國立交通大學 101 學年度第 1 學期

博士班資格考筆試考試試題

土木工程學系　資訊組(己)　科目：資料探勘研究與實務　選考學生數：1　考試時間：60 min

1. A database has six transactions. Let min_sup=60% and min_conf=90%.

TID	DATE	ITEMS_BOUGHT
T1	14/11/11	B, C, D, F
T2	14/12/11	A B, C, E, F
T3	14/01/12	A B, E
T4	14/02/12	A, B, C, D, E, G
T5	14/03/12	C E F G

(10%) List all of the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where X is a variable representing customers and item$_i$ denotes variables representing items (e.g., "A", "B", etc.):

$$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)[s, c]$$

2. (10%) Given the following large (frequent) 3-itemsets:

<1 2 3>
<1 2 4>
<1 3 4>
<1 3 6>
<2 3 4>
<2 3 5>
<2 4 5>
<3 4 5>

(a) Find the candidate 4-itemsets according to the apriori-generate algorithm.

(b) Find the candidate 4-itemsets after pruning.

3. You need to use examples or draw diagrams to aid your explanations.

(a) (5%) Briefly explain what is the Vector Space Model (VSM)

(b) (5%) Briefly explain tf_{ij}, $\log\frac{N}{df_j}$ and the formula $w_{ij} = tf_{ij} \times \log\frac{N}{df_j}$ (for term T_j of the document D_i).

(c) (5%) Explain how to use the k Nearest Neighbors (kNN) approach for document classification.

(d) (5%) Explain how to derive the category vector and use the Category vector for document classification (categorization).

4. There are two classifiers (test drugs) C1 and C2 for tumor. Suppose that there are 40% people having tumor in a city.

(a) (5%) For tumor patients, there are 30% positive by C1. For patients who don't have tumor, there are 40% positive by C1. Compute the ***Precision***, ***Recall,*** and ***Accuracy*** of the classifier C1.

(b) (5%) For patients who don't have tumor, there are 20% negative by C2. Suppose that the precision of the classifier C2 is 20%. What are the ***Accuracy*** and ***Recall*** of the classifier C2?

5. (10%) Use the similarity matrix in the following table to perform **complete link** hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

Table 3

	P1	P2	P3	P4	P5
P1	**1**	**0.33**	**0.96**	**0.54**	**0.78**
P2	**0.33**	**1**	**0.65**	**0.94**	**0.46**
P3	**0.96**	**0.65**	**1**	**0.58**	**0.25**
P4	**0.54**	**0.94**	**0.58**	**1**	**0.37**
P5	**0.78**	**0.46**	**0.25**	**0.37**	**1**

6.

(a) (10%) Explain the concept of support vectors, maximum marginal hyperplane and linear separation between classes in SVM (Support Vector Machines). You should draw a diagram in a 2-D plane to aid your explanation.

(b) (5%) Suppose that the two parallel hyperplanes for the decision boundary are:

$$w \bullet x + b = 1$$

$$w \bullet x + b = -1$$

Explain why maximizing the margin is equivalent to minimizing the following objective function:

$$f(w) = \|w\|^2 / 2.$$

(c) (5%) Explain the usage of Kernel function. Use the following example to assist your explanation.

$$\Phi(u) = (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1); \Phi(v) = (v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1); \Phi(u) \bullet \Phi(v) = ?$$

7.

(a) (10%) Explain the basic concept of EM (Expectation-Maximization) clustering. What are the differences between K-means and EM clustering in terms of the assignment of data points to clusters and the computation of centroids / model parameters?

(b) (10%) Given the following two mixture models (1) and (2).

Which one has higher expected likelihood? Why?

$$P(\mathbf{O}|\Theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j P_j(o_i|\Theta_j)$$

(1)

(2)

國立交通大學 101 學年度第 1 學期

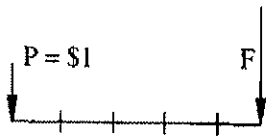
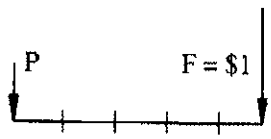
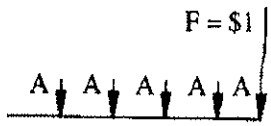
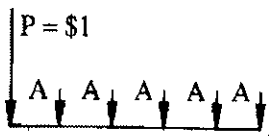
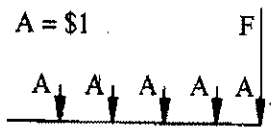
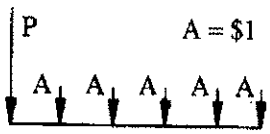
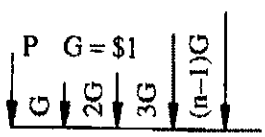
博士班資格考筆試考試試題

土木工程學系　資訊組(己)　科目：水資源規劃　選考學生數：1　考試時間：60 min

1. 某一水資源計劃估計在完成後至第 5 年末時每年增加 20,000 元的收益，第 6 年~第 20 年末收益維持 100,000 元，接著每年逐年遞減 10,000 元直至 30 年末，假設利率 5%，請問收益現值為何？（25%）

TABLE 2.1.1

Summary of discounting factors

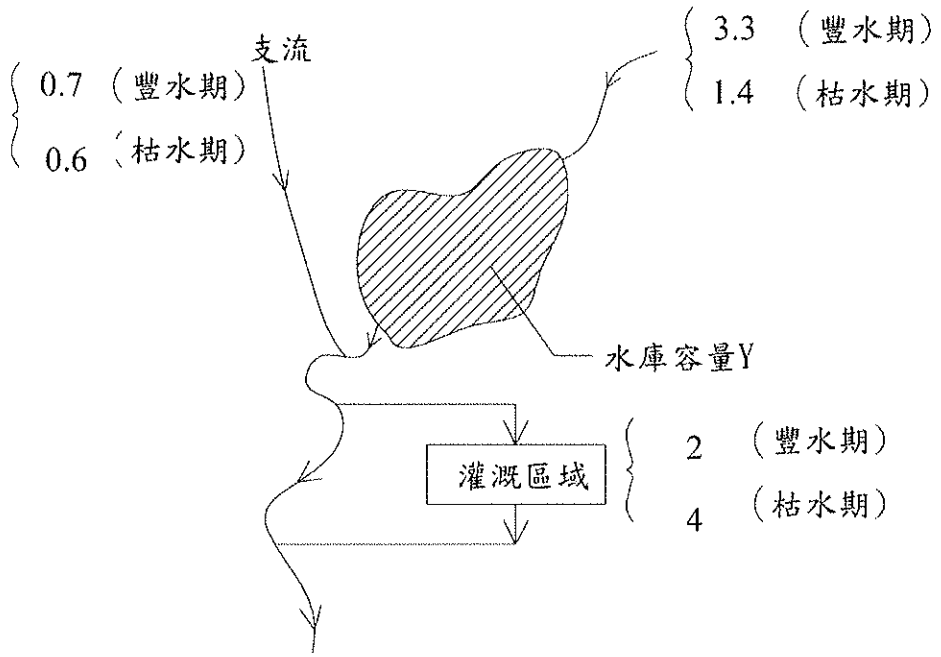
Type of Discount Factor	Symbol	Given*	Find	Factor	
Single-Payment Factors					
Compound-amount factor	$\left(\frac{F}{P}, i\%, n\right)$	P	F	$(1+i)^n$	
Present-worth factor	$\left(\frac{P}{F}, i\%, n\right)$	F	P	$\frac{1}{(1+i)^n}$	
Uniform Annual Series Factors					
Sinking-fund factor	$\left(\frac{A}{F}, i\%, n\right)$	F	A	$\frac{i}{(1+i)^n-1}$	
Capital-recovery factor	$\left(\frac{A}{P}, i\%, n\right)$	P	A	$\frac{i(1+i)^n}{(1+i)^n-1}$	
Series compound-amount factor	$\left(\frac{F}{A}, i\%, n\right)$	A	F	$\frac{(1+i)^n-1}{i}$	
Series present-worth factor	$\left(\frac{P}{A}, i\%, n\right)$	A	P	$\frac{(1+i)^n-1}{i(1+i)^n}$	
Uniform Gradient Series Factors					
Uniform gradient series present-worth factor	$\left(\frac{P}{G}, i\%, n\right)$	G	P	$\frac{(1+i)^{n+1}-(1+ni+i)}{i^2(1+i)^n}$	

*The discount factors represent the amount of dollars for the given amounts of one dollar for P, F, A and G.

2. 兩個水資源開發方案，若已知資金利率為 10%，而且無須考慮稅賦支出，請使用益本比法建議應採取哪一方案？ (25%)

	方案甲	方案乙
經濟壽命	**40** 年	**20** 年
殘值	15	12
每年效益	25	22
每年營運維修成本	5	3
期初成本	300	160

3. 有一水資源系統如下圖所示，水庫之進水量在豐水期為 3.3 單位(每單位10^8 m^3)，枯水期為 1.4 單位，水庫下游支流流入量在豐水期為 0.7 單位，枯水期為 0.6 單位。假設若灌溉需求水量豐水期為 2 單位，枯水期為 4 單位，若水庫之供水效益為(10*供水量)，而水庫投資(含營運、維護)成本之為 22*Y，Y 為水庫容量，(a). 請寫出在最佳利益(效益-成本)考量下之最佳規劃數學模式，包含變數定義、目標函數及限制式。(b).在最佳利益(效益-成本)下之水庫庫容及豐枯兩季之供水量為何？(25%)



水源調配系統圖

4. (a) 請簡述多目標規劃問題與單目標規劃(優選)問題之主要差異，及(b)以簡圖說明何謂多目標規劃問題之非劣勢解(柏拉圖解)。 (25%)