# Datasta Tietoon, Autumn 2007

EXERCISE PROBLEMS 5 [ Dec 7th 2007, Dec 12th 2007 ]

**H5 / 1. (Frequent itemsets)**
Consider 0-1 observation set

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $a$ | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| $b$ | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| $c$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $d$ | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |

There are four variables $a, b, c, d$ and ten observations. Find the frequent itemsets when threshold value is $N = 4$.

**H5 / 2. (Levelwise algorithm)**
What is the time complexity of levelwise algorithm as a function of the size of data and the number of examined candidate sets?

**H5 / 3. (Tšernov border)**
Examine the Tšernov border mentioned in the lectures. How does the border behave as a function of different parameters?

**H5 / 4. (Segmentation)**
Consider one-dimensional time series, whose then points are 1, 3, 6, 2, 4, 2, 8, 9, 7, 8. Find the best way to segment this time series into two segments so that the error

$$\sum_{j=1}^{k} \sum_{i=b_j}^{e_j} (y_i - w_j)^2$$

is minimized.

**H5 / 5. (Segmentation)**
Show that the best representation of one segment is the mean of its items, in other words, the error

$$\sum_{i=b_j}^{e_j} (y_i - w_j)^2$$

is minimized when $w_j$ is the mean of items $y_{b_j}, \ldots, y_{e_j}$. What would be the best choice for value of $w_j$ if

$$\sum_{i=b_j}^{e_j} |y_i - w_j|,$$

were minimized?