D5: Databases and Information Systems
Information Retrieval and Data Mining, WS 2013/14
Dr. Klaus Berberich and Dr. Pauli Miettinen
Third Short Test, Duration: **45 minutes**

max planck institut informatik

Name:					
Matriculation Number:					
Tutorial Group:	A ☐	B ☐	C ☐	D ☐	E ☐

Question:	1 (5 Points)	2 (5 Points)	3 (5 Points)	4 (5 Points)	Total (20 points)
Score:					

General instructions:

- **The written test contains 4 questions and is scheduled for 45 minutes. The maximum amount of points you can earn is 20.**
- **Please verify if your exam consists of 12 pages with 4 questions printed legibly, else contact the examiner immediately.**
- **No electronic devices (calculator, notebook, tablet, PDA, cell phone) are allowed.**
- **Answers without sufficient details are void (e.g.: you can't just say "yes" or "no" as the answer unless the question specifically asks for only yes/no answers).**
- **Last page consists of material that you may use to solve the questions. You may detach the last page for your convenience.**
- **You will be provided additional working sheets if necessary. Make sure to return them along with your solution sheet.**
- **Please provide your ID card when asked by the examiner.**
- **Please fill in name, matriculation number (student registration number) and tutor group in the form above and return the solution sheets into the provided box.**
- **Please sign below.**

Student's Signature _____

Frequent Itemset Mining

Problem 1. Consider the transaction data below.

a) Find all frequent itemsets with *support* ≥ 5. List the itemsets together with their supports. [**3 points**]

b) Which of the above-found frequent itemsets are closed? [**2 points**]

	Items									
tid	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
1	1	1	1	0	0	0	0	1	1	1
2	1	0	1	0	1	1	0	1	1	1
3	1	0	0	1	0	0	1	1	1	1
4	0	1	1	0	1	0	1	1	0	0
5	0	1	1	1	1	0	1	0	0	0
6	1	1	0	0	1	1	1	0	1	0
7	1	1	0	0	0	0	0	0	1	0
8	1	1	1	0	1	1	0	0	1	0
9	1	1	0	0	0	0	0	0	1	1
10	0	0	1	1	1	1	0	0	0	0

Table 1: Transaction data for Problem 1.


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Association Rule Mining

Problem 2.

Consider the transaction data below.

a) What is the confidence and support of the following rules? **[2 points]**

1. $\{A, B\} \to \emptyset$
2. $\{C\} \to \{D\}$

b) Consider the itemset $X = \{A, B, C\}$. List all association rules of type $Y \to Z$, where $\emptyset \neq Y \subsetneq X$ and $Z = X \setminus Y$, and that have *support* ≥ 4 and *confidence* $\geq 2/3$. **[3 points]**

	Items			
tid	A	B	C	D
1	1	1	1	0
2	1	1	1	0
3	0	1	1	1
4	0	1	1	0
5	0	1	1	1
6	1	1	0	0
7	1	1	0	0
8	1	1	1	0
9	1	1	1	0
10	0	0	1	1

Table 2: Transaction data for Problem 2.


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Right or Wrong?

Problem 3. Answer to the following questions either "yes" or "no". You can also leave the question without an answer. For each correct yes or no answer you will gain ½ points; for each wrong answer you will lose ½ points. If you leave a question without an answer, you will neither gain nor lose any points. Your minimum overall points are 0.

[$1/2$ **points for each correct answer,** $-1/2$ **points for each wrong answer**]

a) You can list all frequent itemsets and their support of a data set if you know all maximal itemsets of the data set.

b) If we use cosine measure (IS) to measure the interest of an association rule $\{A\} \to \{B\}$, its value does not change if we add a new transaction that does not contain either A or B.

c) An itemset is called *non-derivable* if we cannot build an upper bound to its support from the supports of its subsets.

d) In order to perform a hierarchical clustering, we do not need to know the coordinates of the elements to be clustered, only their mutual distances (i.e. the distance metric).

e) In order to perform a k-means clustering, we do not need to know the coordinates of the elements to be clustered, only their mutual distances (i.e. the distance metric).

f) If we use k-means to cluster data that is sampled from independent d-dimensional normal distributions with the same variance, the algorithm will take $2^{\Omega(\sqrt{n})}$ iterations with high probability.

g) The k-means++ algorithm improves the standard k-means by changing how the data points are assigned to the clusters.

h) The algorithm to build the decision tree will always return the most accurate tree.

i) In predictive classification, the quality of the classifier is based solely on how well it classifies the training data.

j) An ensamble classifier can only be build if the base classifiers are single-split decision trees.


MAX-PLANCK-GESELLSCHAFT


max planck institut
informatik

Naïve Bayes Classifier

Problem 4.

Consider the training data below. Classify the following test records based on a Naïve Bayes classifier trained on the training data. You only need to compute the probabilities you will need for the classification. For your answer, you need to tell which class has higher posterior probability. You don't need to compute the final posteriors as long as it is clear which one is bigger.

[5 points]

a) $X = (S = \text{CS}, M_1 = \text{Yes}, M_2 = \text{No}, P = \text{No})$

b) $Y = (S = \text{CS}, M_1 = \text{No}, M_2 = \text{Yes}, P = \text{Yes})$

c) $Z = (S = \text{Ph}, M_1 = \text{Yes}, M_2 = \text{No}, P = \text{Yes})$

Attributes				
S	M_1	M_2	P	Class
CS	Yes	Yes	No	+
CS	Yes	Yes	Yes	+
CS	Yes	Yes	No	−
CS	Yes	Yes	No	−
Phys	Yes	No	Yes	+
Phys	No	No	No	−
CS	Yes	Yes	Yes	+
CS	No	Yes	No	−
Econ	Yes	Yes	Yes	+
Econ	No	No	Yes	−

Table 3: Training data for Problem 4.


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT


MAX-PLANCK-GESELLSCHAFT


max planck institut
informatik

ADDITIONAL MATERIAL

Linear algebra

- Identity matrix: n-by-n matrix I such that $I_{ij} = 1$ iff $i = j$ and $I_{ij} = 0$ otherwise
- Product with identity matrix: $AI = IA = A$ for all n-by-n matrices A
- Matrix inverse: $A^{-1}A = AA^{-1} = I$
- Transpose identities: $(A^T)^T = A$ for all A; $(AB)^T = B^T A^T$ when the product is well-defined
- Inverse of a product: $(AB)^{-1} = B^{-1}A^{-1}$ if A and B are invertible
- Inverse of orthogonal matrices: $A^T = A^{-1}$ iff A is orthogonal

Probability & Statistics:

- Bayes' Theorem: $\Pr[A|B] = \dfrac{\Pr[B|A]\ \Pr[A]}{\Pr[B]}$
- Law of Total Probability: $\Pr[B] = \sum_{i=1}^{n} \Pr[B|A_i] \Pr[A_i]$ for disjoint events A_i with $\sum_{i}^{n} \Pr[A_i] = 1$
- Expectation: $\mathbf{E}[X] = \sum_{k=1}^{n} k f_X(k)$ and Variance: $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$ for a discrete RV X with density function f_X
- Markov inequality: $\Pr[X \geq t] \leq \dfrac{\mathbf{E}[X]}{t}$ for $t \geq 0$ and a non-neg. RV X
- Chebyshev inequality: $\Pr[|X - \mathbf{E}[X]| \geq t] \leq \dfrac{\mathbf{Var}[X]}{t^2}$ for $t > 0$ and a non-neg. RV X
- Sample Mean: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and Sample Variance: $S_X^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- For an estimator $\hat{\theta}$ of parameter θ over i.i.d. samples $\{X_1, X_2, .., X_i, .., X_n\}$,
  - If $\mathbf{E}[X_i] = \mu$, then $\mathbf{E}[\hat{\theta}_n] = \mu$
  - If $\mathbf{Var}[X_i] = \sigma^2$, then $\mathbf{Var}[\hat{\theta}_n] = \dfrac{\sigma^2}{n}$
  - Standard Error: $se(\hat{\theta}) = \sqrt{\mathbf{Var}[\hat{\theta}_n]}$
  - Mean Squared Error: $MSE[\hat{\theta}_n] = (\mathbf{E}[\hat{\theta}_n] - \theta)^2 + \mathbf{Var}[\hat{\theta}_n])$


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT