cap5771sp25
April 9, 2025

**Exam 2 version d**

**Question 1.** 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) `Training a sequence classifier to predict the labels from the data.`

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

(d) `Collecting a set of representative training documents.`

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

**Question 2.** 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) `A famous landmark is in Paris, France.`

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

**Question 3.** According to the lecture on classification, what is the main drawback of using accuracy as the sole metric for evaluating classification models?

(a) `Accuracy automatically adjusts for class imbalance.`

(b) `Accuracy directly measures how well a model generalizes to new data.`

(c) `Accuracy can be misleading when dealing with imbalanced datasets, as it may favor the majority class.`

(d) `Accuracy is the only metric needed to evaluate model performance.`

(e) `High accuracy always indicates a well-performing model.`

**Question 4.** According to the lecture on Low-Level Information Extraction (Text Retrieval and Extraction, CAP 5771), which of the following best describes the benefit of using regular expressions in low-level information extraction systems?

(a) `They determine the grammatical or semantic roles (e.g., subject, predicate) of all tokens in a document.`

(b) `They offer a simple and efficient way to extract specific patterns such as phone numbers, dates, or HTML tags from semi-structured or unstructured text.`

(c) `They enable automated query expansion techniques based on real-time user feedback in search engines.`

(d) `They completely replace the need for advanced sequence labeling models like CRFs or LSTMs.`

(e) `They are primarily used to compute TF-IDF weights for named entities across documents.`

**Question 5.** 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 increases the weight of terms that appear frequently in many documents.`

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

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

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

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

**Question 6.** 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 maps terms to the list of documents where they appear, enabling efficient full-text search.`

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

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

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

**Question 7.** According to the clustering lecture and textbook, what is the most logical step to improve clustering performance?

(a) Apply dimensionality reduction techniques (e.g., PCA) to reduce noise and improve cluster separation.

(b) Use a fixed number of clusters without considering data distribution to maintain consistency.

(c) Increase the number of clusters arbitrarily to balance the distribution of data points.

(d) Ignore the small clusters as they are irrelevant to the analysis.

(e) Assign outliers to the nearest cluster to ensure all data points are included in the analysis.

**Question 8.** According to the lecture on pattern mining, which statement best describes the downward closure property in frequent itemset mining?

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

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

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

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

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

**Question 9.** 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 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.

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

(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) `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.`

**Question 10.** According to the lecture "Classification part II.pdf" regarding the principles of multi-layer neural networks, what is the primary role of the hidden layer?

(a) `To cycle back information to previous layers for better weight adjustments.`

(b) `To directly pass the input values to the output layer without modification.`

(c) `To randomly adjust weights without learning from training samples.`

(d) `To transform weighted inputs non-linearly, allowing the network to learn complex patterns.`

(e) `To store intermediate predictions that are directly used as final outputs.`

**Question 11.** 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) `Distance-based methods`

(b) `Density-based clustering method`

(c) `Probabilistic and generative models`

(d) `K-Medoids Clustering`

**Question 12.** 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 Euclidean distance ignores data normalization, which is only addressed by cosine similarity.`

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

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

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

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

**Question 13.** 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 applies to single items, while confidence only applies to pairs of items.`

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

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

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

**Question 14.** According to the lecture on classification part I (slide 33), how does the ROC curve help in classifier evaluation?

(a) `It is used only for regression problems.`

(b) `It calculates the exact error rate of the model.`

(c) `It measures the true positive rate against the true negative rate.`

(d) `It helps you choose the best classifier based on which one has the most stylish logo.`

(e) `It visualizes the trade-off between the true positive rate and false positive rate.`

**Question 15.** 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 its relative support is greater than or equal to a predefined minimum support threshold.`

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

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

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

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

**Question 16.** 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 Boolean model uses cosine similarity to rank retrieved documents, while the Vector Space Model relies on exact logical matching of query terms.

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

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

(e) 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.

**Question 17.** 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) Training error and test error always decrease together as model complexity increases.

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

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

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

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

**Question 18.** What is the key difference between absolute support and relative support for an itemset?

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

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

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

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

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

**Question 19.** 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 is used to extract relations between entities by using logical rule chaining.`

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

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

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

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

**Question 20.** According to the Classification Part 1 slides, which of the following best describes the purpose of a test set for classification model building?

(a) `To help select the most appropriate attributes for creating the model.`

(b) `To tune the parameters of the model after it has been deployed for use.`

(c) `To provide the data used for the initial creation of the classification model.`

(d) `Used to estimate the accuracy of the constructed model on unseen data.`

(e) `To find and remove noisy data from the training set.`

**Question 21.** According to the Lecture on Pattern Mining (Slide 20), how does the Apriori algorithm generate candidate itemsets in each iteration?

(a) `By using a tree-based structure to directly extract frequent itemsets.`

(b) `By considering only itemsets with the highest support in each iteration.`

(c) `By selecting random subsets of transactions and checking for frequent itemsets.`

(d) `By scanning the database once and identifying all frequent patterns immediately.`

(e) `By joining frequent itemsets of size \(k\) to form candidate itemsets of size \(k+1\).`

**Question 22.** According to the text retrieval and extraction lecture, what is a benefit of stemming?

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

(b) `It does not help match similar words.`

(c) `It does not improve recall.`

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

(e) `It decreases the effectiveness of information retrieval.`

**Question 23.** 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 eliminate the need for tokenization and preprocessing.`

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

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

(e) `To compress documents for efficient storage.`

**Question 24.** From the lecture Classification Part I.pdf, Page 7, in decision tree induction, what is the purpose of the information gain heuristic?

(a) `To randomly select attributes for splitting.`

(b) `To measure the reduction in entropy after splitting data on an attribute.`

(c) `To prune overfitting branches post-construction.`

(d) `To assign class labels to leaf nodes.`

(e) `To normalize feature values before constructing the tree.`

**Question 25.** 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 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.

(b) 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.

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

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

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

**Question 26.** According to lecture (Text Retrieval and Extraction, slide 42), which of the following statements is true regarding boundary errors in named entity recognition?

(a) Precision and recall are unaffected by boundary errors and are straightforward to calculate in IE/NER.

(b) Selecting no entity at all is always worse than selecting a partial one.

(c) Boundary errors in IE/NER only affect recall but not precision.

(d) IE/NER evaluation always gives full credit for partial matches to entities.

(e) Selecting an incorrect span like 'Bank of Chicago' instead of 'First Bank of Chicago' can result in both a false positive and a false negative.

**Question 27.** 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) None of the above metrics are suitable for imbalanced datasets, and a different evaluation method is required.

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

(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 28.** According to the lecture on Data Mining II (Slide: "Breadth-First Search vs. Depth-First Search"), if you are analyzing a network graph and wants to find the shortest path between two nodes. Which traversal strategy should you use and why?

(a) `Breadth-First Search (BFS), because it explores all neighbors level by level and is guaranteed to find the shortest path in an unweighted graph.`

(b) `Neither BFS nor DFS, since shortest paths require a weighted graph and both algorithms ignore edge weights.`

(c) `Depth-First Search (DFS), because it is designed for level-order traversal and always considers all neighbors before going deeper.`

(d) `Depth-First Search (DFS), because it explores deeper paths first and therefore reaches the goal node more quickly.`

(e) `Breadth-First Search (BFS), because it uses a stack to prioritize recent nodes, making traversal faster overall.`

**Question 29.** According to Han et al., why does a Support Vector Machine (SVM) aim to maximize the margin between classes?

(a) `Because it helps the model avoid using kernel functions.`

(b) `Because it prevents the model from identifying outliers in the dataset.`

(c) `Because a larger margin reduces the risk of misclassification on unseen data, improving generalization.`

(d) `Because a larger margin ensures the model will always perform better than other models.`

(e) `Because maximizing the margin guarantees 100\% accuracy on the training data.`

**Question 30.** 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) `Ground Coffee`

(b) `Bread, Milk, Eggs.`

(c) `All Purpose Flour, Chocolate Chips, Ground Coffee.`

(d) `Milk`

(e) `All Purpose Flour, Ground Coffee.`

**Question 31.** According to page 420 of the textbook Data Mining Concepts and Techniques, what is the difference between extrinsic and intrinsic methods of evaluating clustering models?

(a) `Extrinsic methods involve a ground truth classification of each data point into ideal clusters, intrinsic methods do not involve a ground truth.`

(b) `Intrinsic methods are probabilistic, extrinsic methods are not.`

(c) `Extrinsic methods use a random sample of the clustered data, intrinsic methods involve all data points.`

(d) `Intrinsic methods only involve calculations within clusters, extrinsic methods also involve calculations with points in different clusters.`

(e) `Extrinsic methods are only used with density-based clustering, intrinsic methods are only used with K-Means and hierarchical clustering.`

**Question 32.** You ran an association rule mining algorithm on transaction data for a supermarket and discovered the following rule: {itemA, itemB} $\rightarrow$ {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 flawed because itemC is an independent item and is not necessarily purchased with itemA and itemB.`

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

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

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

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

**Question 33.** 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) `Gradient Boosting`

(b) `Boosting`

(c) `Bagging`

(d) `Stacking`

(e) `AdaBoost`

**Question 34.** 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 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.`

(b) `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.`

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

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

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

**Question 35.** 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) `(‘COVID-19’ AND ‘Vaccination’ AND ‘Symptoms’)`

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

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

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