**ITE 4005 Data Mining**

# Midterm Exam Outline

---

## Study materials:

- Chapters: ch1, ch6, ch8 and ch9
  - Study lecture notes and algorithm examples.
  - Read textbook: review the topics that were discussed in class

  - Students should thoroughly understand the concepts and implementations of the following algorithms. Study these algorithms through the examples discussed in class:
    - Apriori
    - FP-tree Growth
    - Decision tree
      - Attribute selection measure: Information gain, Gain ratio, Gini index
    - Bayesian classification:
      - Naïve Bayes classifier based on Bayesian theorem and independence among data attributes
      - Conditional probability in Bayesian networks
    - Neural network:
      - Backpropagation algorithm: feed forwarding and backpropagate the errors and reassign the values of weights and bias
    - Rule-based classification: Foil-gain

  - Students also review the following:
    - Interestingness measurement of correlated items
    - Measurement of classification accuracy

## Sample questions:

Here are some sample questions of the topics in chapter 1, 6, 8 and 9. Note that these questions are simply examples but they can help students to understand the format of the exam.

## True/False questions

Q1. Answer whether or not each of the following is a data mining task.

(a) Dividing the customers of company according to their profitability is a data mining task.

True False

(b) Predicting the future stock price of a company using historical data.

True False

(c) Computing the total sales of a company per location.

True False

(d) Monitoring the heart rate of a patient for abnormalities.

True False

Q2. Data mining systems attempt to generate all possible interesting patterns for completeness.

True False

Q3. Association rules are based on subjective interesting measures.

True False

Q4. The Apriori property can be used in the pruning step of the Apriori algorithm to reduce the size of candidate sets as follows. If a (k-1)-subset of a candidate k-itemset does not belong to the set of frequent (k-1)-itemsets, then the candidate is not frequent and therefore, it can be pruned from the set of k-itemset candidates.

True False

Q5. The Frequent Pattern Growth approach generates association rules directly from a FP tree without finding frequent itemsets.

True False

Q6. Classification predicts class labels which are discrete values whose order is insignificant.

True False

Q7. Classification consists of two steps, model construction and model usage.

True False

Q8. Decision Tree induction algorithm works in a bottom-up and divide-and-conquer manner.

True False

Q9. Decision Tree induction algorithm cannot handle categorical data.

True False

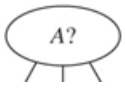
Q10. In post-pruning; a tree is pruned by halting its construction early.

True False

Q11. If all the tuples in a partition are of the same class then it is called “pure”.

True False

Q12. In Decision Tree induction algorithm, if A is discrete-valued, then one branch is grown for each know value of A.




True False

Q13. Bayesian belief networks allow modelling dependencies among attributes. In contrast, naïve Bayesian classifiers do not have this feature although they provide good classification results in most applications.

True False

Q14. In Bayesian theorem, no prior knowledge is needed to apply the formula and get the right probability.

True False

Q15. Bayesian Belief network (Bayesian networks) does not allow a subset of the variables conditionally independent.

True False

Q16. Bayesian classifier is easy to implement, and has high accuracy when the tuple's class conditions are dependent.

True False

Q18. Measuring quality of classification rules considers both coverage and accuracy.

True False

Q19. Bayesian classification is an incremental approach where each training example can incrementally increase/decrease the probability that a hypothesis is correct.

True False

Q20. One of the advantages of Naïve Bayesian classifier, dependencies among variables can be modeled.

True False

**Multiple choices questions**

Q1. The following is used to consolidate data into forms appropriate for data mining:

a. Data cleaning
b. Data integration
c. Data selection
d. Data transformation

Q2. Filling the missing values is a task of:

a. Data cleaning
b. Data integration
c. Data reduction
d. Data interpolation

Q3. Which statement is INCORRECT to describe FP-Growth algorithm?

a. Depth-first search
b. Often generates a huge number of candidates
c. Only 2 passes over dataset
d. Grows long patterns from short ones using local frequent items only

Q4. Which one is NOT a measure of Null-invariance?

a. Lift
b. Kulczynski
c. Cosine
d. Coherence

Q5. Which one is NOT a method of evaluating classification?

a. Accuracy
b. Scalability
c. Interpretability
d. Security

Q6. Which statement is INCORRECT to describe Decision Tree classifier?

a. Its construction does not require any domain knowledge
b. Decision trees can be converted easily to classification rules
c. The steps of decision tree inductions are complicated and slow
d. It has a good accuracy

Q7. Which statement is INCORRECT to describe attribute selection measures in Decision Tree classifier?

a. Information gain (ID3) is biased towards multivalued attributes.
b. Gain ratio (C4.5) tends to prefer unbalanced splits in which one partition is much smaller than the others.
c. Gini index performs well when number of classes is large
d. Gini index tends to favor tests that result in equal-sized partitions and purity in both partitions

Q8. Over-fitting in data classification occurs when:

a. The number of classified tuples differs significantly among classes
b. The classification accuracy is less than a pre-defined threshold
c. The test set and the training set include common tuples
d. Too small size of training data

Q9. When constructing a decision tree for data classification, the “best” splitting attribute is characterized by the following:

a. It maximizes expected information further required to classify tuples.
b. It splits the tuples database with the lowest impurity
c. It splits equal size of partitions
d. It creates a balanced tree

Q10. When used for data classification, neural networks have some drawbacks such as:

a. Poor interpretability.
b. Low tolerance for noisy data.
c. Long training time.
d. Required prior knowledge

Q11. Multilayer feed forward neural network consists of

a. Input layer and output layer
b. Input layer, hidden layer and output layer
c. Input layer, middle layer and output layer
d. Input layer, second layer and output layer

Q12. Which one is NOT a method that can be used to increase overall accuracy of classification

a. Bagging
b. Node distribution
c. Boostrapping
d. Cross validation

Q13. In Decision Tree induction algorithm, which one is NOT one of the conditions for stopping partitioning?

a. All samples for a given node belong to the same class.
b. There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf.
c. There are no samples left.
d. More than 50% of samples for a given node belong to the same class.

Q14. Which statement is INCORRECT to describe weakness of neural network as a classifier?

a. Long training time
b. Require a number of parameters typically best determined empirically, e.g., the network topology or "structure."
c. Poor interpretability: difficult to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network.
d. Low tolerance to noisy data.

Q15. What do we need to specify in order to define a neural network topology?

a. The number of  input layers
b. The number of hidden layers
c. The number of units in each hidden layer
a. The number of units in the output layer

## Problem solving questions

Q1. Let x and y be purchased items in the transactions stored in a DB. Explain the following expression that is used in the frequent itemsets mining process.

$x \Longrightarrow y$ [5%, 80%]

Q2. Assuming the following example dataset that shows 5 transaction and 4 sold items, calculate the confidence of the following rule: {Milk,Bread}$\Longrightarrow${Egg}.

| **ID** | **Items purchased** |
|---|---|
| t1 | {Milk, Bread} |
| t2 | {Egg} |
| t3 | {Cheese} |
| t4 | {Milk, Bread, Egg} |
| t5 | {Bread} |

Q3. Once all frequent itemsets are generated, the Apriori algorithm generates association rules from all frequent itemsets. Suppose itemset X = {A, B, C} is a frequent itemset. Find all association rules where minimum confidence is 70%. Use the following transaction DB.

| ID | Items purchased |
|---|---|
| t1 | A, B, E |
| t2 | B, C |
| t3 | A, B, C, D |
| t4 | A, C |
| t5 | A, C |
| t6 | A, B, C, E |
| t7 | A, B, C |

Q4. Given the following transaction database, construct an FP-tree (min_support = 3)

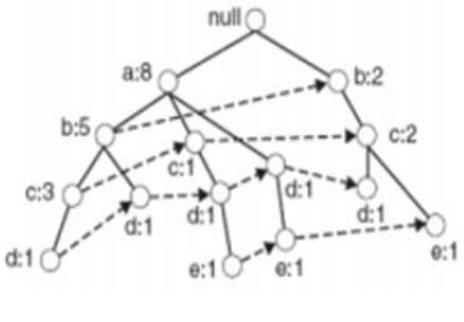
| ID | Items purchased |
|---|---|
| 1 | {f, a, c, d, g, i, m, p} |
| 2 | {a, b, c, f, l, m, o} |
| 3 | {b, f, h, j, o, w} |
| 4 | {b, c, k, s, p} |
| 5 | {a, f, c, e, l, p, m, n} |

Q5. Given the following transaction database (a), the FP-tree (b) is constructed. Extract all frequent itemsets from the prefix path sub-tree ending in ***d*** of the FP-tree.

Note: min_support = 2

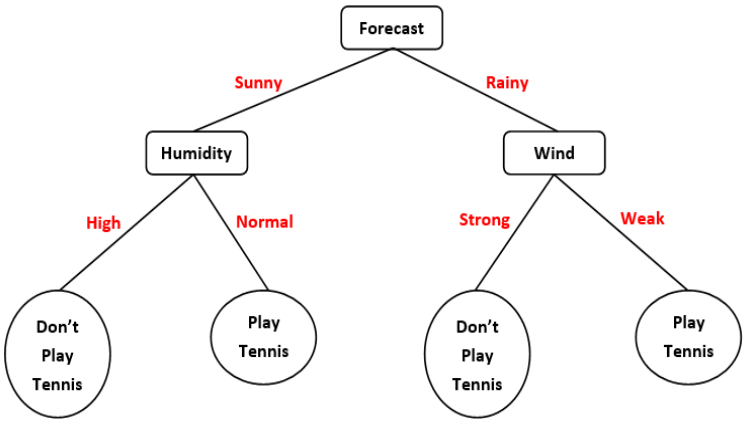
Transaction Data Set

| TID | Items |
|---|---|
| 1 | {a,b} |
| 2 | {b,c,d} |
| 3 | {a,c,d,e} |
| 4 | {a,d,e} |
| 5 | {a,b,c} |
| 6 | {a,b,c,d} |
| 7 | {a} |
| 8 | {a,b,c} |
| 9 | {a,b,d} |
| 10 | {b,c,e} |




*(a)* Transactions (b) FP-tree

Q6. Extract two classification rules from the following decision tree used to classify weather observations (tuples) into two classes: “Play Tennis” and “Don’t Play Tennis”.




Q7. What are the major differences among the three: (1) information gain, (2) gain ratio, and (3) gini index?

Q8. We have discussed the three following classifiers: (1) Naïve Bayesian algorithm, (2) Bayesian Belief Network, and (3) Neural Network.

(a) What are the major differences between (1) and (2)
(b) What are the major differences between (2) and (3)

Q9. Suppose you are requested to classify microarray data with 100 tissues (data size) and 10000 genes (attributes). Does Bayesian Belief Network work well? State your reason(s).

Q10. Both decision-tree induction and associative classification may generate rules for classification. What are their major differences? Why is it that in many cases an associative induction may lead to better accuracy in prediction?

Q11. Given the training dataset, design classification models for prediction of “Play Golf” based on the following: (1) decision tree and (2) Naïve Bayesian.

| Outlook | Temp | Humidity | Windy | Play Golf |
|---|---|---|---|---|
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Mild | Normal | False | Yes |
| Rainy | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Sunny | Mild | High | True | No |