cap5771sp25
April 9, 2025

Exam 2 version a

Question 1. 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) There is no difference between IO and IOB encoding schemes; both use the same labels and are interchangeable in sequence labeling tasks.

(b) 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.

(c) 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.

(d) 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.

(e) 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.

Question 2. According to the lecture on clustering (slide 5), which of the following is a key requirement for a good clustering method?

(a) Maximizing the number of clusters regardless of data structure.

(b) Equal similarity within and between clusters.

(c) High intra-class similarity and low inter-class similarity.

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

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

Question 3. 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) A famous landmark is in Paris, France.

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

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

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

(e) Someone had a sandwich for lunch yesterday.

Question 4. According to lecture slides on Clustering (pgno.13), why is K-Means clustering method sensitive to noisy data and outliers?

(a) Outliers can significantly shift the cluster centroids, leading to poor clustering results.

(b) K-Means assigns each data point to multiple clusters at once.

(c) K-Means uses medoids instead of centroids for cluster representation.

(d) K-Means automatically removes noisy data before clustering.

Question 5. Which of the following best describes the primary goal of cluster analysis?

(a) To determine the probability distribution of a dataset using statistical models.

(b) To reduce data dimensionality by transforming features into a lower-dimensional space.

(c) To assign predefined labels to data points based on a training dataset.

(d) To sequentially label data points using a rule-based decision-making process.

(e) To partition data points into groups that are as similar as possible within clusters and as dissimilar as possible between clusters.

Question 6. According to the lecture on Pattern Mining Part 1, slide 16, which key principle underlies the Apriori Algorithm, and how does it help in reducing computational complexity?

(a) Support Redistribution: It dynamically adjusts the support threshold for itemsets based on their occurrence frequency.

(b) Confidence Boosting: Itemsets with higher confidence scores are prioritized for rule generation, reducing unnecessary computations.

(c) Downward Closure Property: If an itemset is infrequent, all its supersets are also infrequent, allowing pruning of the search space.

(d) Dynamic Hashing: The algorithm uses adaptive hashing techniques to store frequent itemsets, minimizing memory usage.

(e) Recursive Partitioning: It divides the dataset into smaller independent partitions, reducing the number of scans needed.

Question 7. According to the Clustering lecture (Slide 6), what is the main reason Euclidean distance may be inappropriate for high-dimensional data in clustering tasks?

(a) `Because in high dimensions, distances between points become increasingly similar, reducing the effectiveness of separation.`

(b) `Because Euclidean distance always produces non-convex clusters, even in low-dimensional space.`

(c) `Because Euclidean distance in high dimensions causes your computer to become self-aware.`

(d) `Because Euclidean distance requires labeled data to compute similarity.`

(e) `Because Euclidean distance ignores data normalization, which is only addressed by cosine similarity.`

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 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 works only with categorical data while the Dunn Index requires numerical data.`

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

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

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

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

Question 10. According to the clustering.pdf slide 24, what statement best describes the differences between agglomerative (AGNES) and divisive (DIANA) clustering?

(a) `While both clustering methods start with one node and merge until there is only one cluster, they are inverses of each other: AGNES utilizes a bottom-up approach, while DIANA goes from the top down.`

(b) `DIANA separates clusters starting from one big cluster until there are many nodes, AGNES starts with a singular node and merges until there is one large cluster.`

(c) `AGNES is implemented using statistical analysis packages, while DIANA is implemented using different similarity measures.`

(d) `AGNES separates clusters starting from one big cluster until there are many nodes, DIANA starts with a singular node and merges until there is one large cluster.`

(e) `The clustering methods are the same and operate in the exact same fashion.`

Question 11. According to lecture (Text Retrieval and Extraction, Day 3), why is the inverse document frequency (IDF) used in TF-IDF weighting?

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

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

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

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

(e) To assign more weight to frequently occurring terms.

Question 12. What is the key difference between absolute support and relative support for an itemset?

(a) Absolute support requires multiple database scans, while relative support requires only one scan.

(b) Absolute support is used for closed patterns, while relative support is used for max-patterns.

(c) Absolute support is calculated during the Apriori algorithm, while relative support is calculated during FP-Growth.

(d) Absolute support is the count of transactions containing the itemset, while relative support is the percentage of transactions containing it.

(e) Absolute support applies to single items, while relative support applies to itemsets with multiple items.

Question 13. 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) Gradient Descent only works for convex functions and cannot find the minimum of functions with multiple local minima.

(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) 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.

(d) 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.

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

Question 14. According to the Text Retrieval and Extraction lecture (Slide 32: Information Extraction Steps from the Text Retrieval and Extraction PDF), Information Extraction is defined as which combination of processes?

(a) `Segmentation, Classification, Association, and Clustering.`

(b) `Tokenization, Prediction, Association, and Clustering.`

(c) `Segmentation, Translation, Association, and Clustering.`

(d) `Parsing, Classification, Matching, and Clustering.`

(e) `Segmentation, Classification, Association, and Aggregation.`

Question 15. According to the Pattern Mining Part 1 pdf (slide 19), the FP-Growth algorithm is an alternative to Apriori for frequent pattern mining. How does FP-Growth differ from Apriori in its approach?

(a) `FP-Growth uses a tree-based structure to store transactions and avoids candidate generation.`

(b) `FP-Growth directly generates frequent itemsets without requiring multiple database scans.`

(c) `FP-Growth evaluates all possible itemset combinations to determine frequent patterns.`

(d) `FP-Growth dynamically adjusts the minimum support threshold based on item occurrence.`

(e) `FP-Growth only finds maximal frequent patterns, ignoring smaller frequent itemsets.`

Question 16. According to the lecture on Classification Part 1, slide 30, which of the following best describes the relationship between training error, test error, and model complexity?

(a) `Overfitting occurs when both training error and test error are high.`

(b) `Training error remains constant regardless of model complexity, while test error fluctuates randomly.`

(c) `As model complexity increases, training error tends to decrease, while test error first decreases and then increases due to overfitting.`

(d) `Test error is consistently lower than training error because the model generalizes better to unseen data.`

(e) `Training error and test error always decrease together as model complexity increases.`

Question 17. According to the lecture on pattern mining, what is the primary drawback of using a brute-force algorithm for mining frequent itemsets?

(a) `Its tendency to select only the most trivial itemsets, omitting more complex patterns.`

(b) `Its dependency on pre-sorted data, which is rarely available in real-world applications.`

(c) `Its exponential time complexity, as it evaluates all possible candidate itemsets, making it computationally prohibitive.`

(d) `Its exclusive focus on relative support, thereby ignoring absolute support measures .`

(e) `Its failure to generate any candidate itemsets, which results in missing frequent patterns.`

Question 18. 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) `Apriori Algorithm.`

(c) `Max-Patterns.`

(d) `Association Rules.`

(e) `K-Means Clustering.`

Question 19. 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 avoids the need for a minimum support threshold by dynamically selecting itemsets based on their occurrence patterns.`

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

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

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

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

Question 20. According to the lecture, which of the following statements about Kernel K-Means and K-Medoids is incorrect?

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

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

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

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

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

Question 21. As mentioned in slide 14 of pattern mining part1 pdf, why is the brute-force approach to frequent pattern mining computationally expensive?

(a) It requires solving a Rubik's cube before every computation, which delays processing indefinitely.

(b) The brute-force approach requires sorting all transactions before processing, increasing time complexity.

(c) It relies on deep learning models to generate frequent itemsets, making it slow and hardware-dependent.

(d) It only considers a single-item relationship at a time, missing multi-item dependencies.

(e) The number of possible itemsets grows exponentially with the number of items, making it infeasible for large datasets.

Question 22. 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) `Clusterability is easy to assess since all clustering methods, regardless of their approach, produce the same results when applied to any dataset.`

(c) `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.`

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

(e) `The difficulty arises from the fact that clustering methods require pre-defined cluster labels, which are often unavailable in real-world 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 identifies parts of speech like nouns and verbs to improve sentence segmentation .`

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

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

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

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

Question 24. According to slide 16 from our lecture, ”Text Retrieval and Extraction,” which of the following statements is true regarding the IDF (Inverse Document Frequency) component of TF-IDF weighting?

(a) `It normalizes term frequencies within a single document.`

(b) `It reduces the weight of terms that appear frequently across the document collection.`

(c) `It removes all common words from the dataset.`

(d) `It directly measures the cosine similarity between documents.`

(e) It increases the weight of terms that appear frequently in many documents.

Question 25. 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) Manually tagging entities in the test set during evaluation.

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

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

(d) Collecting a set of representative training documents.

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

Question 26. 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) High support and high confidence.

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

(e) High confidence and high correlation.

Question 27. According to the Pattern Mining slides, which of the following is the first step in the Apriori algorithm?

(a) Calculating the correlation coefficient between items.

(b) Removing duplicate transactions from the dataset.

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

(d) Applying clustering techniques to detect outliers.

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

Question 28. According to the 'Text Retrieval and Extraction' lecture slides (slides 19-21), why is precision generally considered a more reliable evaluation metric than recall in the context of web search engines?

(a) `Because recall directly measures the relevance of the top results displayed to the user.`

(b) `Because users tend to be more interested in seeing a few relevant results at the top rather than all relevant documents available.`

(c) `Because precision penalizes documents with high TF-IDF scores.`

(d) `Because search engines rarely return more than 10 documents per query, making recall obsolete.`

(e) `Because web crawlers can accurately determine the total number of relevant documents for any query.`

Question 29. According to slides 25 of the lecture, 'Text Retrieval and Extraction', which of the following statements best describes the relationship between precision and recall for evaluating an information retrieval system?

(a) `Neither precision nor recall can be used to evaluate a retrieval system meaningfully.`

(b) `A higher precision score always implies a higher recall score, and vice versa.`

(c) `Precision measures how many of the relevant documents were successfully retrieved, while recall measures the proportion of retrieved documents that are relevant.`

(d) `Precision measures the proportion of retrieved documents that are relevant, while recall measures how many of the relevant documents were successfully retrieved.`

(e) `Precision and recall both measure only the quantity of documents retrieved, ignoring their relevance.`

Question 30. According to the lecture on Clustering (Slide: "High-Quality Clusters and Influencing Factors"), if someone is assessing several clustering methods to group behavioral data from engineering students, what are the key elements that determine whether a clustering method will produce high-quality clusters?

(a) `A similarity function that always uses cosine distance to ensure consistency across domains.`

(b) `Any method that groups items based on their order in the dataset rather than their features.`

(c) `A clustering algorithm that is efficient.`

(d) High intra-class similarity, low inter-class similarity, a well-chosen similarity measure, thoughtful implementation, and the ability to uncover hidden patterns.

(e) High inter-class similarity and a focus on speed of implementation over interpretability.

Question 31. 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 reduces the need for a dissimilarity matrix.

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

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

(d) It completely avoids the chaining effect.

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

Question 32. 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 measure the similarity between documents and queries using mathematical operations like cosine similarity.

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

(c) To eliminate the need for tokenization and preprocessing.

(d) To compress documents for efficient storage.

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

Question 33. According to lecture (Text Retrieval and Extraction, Day 3), suppose you're building a search engine for a legal document archive. After applying TF-IDF weighting, you notice that common legal terms like 'plaintiff', 'court', and 'case' have high term frequencies but are lowering your retrieval precision. What would be the most effective adjustment to your retrieval pipeline based on the principles discussed?

(a) Increase the weight of query terms using one-hot encoding rather than vector-based methods like TF-IDF.

(b) Apply Inverse Document Frequency to reduce the influence of ubiquitous legal terms and enhance document differentiation.

(c) Install a legal dictionary plugin to translate terms into simpler language for the model to understand.

(d) `Replace TF-IDF with raw term frequency to increase the influence of common legal vocabulary on rankings.`

(e) `Remove all legal-specific terms from the corpus to prevent overfitting and improve model generalization.`

Question 34. According to lecture (slide 31), what defines a 'core point' in the DBSCAN clustering algorithm?

(a) `A point that belongs to a cluster only if it is directly density-reachable from another point.`

(b) `A point that lies in a sparse region and does not contribute to cluster formation.`

(c) `A point that is not within the epsilon-neighborhood of any other point.`

(d) `A point that lies on the border between two clusters, serving as a potential cluster connector.`

(e) `A point with at least MinPts in its epsilon-neighborhood, making its surrounding region dense.`

Question 35. According to the lecture on clustering, what distinguishes the DBSCAN clustering algorithm from K-Means?

(a) `DBSCAN can discover clusters of arbitrary shape and is robust to noise.`

(b) `DBSCAN assigns every point to exactly one cluster.`

(c) `DBSCAN builds clusters by continuously splitting a parent cluster in a top-down fashion.`

(d) `DBSCAN requires specifying the number of clusters beforehand.`

(e) `DBSCAN only works with numerical data in low-dimensional space.`