Department of Artificial Intelligence and Machine Learning
YZZ211 - Introduction to Data Mining
Midterm Exam

5th Nov 2026, 15:00-16:30

Full Name : _____ Student ID: _____

Grade Table (for Lecturer use only)

Page:	2	3	4	5	6	7	Total
Points:	20	18	22	18	18	4	100
Score:							

Instructions for Midterm Exam

Welcome to the midterm exam of YZZ211 - Introduction to Data Mining and good luck! Please read the following rules and confirm by signing that you have read and understood the rules before you receive your exam:

1. The midterm exam shall be conducted between 15:00 and 16:30. Exam duration is 90 minutes. Students must finalise the exam by delivering it before 16:30. Students are not allowed to leave the exam in the first 30 minutes.

2. Student ID cards shall visibly be on the edge of desks till the end of the exam. Students without the student ID cards or Turkish identity cards shall not be participated into the exam.

3. This is a closed-book exam which means that students are not allowed to take notes, books, or any other reference material into the exam. Throughout the exam, students shall not possess mobile phones and electronic devices that are capable of storing, receiving, or transmitting information or electronic signals, such as computerised watches.

4. Students are not allowed to take a glance at the exam questions until told to do so. Students shall not communicate with any other student under any circumstances during the exam period. A student, who cheats, tries to cheat during the exam, or is identified to be cheating after investigating exam documents, is given 0 (zero) for that exam and a disciplinary investigation is opened against the student.

5. All numerical values in the exam shall be calculated according to two decimal digits. Otherwise, there will be a penalty.

6. An incorrect answer to a question is awarded no marks with no consideration of any partial credit. Therefore, no partial credit will be given.

7. This exam contains 50 multiple-choice questions, each worth 2 points. Select exactly one answer (A, B, C, or D) for each question.

In recognition of and in the spirit of the above rules, I certify that I will neither give nor receive unpermitted aid on this examination.

Signature: _____

Department of Artificial Intelligence and Machine Learning
YZZ211 - Introduction to Data Mining
Midterm Exam

5th Nov 2026, 15:00-16:30

1. **(2 points)** What is the main purpose of data mining?
   A. To store every dataset in a single spreadsheet.
   B. To replace all data collection with simulation.
   C. To discover useful patterns and knowledge in data.
   D. To guarantee that every prediction is correct.

2. **(2 points)** Which statement correctly describes the relationship between data mining and knowledge discovery in databases (KDD)?
   A. KDD is only a method for storing data.
   B. Data mining is one stage within the broader KDD process.
   C. Data mining and KDD exclude data preprocessing.
   D. KDD ends before patterns are interpreted.

3. **(2 points)** Which CRISP-DM phase focuses on defining business objectives and success criteria?
   A. Data preparation.
   B. Business understanding.
   C. Modelling.
   D. Deployment.

4. **(2 points)** Predicting wind speed from temperature, humidity, and air pressure is an example of which task?
   A. Classification.
   B. Regression.
   C. Clustering.
   D. Association rule discovery.

5. **(2 points)** A retailer groups customers by similar purchasing patterns without using predefined customer labels. Which task is this?
   A. Classification.
   B. Anomaly detection.
   C. Clustering.
   D. Regression.

6. **(2 points)** Which statement best describes the association rule $\{\text{Milk}\} \rightarrow \{\text{Coke}\}$?
   A. Transactions containing milk tend also to contain Coke.
   B. Buying milk causes a customer to buy Coke.
   C. Milk and Coke are never purchased together.
   D. Every transaction containing Coke must contain milk.

7. **(2 points)** Which task aims to detect credit card transactions that differ substantially from normal behaviour?
   A. Data aggregation.
   B. Association rule discovery.
   C. Anomaly detection.
   D. Dimensionality reduction.

8. **(2 points)** In data warehousing, what does ETL stand for?
   A. Extract, Transform, and Load.
   B. Estimate, Train, and Label.
   C. Encode, Transfer, and Link.
   D. Explore, Test, and Learn.

9. **(2 points)** Which three characteristics are the original three Vs of big data?
   A. Volume, validity, and visualisation.
   B. Volume, velocity, and variety.
   C. Variance, velocity, and verification.
   D. Variety, value, and versioning.

10. **(2 points)** Before collecting and analysing personal data, which action is most appropriate?
    A. Assume that removing names resolves every legal and ethical concern.
    B. Assume that all publicly accessible data can be used without restrictions.
    C. Check privacy only after the results have been published.
    D. Check that both data collection and the intended use are permitted.

11. **(2 points)** Which attribute is nominal, even though its values may be written as numbers?
    A. Height in centimetres.
    B. Temperature in kelvin.
    C. The number of items purchased.
    D. A postal code.

12. **(2 points)** Which attribute is measured on a ratio scale?
    A. Temperature in kelvin.
    B. A customer satisfaction rank.
    C. Temperature in degrees Celsius.
    D. A calendar date.

13. **(2 points)** Which representation is most suitable for shopping baskets containing different sets of purchased products?
    A. Transaction data, with each basket represented as a set of items.
    B. A graph containing no information about basket membership.
    C. A genomic sequence, with each product represented as a nucleotide.
    D. A time series containing only the purchase dates.

14. **(2 points)** What is missing-value imputation?
    A. Deleting every feature that contains any missing values.
    B. Estimating a replacement for a missing value.
    C. Converting every numerical feature into a category.
    D. Replacing every observed value with a missing-value marker.

15. **(2 points)** A merged dataset contains several records referring to the same person. Which data-cleaning operation addresses this problem?
    A. Deduplication.
    B. Feature extraction.
    C. Standardisation.
    D. Discretisation.

16. **(2 points)** Which procedure describes stratified sampling?
    A. Randomly sample from the whole dataset without forming groups.
    B. Divide the data into groups, then randomly sample within each group.
    C. Select observations only from the largest group.
    D. Select every observation more than once.

17. **(2 points)** What commonly happens as the dimensionality of a dataset increases?
    A. The data become sparse, and distances can become less informative.
    B. All missing values disappear.
    C. Every feature becomes equally relevant to the task.
    D. The number of observations automatically increases.

18. **(2 points)** What is the main criterion used by principal component analysis (PCA) to choose its new axes?
    A. Maximise the number of original features retained unchanged.
    B. Guarantee perfect separation between all class labels.
    C. Capture as much variation in the data as possible.
    D. Replace every feature with its most frequent value.

19. **(2 points)** Which statement describes wrapper-based feature selection?
    A. Create new features without evaluating the predictive model.
    B. Use a predictive model to evaluate candidate subsets of features.
    C. Select features only as an automatic part of fitting a model.
    D. Select features solely by examining their names.

20. **(2 points)** Which formula gives the z-score of a value x, using feature mean $\bar{x}$ and standard deviation s?
    A. $z = (x - \bar{x})/s$.
    B. $z = x/s^2$.
    C. $z = (x - s)/\bar{x}$.
    D. $z = x - \bar{x}$.

21. **(2 points)** For $A = (2, 1)$ and $B = (4, 3)$, what are the Euclidean and Manhattan distances, respectively?
    A. 2.83 and 2.00.
    B. 4.00 and 2.83.
    C. 2.00 and 4.00.
    D. 2.83 and 4.00.

22. **(2 points)** For $A = (2, 1)$ and $C = (1, 1)$, what is their cosine similarity, rounded to two decimal places?
    A. 0.71.
    B. 1.00.
    C. 0.89.
    D. 0.95.

23. **(2 points)** Two binary vectors have $f_{11} = 0$, $f_{10} = 1$, $f_{01} = 2$, and $f_{00} = 7$. What is their Jaccard similarity?
    A. 0.30.
    B. 0.70.
    C. 0.00.
    D. 1.00.

24. **(2 points)** Which measure is designed to combine attribute-wise comparisons when a dataset contains mixed attribute types?
    A. Gower's similarity or dissimilarity.
    B. Manhattan distance applied without accounting for attribute types.
    C. Euclidean distance applied directly to unencoded category names.
    D. Cosine similarity applied directly to arbitrary text labels.

25. **(2 points)** Which property must a distance satisfy to be a metric?
    A. $d(p, p) > 0$ for every point.
    B. $d(p, r) \leq d(p, q) + d(q, r)$.
    C. $d(p, r) > d(p, q) + d(q, r)$ for all points.
    D. $d(p, q) = -d(q, p)$ for all points.

26. **(2 points)** In kernel density estimation (KDE), what does the bandwidth mainly control?
    A. The number of class labels in the dataset.
    B. The amount of smoothing in the estimated density.
    C. The number of features selected for a classifier.
    D. The maximum number of missing values allowed.

27. **(2 points)** For a categorical attribute, what is the mode?
    A. The difference between the largest and smallest labels.
    B. The most frequently occurring value.
    C. The arithmetic mean of the category labels.
    D. The least frequently occurring value.

28. **(2 points)** Which measure of location is generally more robust to extreme outliers than the arithmetic mean?
    A. The range.
    B. The median.
    C. The variance.
    D. The standard deviation.

29. **(2 points)** How is the interquartile range (IQR) calculated?
    A. $Q_3 - Q_1$.
    B. $Q_3 + Q_1$.
    C. $\max(x) - \min(x)$.
    D. $(Q_1 + Q_3)/2$.

30. **(2 points)** What does the Pearson correlation coefficient measure?
    A. Whether one variable causes changes in another.
    B. The number of clusters in a dataset.
    C. The direction and strength of a linear relationship between two variables.
    D. Every possible nonlinear relationship between two variables.

31. **(2 points)** How is Spearman's rank correlation calculated?
    A. As the mean of the original observations.
    B. As the Euclidean distance between two category labels.
    C. As the difference between two sample variances.
    D. As the Pearson correlation between the ranks of two variables.

32. **(2 points)** Which choice can substantially change the appearance of a histogram for the same dataset?
    A. The names assigned to the dataset rows.
    B. The number of characters in the variable name.
    C. The number and width of its bins.
    D. The order in which the original records were entered.

33. **(2 points)** Which plot is particularly useful for comparing the median and spread of a numerical variable across several groups?
    A. A box plot for each group.
    B. A graph of category names without numerical values.
    C. A scatter plot of record identifiers only.
    D. A table containing only the group names.

34. **(2 points)** In a parallel-coordinates plot, how is one data object represented?
    A. As a single category label without coordinates.
    B. As a contour connecting locations with the same elevation.
    C. As one histogram bin.
    D. As a line joining its values on several parallel feature axes.

35. **(2 points)** In a classification decision tree, what does a leaf node usually provide?
    A. The complete list of all training records.
    B. A test that must always create two further child nodes.
    C. The name of the database storing the data.
    D. A predicted class label.

36. **(2 points)** What makes standard top-down decision tree induction a greedy procedure?
    A. It guarantees the globally optimal tree for every dataset.
    B. It always selects the attribute with the longest name.
    C. It evaluates every possible complete tree before making a split.
    D. It chooses the best available split at the current node.

37. **(2 points)** A node contains two records from class 1 and four from class 2. What is its Gini impurity, rounded to two decimal places?
    A. 0.44.
    B. 0.33.
    C. 0.00.
    D. 0.50.

38. **(2 points)** A node contains equal numbers of records from two classes. What is its entropy when logarithms are base 2?
    A. 0.00.
    B. 1.00.
    C. 0.50.
    D. 2.00.

39. **(2 points)** When choosing a decision tree split, how is information gain calculated?
    A. Parent entropy plus the entropy of the largest child node.
    B. Parent entropy minus the size-weighted average entropy of the child nodes.
    C. The unweighted sum of all child-node entropies.
    D. The number of child nodes divided by the number of classes.

40. **(2 points)** Which pattern most strongly suggests overfitting?
    A. High training error and equally high error on unseen data.
    B. A simple model with similar performance on training and unseen data.
    C. Low error on both training and unseen data.
    D. Very low training error but substantially higher error on unseen data.

41. **(2 points)** Which action is an example of pre-pruning a decision tree?
    A. Replace every numerical feature with a class label.
    B. Grow a complete tree and then remove a subtree.
    C. Stop splitting a node when it contains fewer records than a chosen threshold.
    D. Evaluate a completed tree on an independent test set.

42. **(2 points)** Which action describes post-pruning a decision tree?
    A. Add a new branch for every misclassified training record.
    B. Stop tree growth before a split is created.
    C. Remove selected subtrees after the tree has been grown.
    D. Use the final test set to train each new branch.

43. **(2 points)** A fraudulent transaction is incorrectly classified as legitimate. Treating fraud as the positive class, what type of result is this?
    A. A true negative.
    B. A false positive.
    C. A true positive.
    D. A false negative.

44. **(2 points)** A dataset contains 9,990 negative and 10 positive cases. A classifier predicts every case as negative. Why is its 99.9% accuracy misleading?
    A. It has 99.9% recall for the positive class.
    B. It correctly identifies all positive cases.
    C. It fails to detect any positive cases.
    D. It has no true negative predictions.

45. **(2 points)** A classifier has $TP = 150$ and $FP = 60$. What is its precision, rounded to two decimal places?
    A. 0.80.
    B. 0.29.
    C. 0.79.
    D. 0.71.

46. **(2 points)** A classifier has $TP = 150$ and $FN = 40$. What is its recall, rounded to two decimal places?
    A. 0.79.
    B. 0.80.
    C. 0.71.
    D. 0.21.

47. **(2 points)** A classifier has $TP = 150$, $FP = 60$, and $FN = 40$. What is its F1-score, rounded to two decimal places?
    A. 0.79.
    B. 0.71.
    C. 0.75.
    D. 0.80.

48. **(2 points)** Which quantities are plotted on the horizontal and vertical axes of a standard ROC curve, respectively?
    A. Training error and model complexity.
    B. True positive rate and false positive rate.
    C. False positive rate and true positive rate.
    D. Precision and recall.

49. **(2 points)** For a classifier whose ROC curve follows the random-guessing diagonal, what is the area under the ROC curve (AUC)?

    A. 0.00.

    B. 1.00.

    C. 0.75.

    D. 0.50.

50. **(2 points)** Which procedure correctly describes model selection and final evaluation for independent, identically distributed data?

    A. Use cross-validation within the training data to select hyperparameters; then evaluate the final model on an untouched test set.

    B. Fit the model to all data before selecting a test set.

    C. Report only the training accuracy of the selected model.

    D. Use the final test set repeatedly to choose hyperparameters.