducing the need for multiple scans of the database.

(b) It limits the number of frequent itemsets by setting a maximum threshold on the number of items considered in each transaction.

(c) It avoids the need for a minimum support threshold by dynamically selecting itemsets based on their occurrence patterns.

(d) It guarantees that if an itemset is frequent, all of its supersets will also be frequent, eliminating the need for further pruning.

(e) It states that if an itemset is frequent, all of its subsets must also be frequent, allowing the pruning of supersets of infrequent itemsets.

Question 14. According to a lecture on Pattern Mining, CAP5771 - Introduction to Data Science, University of Florida, 2025 (Slide 10, Closed Patterns and Max-Patterns), what is the key distinction between closed patterns and max-patterns in frequent itemset mining?

(a) A max-pattern is a frequent itemset that has no superset with the same support, whereas a closed pattern is a frequent itemset that has no frequent superset.

(b) Closed patterns are always subsets of max-patterns in any given dataset.

(c) Closed patterns and max-patterns are the same and can be used interchangeably in pattern mining.

(d) A closed pattern is a frequent itemset that has no superset with the same support, whereas a max-pattern is a frequent itemset that has no frequent superset.

(e) A closed pattern is a pattern that only appears once in the dataset, while a max-pattern appears multiple times.

Question 15. According to the text retrieval and extraction lecture, what is a benefit of stemming?

(a) It reduces indexing size by a substantial amount.

(b) It does not help match similar words.

(c) It does not improve recall.

(d) It would be used to reduce like and line because they share many letters.

(e) It decreases the effectiveness of information retrieval.

Question 16. According to the lecture on Pattern Mining Part 1, slide 10, which of the following best defines the structural and mathematical properties of a closed pattern within a transaction database?

(a) A closed pattern is a subset of a maximal pattern, where the subset retains only the most statistically significant co-occurring items while discarding those that do not contribute to higher confidence in association rules.

(b) A closed pattern is defined as the largest possible frequent itemset within a given dataset, ensuring that all of its subsets are also frequent and can be directly derived from its occurrence statistics.

(c) A frequent itemset qualifies as a closed pattern when its absolute support exceeds a dataset-specific entropy threshold, thereby distinguishing itself as an essential pattern for association rule mining.

(d) A frequent itemset \(X\) is considered a closed pattern if and only if there exists no proper superset of \(X\) with an identical support count, thereby ensuring that all significant item relationships are captured without unnecessary redundancy.

(e) A closed pattern is any frequent itemset where the sum of its subset supports is equal to its own support count, indicating a perfect correlation between the items it contains.

Question 17. According to the lecture based on Text Retrieval and Extraction.pdf, which of the following is NOT a key step in training a machine learning-based Named Entity Recognition (NER) system?

(a) Designing feature extractors appropriate to the text and classes.

(b) Manually tagging entities in the test set during evaluation.

(c) Labeling each token for its entity class or other (O).

(d) Training a sequence classifier to predict the labels from the data.

(e) Collecting a set of representative training documents.

Question 18. According to the lecture On Pattern Mining, given the following dataset of transactions and a minsup threshold of 50 percent {{Bread, Milk, Eggs}, {Bread, Milk}, {Eggs, All Purpose Flour, Chocolate Chips}, {Milk, Ground Coffee}} which items would NOT be frequent?

(a) All Purpose Flour, Chocolate Chips, Ground Coffee.

(b) Milk

(c) Ground Coffee

(d) All Purpose Flour, Ground Coffee.

(e) Bread, Milk, Eggs.

Question 19. According to the lecture and "Text Retrieval and Extraction.pdf", in the Vector Space Model, what is the primary role of representing documents and queries as vectors?

(a) To ensure that documents have a fixed number of words.

(b) To visualize documents in a two-dimensional plot.

(c) To compress documents for efficient storage.

(d) To measure the similarity between documents and queries using mathematical operations like cosine similarity.

(e) `To eliminate the need for tokenization and preprocessing.`

Question 20. According to the lecture, which of the following statements about agglomerative clustering and its similarity measures is most accurate?

(a) `The Manhattan distance is the preferred similarity measure for all agglomerative clustering methods due to its robustness to high-dimensional data.`

(b) `The complete link method is more sensitive to outliers than the single link method, potentially leading to more compact and tightly bound clusters.`

(c) `The average link method always produces the same clustering results as the centroid link method, regardless of data distribution.`

(d) `The single link method always produces elongated, chain-like clusters regardless of the dataset's characteristics.`

(e) `The centroid link method is computationally more efficient than other methods because it only considers the center points of clusters.`

Question 21. According to the lecture in pattern mining, what is a potential pitfall of relying solely on high confidence values when selecting strong association rules?

(a) `Rules with high confidence have a greater likelihood of being spurious, making them unreliable for decision-making.`

(b) `Confidence scores can predict future customer purchases with 100\% accuracy.`

(c) `Confidence does not take into account the baseline probability of the consequent, potentially leading to misleading conclusions.`

(d) `High confidence always guarantees a strong and meaningful association between items in a dataset.`

(e) `Confidence values are directly proportional to support values, ensuring that high-confidence rules are also frequent.`

Question 22. According to lecture (Text Retrieval and Extraction, Day 3), why is the inverse document frequency (IDF) used in TF-IDF weighting?

(a) `To reduce the influence of stopwords.`

(b) `To increase the weight of common words across documents.`

(c) `To normalize word count by document length.`

(d) `To assign more weight to frequently occurring terms.`

(e) `To remove punctuation and symbols from the index.`

Question 23. Lecture on pattern mining (slide 11-13) Given a transaction database containing the following two transactions:
T1: {b1, b2, b3, b4} T2: {b2, b3, b4, b5}
If the minimum support threshold (min_sup) is set to 2, which of the following itemsets is a frequent itemset?

(a) `\{b4, b5\}`

(b) `\{b1, b5\}`

(c) `\{b3, b4, b5\}`

(d) `\{b1, b3\}`

(e) `\{b2, b3\}`

Question 24. According to lecture (slide 7 of Classification-part-I.pdf), what is a key stopping condition for splitting nodes in decision tree induction?

(a) `When all samples for a given node belong to the same class.`

(b) `When the tree reaches a predefined depth.`

(c) `When all attributes have been used at least once.`

(d) `When the tree contains more than 100 nodes.`

(e) `When accuracy on the training set exceeds 95\%.`

Question 25. According to the text retrieval and extraction lectures (slides 13- 17), which of the following query examples would produce poor retrieval results when using the Boolean Model, due to lack of term frequency consideration?

(a) `(‘Python’ AND ‘Programming’) AND (‘Web’ OR ‘Scraping’)`

(b) `("Deep" OR "Learning") AND ("Neural" OR "Networks")`

(c) `("Artificial Intelligence" AND "Machine Learning") AND (NOT "Supervised")`

(d) `(‘COVID-19’ AND ‘Vaccination’ AND ‘Symptoms’)`

(e) `((‘Data’ AND ‘Science’) AND (‘Analysis’ OR ‘Visualization’))`

Question 26. According to the lecture on Classification II (Slide: "Gradient Descent and Optimization"), which of the following statements correctly describes the process and mechanics of gradient descent?

(a) `In Gradient Descent, the gradient is ignored, and the algorithm simply moves in a predetermined direction regardless of the function's slope to minimize the error.`

(b) `Gradient Descent is an iterative optimization algorithm that minimizes a function by moving in the direction of the negative gradient, adjusting the step size at each iteration until the gradient becomes zero, indicating a local minimum.`

(c) `Gradient Descent works by randomly selecting the direction to move and adjusting the step size based on the overall error at each iteration, aiming to reach a global maximum.`

(d) `Gradient Descent only works for convex functions and cannot find the minimum of functions with multiple local minima.`

(e) `The algorithm continues indefinitely until the function reaches its global minimum, regardless of the gradient direction or step size.`

Question 27. According to the lecture on pattern mining (patternMining-part1.pdf, slide-10), which of the following is considered a lossless compression method for frequent itemsets?

(a) `Closed Patterns.`

(b) `Max-Patterns.`

(c) `Association Rules.`

(d) `Apriori Algorithm.`

(e) `K-Means Clustering.`

Question 28. According to the text retrieval and extraction lectures (slides 29-31), which of the following sentences would make the task of Named Entity Recognition quite difficult?

(a) `Someone had a sandwich for lunch yesterday.`

(b) `I had a great time visited Jordan last year.`

(c) `Dr. Someone Vance presented her research at a university.`

(d) `Someone debated the economic philosophy of open-source software with Orwellian fervor.`

(e) `A famous landmark is in Paris, France.`

Question 29. You ran an association rule mining algorithm on transaction data for a supermarket and discovered the following rule: {itemA, itemB} → {itemC} The store manager says: 'Since itemA and itemB cause customers to buy itemC, we should increase the price of itemB to maximize profits.' According to the Pattern Mining Part 1 lecture (page 6), which of the following best evaluates the manager's reasoning?

(a) `The conclusion is valid because increasing the price of itemC will likely not affect the sale of itemA and itemB.`

(b) `The conclusion is correct because high-confidence rules always indicate consumer intent.`

(c) `The conclusion is flawed because association rules only indicate co-occurrence, not causation.`

(d) `The conclusion is valid because association rules imply a strong causal relationship between itemA, itemB, and itemC.`

(e) `The conclusion is flawed because itemC is an independent item and is not necessarily purchased with itemA and itemB.`

Question 30. According to the lecture on "Inverted Indexes," what is the primary purpose of using an inverted index in text retrieval systems?

(a) `To efficiently map each term to the list of documents in which it occurs, enabling fast query processing.`

(b) `To track the sequence in which users access documents during search sessions.`

(c) `To rank documents by their publication date for recency-based retrieval.`

(d) `To arrange documents in alphabetical order for easier lookup by title.`

(e) `To compress text data and reduce storage requirements.`

Question 31. According to the Pattern Mining slides, which of the following is the first step in the Apriori algorithm?

(a) `Generating candidate itemsets of size \(k+1\) from frequent \(k\)-itemsets.`

(b) `Removing duplicate transactions from the dataset.`

(c) `Calculating the correlation coefficient between items.`

(d) `Scanning the database to identify frequent 1-itemsets.`

(e) `Applying clustering techniques to detect outliers.`

Question 32. Based on the lecture on clustering, imagine this scenario: A data analyst is working on a dataset containing customer information and wants to segment customers into meaningful groups based on their purchasing behavior. Which of the following clustering methods is the best for this task?

(a) `Utilizing hierarchical methods to build a tree-like structure of nested clusters, allowing flexible exploration of customer segments.`

(b) `Applying partitioning methods to divide the dataset into \(k\) distinct clusters, ensuring each data point belongs to exactly one cluster.`

(c) `Implementing density methods to detect clusters of varying shapes and sizes by identifying dense regions in the data.`

(d) `Using basic concepts of cluster analysis to understand different clustering techniques before selecting an appropriate method.`

(e) `Using density-based and grid-based methods to improve clustering efficiency by organizing data into grids before detecting dense regions.`

Question 33. According to the lecture on Clustering (Slide: "DBSCAN and Sensitivity to Parameter Settings"), what is a key challenge when using DBSCAN for clustering, especially in datasets with varying density?

(a) `DBSCAN requires an extremely high value for minPts to function correctly, which can lead to excessive noise detection in dense datasets.`

(b) `DBSCAN automatically adjusts its parameters during the clustering process to ensure that all clusters are of equal size, regardless of density.`

(c) `DBSCAN's clustering results are always consistent, as it uses hierarchical clustering to determine the best number of clusters regardless of parameter settings.`

(d) `DBSCAN is sensitive to the setting of parameters such as epsilon and minPts, which can significantly affect the number and quality of clusters, especially when density varies across regions of the data.`

(e) `DBSCAN is not sensitive to parameter settings and can handle varying densities without requiring adjustments to epsilon or minPts.`

Question 34. According to Example 4.7 in the pattern mining section of the textbook, suppose you analyze some transaction data and find that of the 10,000 transactions analyzed, the data shows that: 1. 6000 of the customer transactions included computer games 2. 7500 included videos 3. 4000 included both computer games and video You discover the rule: buys(X, "computer games") $\rightarrow$ buys(X, "videos")with support = 40 percent and confidence = 66 percent. Let the minimum support be 30percent and minimum confidence be 60 percent. What is your next step?

(a) `Report the rule with a cautionary note and take no further action.`

(b) `The rule might be misleading, so investigate further using additional metrics.`

(c) `Adjust the thresholds to force a higher confidence relative to the baseline.`

(d) `Discard the rule since its confidence is below the overall rate.`

(e) `Accept the rule because it meets preset thresholds.`

Question 35. Which of the following sentences would most likely confuse a Named Entity Recognition (NER) system due to coreference and ambiguous entity boundaries?

(a) `After discussing with the client, she emailed the documents to her boss.`

(b) `A major company recently acquired a startup based in San Francisco.`

(c) `Tesla opened a new Gigafactory in Berlin, expanding its European operations.`

(d) `The decision of the company's CEO was widely debated in tech circles.`

(e) `The company's new product launch was covered by all major news channels.`