# Försättsblad till skriftlig tentamen vid Linköpings universitet

| | |
|---|---|
| **Datum för tentamen** | 2015-08-25 |
| **Sal (1)** | G36 |
| **Tid** | 8-12 |
| **Kurskod** | TDDD41 |
| **Provkod** | TEN1 |
| **Kursnamn/benämning**<br>**Provnamn/benämning** | Data Mining - Clustering and Association Analysis<br>Skriftlig tentamen |
| **Institution** | IDA |
| **Antal uppgifter som ingår i tentamen** | 7 |
| **Jour/Kursansvarig**<br>Ange vem som besöker salen | Patrick Lambrix / Jose Pena |
| **Telefon under skrivtiden** | 2605 / 1651 |
| **Besöker salen ca klockan** | 9:30, 10:45 |
| **Kursadministratör/kontaktperson**<br>(namn + tfnr + mailaddress) | Carita Lilja, 1463, carita.lilja@liu.se |
| **Tillåtna hjälpmedel** | dictionary |
| **Övrigt** | |
| **Antal exemplar i påsen** | |

Institutionen för datavetenskap
Linköpings universitet

# EXAM

# 732A31and TDDD41
# Data Mining –
# Clustering and Association Analysis
# August 25, 2015, kl 8-12

*Teachers:* Patrick Lambrix, José M Pena

*Instructions:*

- Start each question at a new page.
- Write at one side of a page.
- Write clearly.
- If you make assumptions about a question, that are not explicitly stated, you need to write these down. (These assumptions cannot change the exercise or question.)

*Help:* dictionary

GOOD LUCK!

## 1. Clustering by partitioning (2p+2p=4p)

a. Describe the principles and ideas regarding PAM.
- Give a sketch of the algorithm.
- Define swapping cost.

b. Given the graph representation of the clustering problem where n is the number of data points and k is the number of clusters.

(i) What does a node represent?
(ii) How can this graph be used for finding a solution for the clustering problem?
(iii) When are two nodes neighbors and how many neighbors does a node have?
(iv) Considering PAM, CLARA and CLARANS, the graph for which algorithm/algorithms contains/contain the most nodes? Explain.

## 2. Hierarchical clustering (3+2=5p)

a. Describe the principles and ideas regarding Agglomorative Hierarchical Clustering. Show the different steps of the algorithm using the dissimilarity matrix below and complete link clustering. Give partial results after each step.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 5 | 0 | | | |
| 3 | 9 | 10 | 0 | | |
| 4 | 3 | 2 | 6 | 0 | |
| 5 | 7 | 1 | 4 | 8 | 0 |

b. Describe the principles and ideas regarding the CHAMELEON algorithm. Explain the major steps.

## 3. Clustering categorical data (2+1=3p)

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to give a sketch of the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.
b. What is the main idea behind OPTICS?

**4. Data mining concepts (1p+1p+1p+1p=4p)**

a. The purpose of data mining is to extract interesting patterns from a huge amount of data. When is a pattern 'interesting' in this case?
b. Data in the real world can be dirty. Give 3 reasons and an example for each.
c. Show how an interval-based distance measure can be defined for ordinal variables.
d. Show how a distance measure can be defined for categorical (or nominal) variables.

**5. FP grow algoritm (2p+2p+2p=6p)**

a. Describe the FP grow algorithm, i.e. describe how it works in general not in a particular example.

b. Describe how to incorporate a monotone constraint into the FP grow algorithm.

c. Describe how to incorporate an antimonotone constraint into the FP grow algorithm.

**6. Apriori algorithm (2p+1p+1p+1p=5p)**

d. Describe the Apriori algorithm, i.e. describe how it works in general not in a particular example.

e. Describe how to incorporate a monotone constraint into the Apriori algorithm.

f. Describe how to incorporate an antimonotone constraint into the Apriori algorithm.

a. Prove formally the correctness of the Apriori algorithm.

**7. Constraints and lift (1p+1p+1p=3p)**

a. Describe what a monotone constraint is. Do not give a particular example. Describe the constraint in general terms.

b. Describe what a convertible antimonotone constraint is.

c. Describe what the lift of an association rule is.