cap5771sp25
April 9, 2025

Exam 2 version c

Question 1. Which of the following best explains the distinction between feature selection and feature engineering in the context of building a predictive model?

(a) `Feature selection focuses on identifying the most relevant and non-redundant features from the existing dataset, while feature engineering involves transforming or creating new features to improve model performance.`

(b) `Feature engineering is only applicable when using deep learning, while feature selection is used for traditional machine learning models.`

(c) `Feature selection creates new variables by combining raw inputs, while feature engineering removes irrelevant data to simplify the model.`

(d) `Feature selection and feature engineering are identical and both refer to reducing the size of the dataset by removing noise.`

Question 2. Use the information on slide page 9 from the Clustering lecture slides to answer the following question. A supermarket wants to optimize its store layout by clustering customer movement data. If shoppers tend to linger around specific sections rather than moving evenly throughout the store, what clustering method would be preferable?

(a) `Density-based clustering method`

(b) `Distance-based methods`

(c) `Probabilistic and generative models`

(d) `K-Medoids Clustering`

Question 3. According to the lecture, which of the following statements about Kernel K-Means and K-Medoids is incorrect?

(a) `K-Medoids is less sensitive to outliers compared to K-Means.`

(b) `Kernel K-Means is computationally less intensive than standard K-Means.`

(c) `Both Kernel K-Means and K-Medoids require specifying the number of clusters in advance.`

(d) `K-Medoids uses medoids, which are actual data points, as cluster representatives.`

(e) `Kernel K-Means can handle non-linearly separable clusters by mapping data to a high-dimensional space.`

Question 4. According to the Pattern Mining lecture slides, which of the following choices shows an example of the downward closure property?

(a) `The itemset \{A\} is frequent and the itemset \{B\} is frequent, therefore the itemset \{A, B\} must be frequent.`

(b) `The itemset \{A, B, C\} is frequent and so are the itemsets \{A\}, \{B\}, and \{C\}.`

(c) `The itemset \{A, B, C\} is frequent and so are the itemsets \{A, B\}, \{B, C\}, \{A, C\}, \{A\}, \{B\}, and \{C\}.`

(d) `The itemset \{A\} is frequent and the itemset \{A, B, C\} is not frequent.`

(e) `The support of the itemset \{A, B\} is greater than the minimum support, but no other itemset has support greater than the minimum support.`

Question 5. According to the lecture on Text Retrieval and Extraction.pdf (slide 17), which text processing technique reduces word variations like "users" to "use" and significantly improves recall while reducing indexing size?

(a) `Named Entity Recognition`

(b) `POS Tagging`

(c) `Stopword Removal`

(d) `TF-IDF Weighting`

(e) `Stemming`

Question 6. According to the Pattern Mining Part 1 pdf (slide 6), a frequent itemset is defined based on a support threshold. What determines whether an itemset is considered frequent?

(a) `If it contains at least one item that appears frequently in the database.`

(b) `If its absolute support count is greater than the average support count of all itemsets in the dataset.`

(c) `If it has more items than any other itemset in the database.`

(d) `If it appears in at least 10\% of the transactions in the dataset.`

(e) `If its relative support is greater than or equal to a predefined minimum support threshold.`

Question 7. According to Classification part I.pdf (slide 32) and the associated lecture, which of the following is an appropriate list of subset sizes when using 5-fold cross-validation on data that has a total of 60 tuples?

(a) `30, 15, 8, 4, 3`

(b) `20, 20, 10, 5, 5`

(c) `22, 17, 12, 7, 2`

(d) `56, 1, 1, 1, 1`

(e) `12, 12, 12, 12, 12`

Question 8. According to the lecture on Classification II (Slide: "Backpropagation Algorithm"), which of the following correctly describes the process of backpropagation in neural networks?

(a) `Backpropagation modifies the network weights by applying the activation function during the forward pass and uses gradient descent during the backward pass to adjust weights for minimizing errors.`

(b) `Backpropagation is a method for calculating the gradients of weights in convolutional neural networks and is not applicable to feedforward neural networks.`

(c) `Backpropagation involves a one-time forward pass through the network, where the weights are fixed and no further adjustments are made. It is used to compute the final output without any error correction.`

(d) `In backpropagation, weights are initialized with large random numbers, and the error is propagated only in the forward direction to refine the model's predictions.`

(e) `Backpropagation is an iterative process where the network's weights are adjusted by calculating the error at the output layer and propagating it backward through the hidden layers to minimize the mean squared error between the predicted output and the actual target values.`

Question 9. According to the lecture, which of the following best explains how the TF-IDF weighting scheme enhances retrieval performance in the vector space model, especially in distinguishing between documents with overlapping vocabulary?

(a) `TF-IDF transforms every document into a binary feature vector, facilitating fast exact-match lookups and enabling Boolean-style retrieval with ranked output.`

(b) `TF-IDF, by relying solely on inverse document frequency, ensures that rare terms are always more important regardless of their presence in the query.`

(c) By combining normalized term frequency with a logarithmically scaled inverse document frequency, TF-IDF ensures that terms both frequent in a document and rare in the corpus contribute more significantly to the cosine similarity score.

(d) TF-IDF assigns lower weights to rare terms and boosts the influence of frequent terms across the corpus, ensuring that general topics dominate document similarity computation.

(e) The primary role of TF-IDF is to eliminate all stopwords and perform stemming, which directly influences the cosine similarity between query and document vectors.

Question 10. According to the lecture on ensemble methods, which ensemble method involves training multiple models on bootstrapped samples of the data and then averaging their predictions?

(a) Stacking

(b) AdaBoost

(c) Boosting

(d) Bagging

(e) Gradient Boosting

Question 11. From the Clustering Lecture PDF (slide 23), which of the following is NOT a benefit of the dendrogram visualization that can be created when using hierarchical clustering techniques?

(a) Seeing how clusters merge or divide at different levels of similarity allows you to determine the number of true clusters instead of needing to input a cluster number.

(b) Dendrograms display the statistical significance of each cluster formation, providing p-values for the validity of each merging step.

(c) Dendrograms allow you to get a better grasp of the relationship between the clusters in the dataset.

(d) Merging clusters at a later point is made easier by seeing how closely certain clusters are related.

(e) Dendrograms increase interpretability by allowing you to visually inspect the cluster creation process.

Question 12. According to the lecture on information retrieval, in the context of Information Retrieval (IR), what is the primary role of an inverted index in supporting efficient document retrieval?

(a) `It parses queries into logical expressions and evaluates them using rule-based natural language understanding systems.`

(b) `It maps each term in the vocabulary to a list of documents in which the term appears, allowing for fast lookup of matching documents for a given query.`

(c) `It stores the frequency of every word in a document to facilitate named entity recognition using IOB tagging.`

(d) `It maintains a binary matrix where each row represents a document and each column represents a term, primarily for document classification tasks.`

(e) `It compresses the entire document collection into a single vector space to enable cosine similarity computations for every pair of documents.`

Question 13. 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) `((‘Data’ AND ‘Science’) AND (‘Analysis’ OR ‘Visualization’))`

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

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

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

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

Question 14. According to the clustering lecture notes, which of the following is NOT a weakness of the K-Means clustering method?

(a) `High computational complexity for large datasets`

(b) `Requires specifying the number of clusters (k) in advance`

(c) `Struggles with non-convex cluster shapes`

(d) `Assumes clusters are spherical and equally sized`

(e) `Sensitive to noisy data and outliers`

Question 15. Based on the discussion in class on Slide 9 in the patternMining-part1.pdf, what is the key difference between Support and Confidence in the context of Association rules?

(a) `Support count is always a percentage, while confidence is always a whole number.`

(b) `Support count measures the frequency of an itemset, while confidence measures the strength of implication between itemsets.`

(c) `Support count applies to single items, while confidence only applies to pairs of items.`

(d) `Support count decreases as itemset size increases, while confidence always increases.`

(e) `Support count is calculated after mining, while confidence is used to prune candidates during mining.`

Question 16. According to the lecture materials on Text Retrieval, which of the following best describes the primary role of the Inverted Index in a search engine system?

(a) `It ranks documents based on PageRank scores before user queries are processed.`

(b) `It stores the full content of each document in a compressed format for retrieval.`

(c) `It evaluates term frequency-inverse document frequency (TF-IDF) weights for all documents.`

(d) `It maps terms to the list of documents where they appear, enabling efficient full-text search.`

(e) `It detects named entities in documents and links them to a knowledge base.`

Question 17. According to the lecture on classification, what is a defining characteristic of a Bayesian Belief Network?

(a) `It requires all input features to be linearly independent.`

(b) `It learns from data by propagating error signals backward through layers.`

(c) `It represents probabilistic relationships using a directed acyclic graph (DAG).`

(d) `It classifies data using maximum margin hyperplanes.`

(e) `It automatically reduces irrelevant features using L1 regularization.`

Question 18. According to the lecture on Classificatio I (Slide: "Handling Rare Cases with Laplacian Correction"), what is the purpose of Laplacian correction in Naive Bayes prediction, and how does it impact the probability estimates?

(a) `Laplacian correction is used to prevent zero probabilities by adding a small value (e.g., 1) to each count, ensuring that no probability is zero, even for rare events.`

(b) `Laplacian correction increases the probability of rare events by decreasing the count of frequent events, thus leading to more balanced predictions.`

(c) `Laplacian correction is used to adjust the decision threshold of the model, allowing it to classify more instances as belonging to the rare class.`

(d) `Laplacian correction removes rare events from the dataset to improve the model's predictive power and efficiency.`

(e) `Laplacian correction is used to adjust for class imbalance by changing the prior probabilities of each class, making the model less sensitive to the majority class.`

Question 19. 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) `Clusterability is easy to assess since all clustering methods, regardless of their approach, produce the same results when applied to any dataset.`

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

Question 20. 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 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.`

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

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

(e) `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.`

Question 21. According to the lecture on "Inverted Indexes," what is the primary purpose of using an inverted index in text retrieval systems?

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

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

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

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

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

Question 22. According to the lecture on Classification I (Slide: "Accuracy, Error Rate, Sensitivity, and Specificity"), which metric is most appropriate to evaluate the performance of a classifier when dealing with an imbalanced dataset where one class (e.g., fraud) is much rarer than the other?

(a) `Sensitivity (recall) is the most appropriate metric, as it focuses on the true positive recognition rate for the minority class, which is typically the class of interest in imbalanced datasets.`

(b) `Error rate is the best metric because it accounts for both false positives and false negatives equally, making it ideal for imbalanced classes.`

(c) None of the above metrics are suitable for imbalanced datasets, and a different evaluation method is required.

(d) Accuracy is the most appropriate metric because it measures the overall proportion of correctly classified instances, regardless of class distribution.

(e) Specificity is the most appropriate metric, as it evaluates the true negative recognition rate, which is more relevant for imbalanced datasets.

Question 23. According to the lecture regarding Named Entity Recognition (Text Retrieval and Extraction.pdf, Slide 51), what is the main purpose of IOB encoding in sequence labeling tasks like NER?

(a) It helps differentiate between the beginning and continuation of named entities, improving the accuracy of entity boundary detection.

(b) It clusters documents based on shared entities using cosine similarity.

(c) It identifies parts of speech like nouns and verbs to improve sentence segmentation .

(d) It is used to extract relations between entities by using logical rule chaining.

(e) It transforms raw documents into bag-of-words vectors for classification tasks.

Question 24. What is a key characteristic that distinguishes Information Extraction (IE) from Information Retrieval (IR)?

(a) IE is mainly used for web search and e-discovery, whereas IR is applied to more specific domains like extracting information from medical literature.

(b) IE evaluates its performance using precision and recall at the document level, while IR uses metrics like F1 score per entity.

(c) IE focuses on basic text processing tasks such as tokenization and stemming, while IR deals with higher-level tasks like sentiment analysis.

(d) IE primarily uses statistical models like TF-IDF, while IR relies more on rule-based systems and regular expressions.

(e) IE aims to produce a structured representation of specific information from text, whereas IR focuses on retrieving relevant documents in response to a query.

Question 25. According to the lecture, which of the following statements about agglomerative clustering and its similarity measures is most accurate?

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

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

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

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

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

Question 26. According to the Pattern Mining slides, which of the following is the first step in the Apriori algorithm?

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

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

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

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

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

Question 27. 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) `Confidence values are directly proportional to support values, ensuring that high-confidence rules are also frequent.`

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

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

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

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

Question 28. Which of the following clustering quality metrics is based on intra-cluster compactness and inter-cluster separation?

(a) `Kullback-Leibler Divergence`

(b) `Entropy`

(c) `Jaccard Coefficient`

(d) `Silhouette Coefficient`

(e) `Hopkins Statistic`

Question 29. According to the Clustering lecture based on Slide 24 (Clustering.pdf), why might one prefer the average-link method over single-link and complete-link clustering in AGNES?

(a) `It completely avoids the chaining effect.`

(b) `It reduces the need for a dissimilarity matrix.`

(c) `It provides faster convergence compared to other methods.`

(d) `It minimizes the impact of outliers while preventing elongated clusters.`

(e) `It balances compactness and connectivity better than single-link and complete-link methods.`

Question 30. 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) `\{b1, b3\}`

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

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

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

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

Question 31. According to the lecture on pattern mining, which statement best describes the downward closure property in frequent itemset mining?

(a) `Any superset of a frequent itemset must also be frequent.`

(b) `Any subset of a frequent itemset must also be frequent.`

(c) `An infrequent itemset eventually becomes frequent when more data is collected.`

(d) `The minimum support threshold can never change once it has been set.`

(e) `It's a key principle used in decision trees to determine the best split at each node.`

Question 32. According to the Pattern Mining Part 1 pdf (slide 9), association rules describe relationships between items in transactions. What does a strong association rule require?

(a) `High frequency and at least one item with maximum support.`

(b) `A high lift value greater than 2.`

(c) `A minimum number of transactions containing both antecedent and consequent.`

(d) `High confidence and high correlation.`

(e) `High support and high confidence.`

Question 33. According to the Text Retrieval and Extraction.pdf, which of the following best describes the primary goal of Information Extraction systems?

(a) `To retrieve a ranked list of documents that are relevant to a search query.`

(b) `To perform basic text processing steps like tokenization, stemming, and removal of stop words.`

(c) `To find and understand limited parts of texts and produce a structured representation of relevant information such as entities and relations.`

(d) `To evaluate the performance of a search engine based on measures like precision and recall.`

(e) `To classify the overall topic of a given document.`

Question 34. According to the lecture on text retrieval and extraction, what is the key difference between IO and IOB encoding schemes in sequence labeling?

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

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

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

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

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

Question 35. According to the lecture on Text Retrieval and Extraction (Slide: "Encoding Classes for Sequence Labeling"), what is the primary difference between the IO and IOB encoding schemes?

(a) `In the IOB encoding, 'B' is used to mark the beginning of an entity and 'O' is used for non-entity words, while in IO encoding, 'B' and 'I' are not used, and only 'O' is used for all tokens.`

(b) `IO encoding is more complex than IOB encoding because it requires additional labels for entity relationships, whereas IOB encoding only uses 'B' and 'I' tags.`

(c) `There is no difference between IO and IOB encoding schemes; both use the same labels and are interchangeable in sequence labeling tasks.`

(d) `In IO encoding, entities are labeled starting from the inside ('I') rather than from the beginning ('B') as in IOB encoding, and the 'O' tag is not used in IO encoding.`

(e) `In the IO encoding, 'I' is used to mark entity tokens and 'O' is used for non-entity tokens, while in IOB encoding, 'B' marks the beginning of an entity, 'I' marks the continuation of the entity, and 'O' is used for non-entity tokens.`