Data Mining 2024 Fall 期末考考試題綱

* 這是題綱，不是題庫。主要供同學複習參考檢核用，考題與題型與題綱沒有直接關係。

* 考題類型主要以概念的理解分析、程序步驟、綜合應用為主，而非背誦記憶。

Data Preprocessing

1. 資料集有哪些常見的類型？
2. 評評估資料品質有哪些面向？
3. Data Preprocessing 有哪些 Tasks？
4. 常見的 Dirty Data 有哪些情形？Data Cleaning 如何處理?
5. Data Integration 做哪些 Tasks?
6. 何謂 Feature Scaling, Feature Normalization? 何謂要做這處理？有哪些不同做法？
7. 何謂 Feature (Attribute) Discretization? 為何要做這處理？有哪些不同做法？
8. 何謂 Feature Selection? 為何要做這處理？

Association Rule Mining

9. 何謂 Association Rule？(包括 Transaction, Transaction Database, Association Rule, Support, Confidence 的定義)。
10. 何謂 Association Rule Mining 的問題？其 Input 為何？Output 為何？何謂 Frequent Itemsets？ 與 Association Rules 有何不同？與 Association Rules 的關係為何？
11. 給定一 Transaction Database 與 Minimum Support，請舉例說明 Apriori 演算法 Mine Frequent Itemsets 的過程？Apriori 演算法的核心精神為何？為何稱之為 Apriori？
12. 請說明 DHP 演算法的核心精神為何？為何 DHP 演算法可以改進 Apriori 演算法的效率？ DHP 演算法中 Hashing Function 的 Collision Ratio 與其演算法效率的關係為何？
13. FP-Tree 演算法的核心精神為何？請舉例說明 FP-Tree 的 Data Structures。
14. 何謂 Quantitative Association Rules，請舉例說明，並說明如何利用 Association Rules 來求解 Quantitative Association Rules。Discretization 會影響 Quantitative Association Rule 的產生嗎?
15. Association Rules 的 Interestingness Measures，除了 Support, Confidence 之外，還有 Lift。Lift 的定義為何？為何必須多考慮 Lift？

16. 請舉例說明何謂 Temporal Association Rules?

17. 請舉例說明何謂 Intra-transaction Association Rules? 給定每日各上市公司的股價，如何求解 Intra-transaction Association Rules?

18. 除了購物交易資料，Association Rule Mining 還有哪些應用?

Clustering

19. 何謂 Clustering? Clustering 的目標為何? Clustering 與 Classification 的差別為何?

20. 何謂好的 Clustering? 影響 Clustering 效果的因素有哪些?

21. Clustering 常見的輸入資料包括 One Mode 與 Two Mode，何謂 One Mode 與 Two Mode?

22. Clustering 輸入的資料中，Attribute 的 Data Type 包括哪幾種? 不同的 Data Type 與 Clustering 的關係為何?

23. K-Means 演算法的核心精神為何? 執行過程為何? 其優缺點為何?

24. Hierarchical Clustering 演算法的核心精神為何? 請舉例說明 Single Link, Complete Link 其演算法執行的過程?

25. DBSCAN 演算法的核心精神為何? 執行過程為何? DBSCAN 演算法將資料分為哪三種?

26. Expectation Maximization 演算法的核心精神為何? 執行過程為何? 與 K-means 有合異同 ?

27. Clustering 與 Outlier (anomaly) Detection 的關係為何?

28. 如何評估 Clustering 的效果? 有哪些評估方法? 如何評估?

Classification

29. 何謂 Classification 的問題? 其 Input 為何? Output 為何? 請舉例說明其應用。

30. 請舉例說明 ID3 演算法如何求解出 Decision Trees? 如何利用 Information Gain (Entropy)求解 Decision Tree 產生過程中每個 node 的 Attributes。

31. 何謂 Rule-based classifier ? Rule-based classifier 何謂 Exhaustive ? 何謂 Exclusive ?

32. 如何由 Decision Tree 產生 Rule-based Classifier ? 所產生的 Classifier 具備 Exhaustive 及 Exclusive 的性質嗎 ?

33. 如何由 Association Rules 產生 Rule-based Classifier ? 所產生的 Classifier 具備 Exhaustive 及 Exclusive 的性質嗎 ?

34. 請舉例說明 Bayesian 演算法如何求解 Classification 的問題? Naive Bayesian 有何假設?

35. K-Nearest Neighbor 演算法的核心精神為何？影響其 Performance 的因素為何？ 為何 K-Nearest Neighbor 稱之為 Lazy Classification ?
36. Support Vector Machine 演算法的核心精神為何？SVM 的 Objective Function 為何？何謂 Support Vector? 何謂 Regularization ?
37. 如果 Decision Boundary 不是 Linear，SVM 如何求解？SVM 有哪些常見的 Kernel Function ?
38. SVM 如何求解 Multi-class Classification ?
39. SVM 也可求得分類的機率，其原理為何？
40. Ensemble Classifier 表現比 Single Classifier 準確高的 Necessary Condition 為何？
41. N 個 error rate e 的 Base Classifier 所組成的 Ensemble Classifier，假設每個 Base Classifier 都是獨立的，若此 Ensemble Classifier 採 Majority Vote，其錯誤率為何？
42. Bagging 如何透過 Ensemble 來提升準確率？
43. AdaBoost 的運作原理為何？AdaBoost 如何調整權重？其權重扮演什麼角色？
44. Random Forest 如何透過 Ensemble 來提升準確率？
45. 何謂 Class Imbalance？會導致什麼問題？有哪些解決辦法？
46. 何謂 K-Fold Cross Validation？何謂 Leave-One-Out? 評估 Classification 為何需要做這處理？
47. 何謂 Confusion Matrix？何謂 True Positive, True Negative, False Positive, False Negative ?
48. 何謂 Sensitivity, Specificity, True Positive Rate, False Negative Rate, False Positive Rate, False Negative Rate ?
49. 何謂 Precision, Recall, F-Measure ?
50. 何謂 ROC Curve, 何謂 AUC? ROC, AUC 的功用為何？如何產生 ROC Curve? 使用時機為何？
51. 何謂 Data Leakage? 何謂 Feature Leakage? 何謂 Training Example Leakage? Data Leakage 會產生什麼問題？舉例說明 Data Leakage 的問題。
52. 有哪些方法可提升 Classification 的準確率？ 何謂 Feature Engineering ？何謂 Grid Search ？Nominal Attributes 如何轉換成 SVM 的 Features ?
53. 何謂 Feature Importance？如何由 ID3, SVM 判斷 Feature Importance？哪些 Classification 演算法比較不受 Irrelevant 及 Redundant Attributes 的影響？為什麼？

Recommendation

54. 何謂 Recommendation？與 Personalization 的關係為何？
55. 何謂 Content-based Recommendation？Content-based Recommendation 的限制為何？
56. 何謂 Collaborative Recommendation？Collaborative-based Recommendation 的限制為何？
57. User-based Collaborative Recommendation 如何進行推薦？Item-based Collaborative Recommendation 如何進行推薦？Item-based Collaborative Recommendation 與 User-based Collaborative Recommendation 差別為何？
58. Recommendation 中，何謂 Cold Start 的問題？如何解決？
59. 在 Information Retrieval 中，何謂 Boolean Model？何謂 Vector Space Model？
60. Boolean Model, Vector Space Model 的關鍵詞如何自動取得？
61. 何謂 Stop Word？何謂 Stemming？何謂 N-Gram？何謂 Term Frequency？何謂 Document Frequency？
62. 用 Singular Vector Decomposition 降維的目的為何？如何透過 Singular Vector Decomposition 解決同義詞的問題？

1. relational, transactional, spatial, temporal, mobility, textual, graph, multimedia ... (Ch1, Last page)

2. accuracy, completeness, consistency, timeliness, believability, interpretability (Ch2, P8))

3. data Cleaning, Integration, transformation, reduction)

4. missing : imputation, delete entry ...
   noisy : binning, regression, clustering
   inconsistent :
   (ch2, P.18)

5. Schema integration, entity linking, handling redundancy

6. min-max normalization, Z-score normalization, normalization by decimal scaling (Ch2 P44)

7. binning, histogram analysis, clustering analysis (Ch2 P47)

8. –

9. I : item set, $T \subseteq I$: a transaction

D: a set of T, transaction database

Association Rule: $A \rightarrow B$, $A \subseteq I$, $B \subseteq I$

$$A \cap B = \phi$$

* $A \rightarrow B$

Support = $P(A \cup B)$, Confidence = $\frac{P(A \cup B)}{P(A)}$

10. Input: transaction database D,
minsup, min conf

Output: (Strong) association rules

frequent itemset (support > minsup)

11. min sup. cnt. = 2

D

A C D
B C E
A B C E
B E

C_1

A	2
B	3
C	3
D	1
E	3

→

F_1

A	2
B	3
C	3
E	3

→

C_2

A B	1
A C	2
A E	1
B C	2
B E	2
C E	2

F_2

A C	2
B C	2
B E	2
C E	2

→

C_3

B C E	1

* apriori property:
  1. all non-empty subsets of a frequent itemset must be frequent
  2. all supersets of a infrequent itemset must be infrequent

13. Motivation

① Mining in main memory to reduce # (DB scans)

② no candidate generation

③ more frequently occuring items will have better chances of sharing items

$D =$

a	c	d	f	g	i	m	p	
a	b	c	f	i	m	o		
b	f	h	j	o				
b	c	k	s	p				
a	c	e	f	l	m	n	p	

min. sup. cnt. = 3

a	3
b	3
c	4
d	1
e	1
f	4
g	1
h	1
i	2
k	1
l	1
m	3
n	1
o	2
p	3

frequent → 1-itenset

c	4
f	4
a	3
b	3
m	3
p	3

D (ordered) =

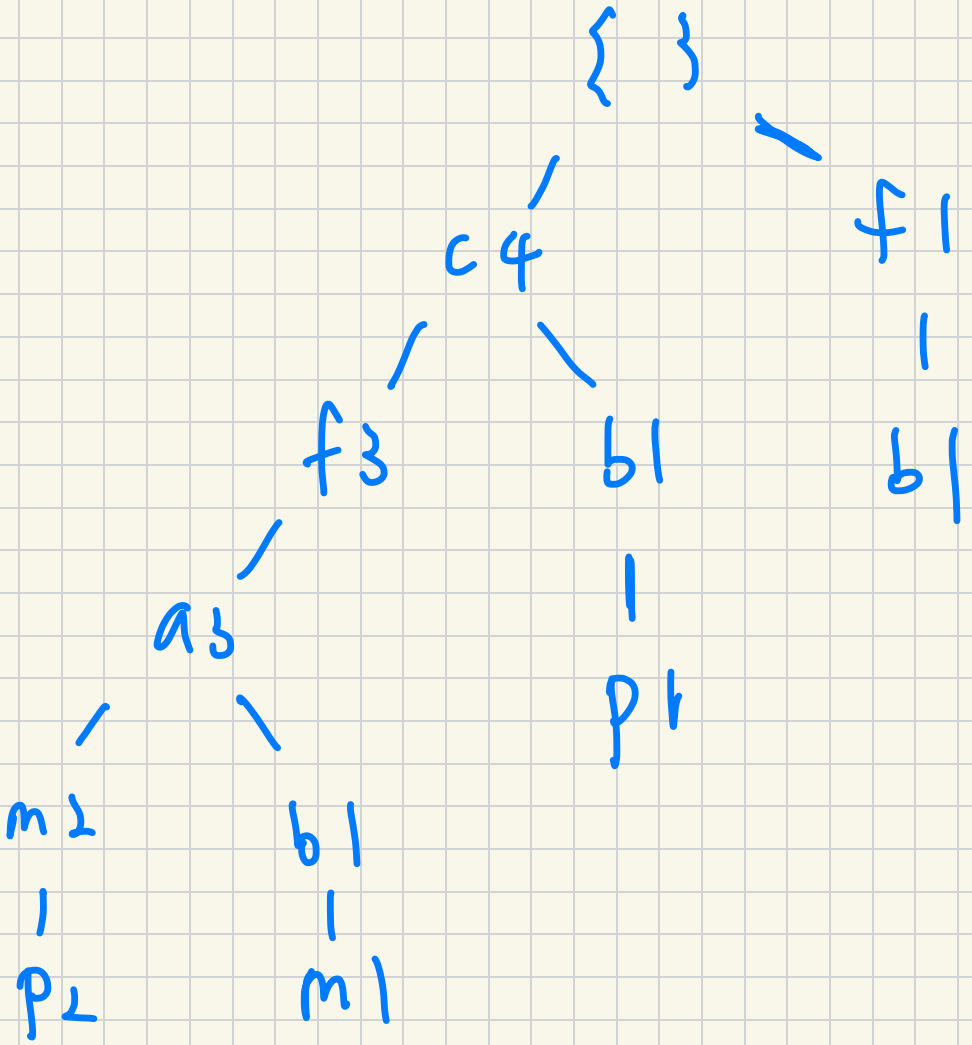
c f a m p

c f a b m

f b

c b p

c f a m p



suffix		
c		c:4
f	c:3	f:4, cf:3
a	cf:3	a:3, ca:3, fa:3
b		b:3
m	cfa:3	m:3, cm:3, fm:3 am:3, cfm:3, cam:3 fam:3, cfam:3
p	c:3	p:3, cp:3

15. $Lift = P(A \cup B) / P(A)\,P(B)$

16. ① Similarity measure & its implementation
② def. & representation of cluster chosen
③ clustering algo.

22. interval-scaled, ordinal-scaled, ratio-scaled, boolean, nominal

24.

	1	2	3	4	5
1	0	2.3	3.4	1.2	3.7
2		0	2.0	1.8	2.2
3			0	4.2	0.7
4				0	4.4
5					0

Proximity matrix

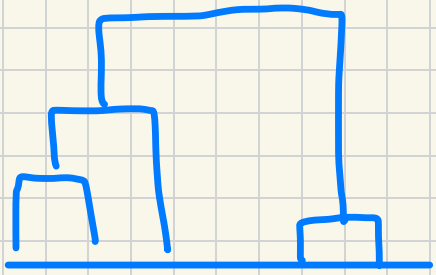
a) Single Link:

	1	2	3, 5	4
1	0	2.3	3.4	1.2
2		0	2.0	1.8
3, 5			0	4.2
4				0

	1, 4	2	3, 5
1, 4	0	1.2	3.4
2		0	2.0
3, 5			0

1° 3 & 5
2° 1 & 4
3° 2 & 1, 4

	1, 2, 4	3, 5
1, 2, 4	0	2.0
3, 5		0

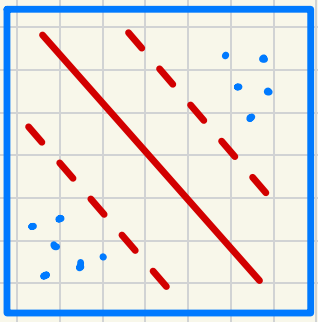


dendrogram

② Complete link

$$\max(a, b)$$

36.



$$\vec{w} \cdot \vec{x} + b = 0$$

$$D(x) = \frac{w^T x + b}{\|w\|} \begin{cases} \geq k^+, \text{ if } y_i = 1 \\ \leq k^-, \text{ if } y_i = -1 \end{cases}$$

$$y_i (w^T x + b) \geq M \|w\| \quad (\text{margin} = 2M)$$

①

⇒ Objective fnc. $\max (2M)$ subject to

$$y_i (w^T x + b) \geq M \|w\|$$

choose $\|w\| = \frac{1}{M}$

$$\min \left(\frac{\|w\|^2}{2} \right) \text{ subject to}$$

$$y_i (w^T x + b) \geq 1$$

③ $$\min \left(\frac{\|w\|^2}{2} + C \left(\sum_{i=1}^{n} \varepsilon_i^k \right) \right)$$

37. transform data into higher dimension

kernel trick : $\phi(x_i) \cdot \phi(x_j) = K(x_i, x_j)$

kernel fnc: linear, polynomial,
RBF, sigmoid, ...

39. Platt Scaling: maps the raw decision scores into $[0, 1]$ using logistic regression model.

41. $$\sum_{i=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{i} e^i (1-e)^{N-i}$$

43. error rate $\varepsilon_i = \frac{1}{N} w_j^{(i)} \delta(C(x_j) \neq y_j)$

amount of say $\alpha = \frac{1}{2} \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$

$$w_j^{(i+1)} = \frac{w_j}{z_i} \times \begin{cases} e^{-\alpha_i}, & \text{if } C_i(x_j) = y_j \\ e^{\alpha_i}, & \text{if } C_i(x_j) \neq y_j \end{cases}$$