COMP5310: Principles of Data Science - Final Examination

Time Allowed: 2 hours

Semester: Semester 1, 2024

Section A: Multiple Choice Questions (MCQ)

1. Consider the following scenario: H0 is true, and the test results are in the rejection state. Choose the correct answer.

(a) Type I error

(b) Type II error

(c) Matrix

(d) No correct answer

2. What is the range of the coefficient of determination R^2?

(a) Between 0 and 1

(b) Greater than 1

(c) Less than 0

(d) All of the above

3. Which of the following statements is incorrect regarding association rule metrics?

(a) High confidence implies low support

(b) High support implies low confidence

(c) Low confidence and low support can occur simultaneously

(d) None of the above

4. The loss function in logistic regression is based on which of the following properties of :

(a) Differentiable

(b) Non-differentiable

(c) Differential

(d) Cannot be determined

2. What is the range of the coefficient of determination R^2?

(a) Between 0 and 1

(b) Greater than 1

(c) Less than 0

(d) All of the above

3. Which of the following statements is incorrect regarding association rule metrics?

(a) High confidence implies low support

(b) High support implies low confidence

(c) Low confidence and low support can occur simultaneously

(d) None of the above

4. The loss function in logistic regression is based on which of the following properties of :

(a) Differentiable

(b) Non-differentiable

(c) Differential

(d) Cannot be determined

C 5. When choosing between L1 and L2 regularization, which scenario is most appropriate for using L1 regularization?

(a) Data with highly correlated features L2：权重平滑分配，Avoid丢失信息

(b) Data with low feature correlation

(c) Data where feature selection is desired L1，把不重要的 attr 权重变成0

(d) Both B and C

C ✓ 6. Which of the following ensemble methods is based on sequentially applying weak learners to the data, where each subsequent learner focuses more on the instances that were previously misclassified?

(a) Bagging

(b) Random Forest → 不同于 Bagging 的是随机抽取一部分 attributes 构建 model

(c) AdaBoost

(d) Stacking

b ✓ 7. In Principal Component Analysis (PCA), which of the following statements is true?

(a) PCA is a supervised learning technique. ✗

(b) PCA maximizes the variance along the principal components.

(c) PCA minimizes the reconstruction error without reducing dimensionality. ✗

(d) PCA always results in integer principal components. ✗

C ✓ 8. Which of the following model selection criteria incorporates both the goodness of fit and the complexity of the model?

(a) R-squared

(b) Mean Squared Error ✗

(c) Akaike Information Criterion (AIC)

(d) Confusion Matrix ✗

C ✗ 9. Decision tree is

(a) Not sensitive to noise data 有 "loss"

(b) Non-parametric model → 不用对 train data 的 distribution 有任何 Assumption

(c) No loss function model

(d) Bias tree model

Section B: True and False (10 Marks)

1. A decreasing learning rate during the training of a machine learning model leads to a flatter convergence landscape, which helps in achieving more stable and precise convergence towards the minimum.

[ ] True [ ] False

2. Gradient descent is a globally optimal optimization algorithm, ensuring that it always converges to the global minimum of any differentiable function.

[ ] True [ ] False

3. The Sum of Squares (SS) is utilized as a measure of variance within statistical models to assess the goodness of fit.

[ ] True [ ] False

4. Inverse Document Frequency (IDF) assigns higher weights to terms that appear frequently across all documents in a corpus, emphasizing their importance in distinguishing document relevance.

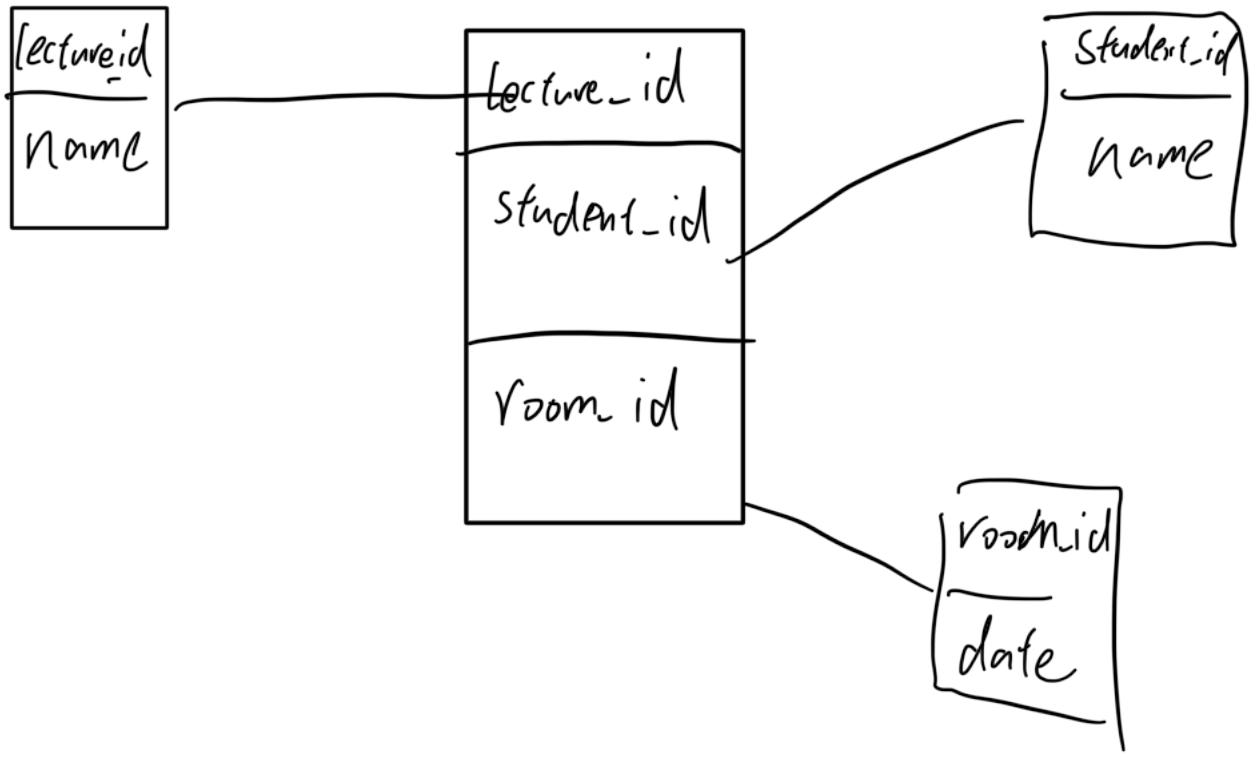
[ ] True [ ] False

Section C: SQL and Database Design

5. Given the following tables:

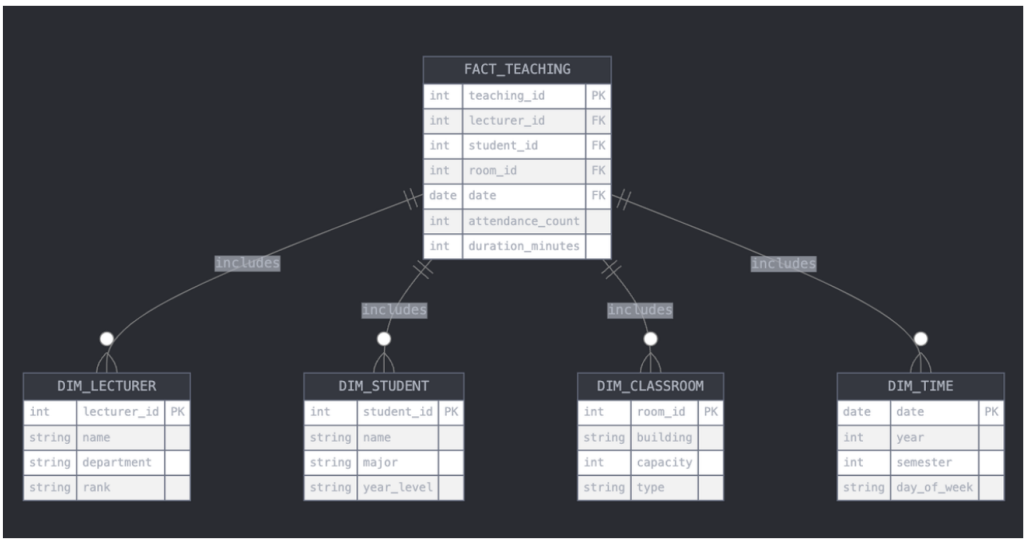
- Lecturer: lecturer_id, name
- Student: student_id, name
- Classroom: date, room_id

Construct a star schema for the university database, identifying the fact and dimension tables. Provide an Entity-Relationship diagram or a descriptive explanation outlining the relationships between tables.



code

Q1: Star Schema Design



1. **Fact Table (FACT_TEACHING)**:
   - Contains the measurable events (facts): attendance_count, duration_minutes
   - Foreign keys linking to all dimension tables
   - Represents the actual teaching events

2. **Dimension Tables**:
   - `DIM_LECTURER` : Contains lecturer details
   - `DIM_STUDENT` : Contains student information
   - `DIM_CLASSROOM` : Contains room details
   - `DIM_TIME` : Contains temporal information

6. Consider the following tables from an ER (Entity-Relationship) tutorial:

- Film: film_id, title
- Actor: actor_id, name
- Film_Actor: film_id, actor_id

Write an SQL query to find all films that have more than five actors.

The query should list the film_id, title, and the number of actors, ordered by the film title.

```
select
    film_id, title,
    count(actor_id)
From
    Film
Natural Join
    Film_Actor
Group by film_id, title
Having count(actor_id)
    >5
Order by title;
```

Section D: Short Answer Questions

7. Outline the steps involved in Principal Component Analysis (PCA). Provide a brief explanation of each step.

8. A research study aims to observe the effect of high-sugar, high-carbohydrate, and high-fat diets on ant population numbers.

Define the null hypothesis (H), specify an appropriate statistical test to evaluate H, and describe the criteria for rejecting or accepting H.

9. You are provided with a dataset containing information about various materials.

The dataset includes the following three features for each material. Describe the steps you would take to preprocess each of the three features to prepare the dataset for a classification algorithm.

Include any assumptions you make about the nature of the sorted numerical feature.

(a) A binary string representation derived from converting a sequence of categorical information into codes, for example, '01001110'.

(b) A numerical value representing the density of the metal in grams per cubic centimeter (g/cm^3).

(c) A numerical attribute that has been processed by sorting a related set of measurements.

(Note: The specific nature of this sorted data has been omitted.)

Q7

Short answer

Q1: Steps in Principal Component Analysis (PCA)

1. **Data Standardization**
   - Mean centering: Subtract mean from each feature
   - Scaling: Divide by standard deviation
   - Formula: $z = \frac{x - \mu}{\sigma}$

2. **Compute Covariance Matrix**
   - Calculate: $\Sigma = \frac{1}{n-1} X^T X$
   - Where X is standardized data matrix
   - Results in $p \times p$ symmetric matrix

3. **Calculate Eigenvectors and Eigenvalues**
   - Solve: $\Sigma v = \lambda v$
   - λ = eigenvalues
   - v = eigenvectors (principal components)

4. **Sort and Select Components**
   - Order eigenvalues in descending order
   - Select top k components based on:
     - Explained variance ratio
     - Cumulative variance threshold (e.g., 95%)

5. **Project Data**
   - Transform original data: $X_{new} = XW$
   - Where W is matrix of selected eigenvectors

Q2: Diet Effect on Ant Population

1. **Null Hypothesis (H_0):**
   - H_0 : There is no significant difference in ant population numbers among different diet types (high-sugar, high-carbohydrate, and high-fat)
   - H_1 : At least one diet type has a significantly different effect on ant population numbers

2. **Appropriate Statistical Test**: One-way ANOVA
   - Reasons:
     - Three independent groups (diet types)
     - Continuous dependent variable (population numbers)
     - Comparing means across groups

3. **Rejection Criteria**:
   - Reject H_0 if $p < \alpha$ (typically $\alpha = 0.05$)
   - F-statistic > Critical F-value
   - Assumptions to check:
     - Normality of residuals
     - Homogeneity of variance
     - Independence of observations

Q7

1. Calculate covariance matrix for every Attributes
2. if their cov is not 0, then we need to do PCA
3. Using eignvector and eignvalues to calculate PCA matrix
4. we usually do 95% importance, thats use diagonal component to calculate

Q8.. H_0: 3 High have no influence to Ant Population Number

. Experimental study is a good choice.

. We usually set α be 5% to reject H_0

Q9.

. First we need to deal with NAN, we can just remove them

. For (a) we can covert it into a one-hot code encoding

Q3: Data Preprocessing Steps

1. **Binary String Feature**
   - Convert to numeric array using binary encoding
   - Example: '01001110' → [0,1,0,0,1,1,1,0]
   - Consider dimensionality reduction if string is long
   - Check for consistent string length across samples

2. **Density Values**
   - Standard scaling: $z = \frac{x - \mu}{\sigma}$
   - Check for outliers using IQR or z-score method
   - Handle missing values if any
   - Consider log transformation if distribution is skewed

3. **Sorted Numerical Feature**
   - Min-max scaling: $x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$
   - Check for uniform distribution
   - Verify ordinal nature of data
   - Consider rank transformation if distribution is highly skewed

10. Describe how the Apriori algorithm works for mining frequent itemsets and association rules.

Include an explanation of how support and confidence are utilized in the algorithm.

we first use support to find frequent k-itemset
then for each k-itenset, we use confidence to find rule

11. In the context of skewed data distributions, explain the relationship between mean, median, and mode.

mean < Median < mode

Provide an example of a left-skewed distribution and indicate which measure of central tendency (mean, median, mode) would be most appropriate to represent the data.

12. Explain the relationship between a relational model and referential integrity.

How do they interact in the context of database design?

referential integrity refer to FK
and in a RM, we know the
relationship between each table,

13. Choose an appropriate machine learning algorithm for spam email classification.

Justify your choice based on the characteristics of the data and the requirements of the classification task.

NB, As NB ~~assume each record is independent~~
Simple, fast, good with missing value

Section E: Practical Calculation

14. Given the following dataset with three attributesSize (Large, Medium, Small), Colour (White, Brown),

and a binary Class attribute (Yes, No). Use Information Gain (IG) to determine which attribute should be used first to split the data.

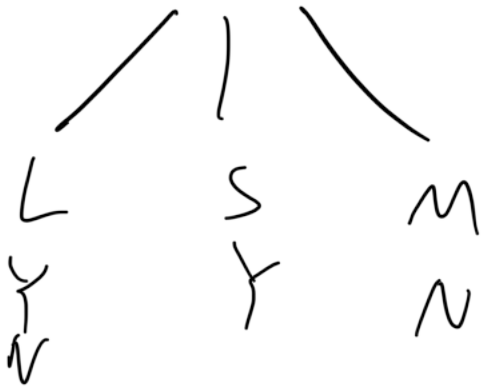
Explain your reasoning.

Table:

Size	Colour	Class
Large	White	Yes
Medium	Brown	No
Small	White	Yes
Large	Brown	No

14. $T_1 = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

For Size:



$$\frac{2}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{4}\left(-1\log 1 - 0\log 0\right)$$

$$+\frac{1}{4}\left(-0\log 0 - 1\log 1\right) = \frac{1}{2} = 0.5$$

$IG = 0.5$

For color:

$T_2 = 0$

$T_1 - T_2 = 1$

So using color

Medium White No

Small Brown Yes

Large White Yes

Medium Brown Yes

15. Given the following true labels and predicted labels for a classification task:

True Labels: [b, a, a, a, a, c, c]

Predicted Labels: [c, a, a, c, a, c, b]

Construct the confusion matrix for this classification problem.

Assume the classes are A, B, and C.

Pred \ Real	A	B	C
A	3	0	0
B	0	0	1
C	1	1	1

16. You have a dataset of vehicles with the following attributes:

- Color: Red, Blue, Black, etc.
- Origin: Domestic or Imported
- Model Type: SUV, Sedan, Truck, etc.
- Stolen: Yes or No

Given Probabilities:

	Yes	No
P(Stolen = Yes) = 0.05 / P(Stolen = No) = 0.95	0.05	0.95
P(Red | Stolen = Yes) = 0.30 / P(Red | Stolen = No) = 0.10	Red 0.3	0.1
P(Domestic | Stolen = Yes) = 0.60 / P(Domestic | Stolen = No) = 0.80	Dom 0.6	0.8
P(SUV | Stolen = Yes) = 0.40 / P(SUV | Stolen = No) = 0.30	SUV 0.4	0.3

Yes: $0.05 \times 0.3 \times 0.4 \times 0.6$

No: $0.95 \times 0.1 \times 0.3 \times 0.8$ So No

A vehicle is observed with:

- Color: Red

- Origin: Domestic

- Model Type: SUV

Determine if the car was stolen given red color SUV.

17. You are given a dataset of transactions in a retail store. Each transaction consists of a subset of items from the item set {I, I, I, I, I}. The minimum support threshold is set to 2.

Transactions:

T1: {I, I, I}

T2: {I, I}

T3: {I, I}

T4: {I, I, I}

T5: {I, I}

T6: {I, I}

T7: {I, I}

T8: {I, I, I}

T9: {I, I}

T10: {I, I, I}

(a) List all candidate 1-itemsets. Determine which 1-itemsets are frequent based on the minimum support threshold.

(b) Using the frequent 1-itemsets, generate candidate 2-itemsets. Identify the frequent 2-itemsets.

(c) From the frequent itemsets obtained, generate possible association rules.

End of Examination