**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
Prof. Dr. Peer Kröger
Yifeng Lu

# Knowledge Discovery in Databases II

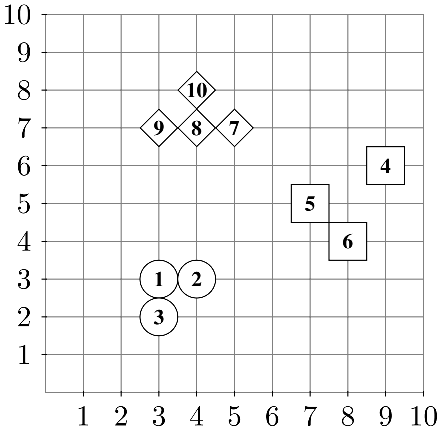
SS 2017

## Exercise 8: Data Stream Clustering and Classification

### Exercise 8-1 Cluster Features

Given the following dataset:

| ObjID | Cluster | $X$ | $Y$ | $t$ |
|---|---|---|---|---|
| 1 | A | 3 | 3 | 1.7 |
| 2 | A | 4 | 3 | 3.5 |
| 3 | A | 3 | 2 | 1.2 |
| 4 | B | 9 | 6 | 4.1 |
| 5 | B | 7 | 5 | 5.0 |
| 6 | B | 8 | 4 | 1.2 |
| 7 | C | 5 | 7 | 4.7 |
| 8 | C | 4 | 7 | 2.3 |
| 9 | C | 3 | 7 | 2.2 |
| 10 | C | 4 | 8 | 2.2 |



Compute the CluStream cluster features CFT for each of these three clusters.

A new observation in the stream is $p = (X = 8, Y = 5, t = 6.1)$.

Run the "online micro-cluster maintainance" of CluStream for this Point $p$.

### Exercise 8-2 Hoeffding trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1 - 2$ years, $2 - 7$ years, $> 7$ years)
- Gender (male, female)
- Residential area (urban, rural)

These are the first 8 examples.

| Person | Time since license | Gender | Area | Risk class |
|---|---|---|---|---|
| 1 | $1 - 2$ | m | urban | low |
| 2 | $2 - 7$ | m | rural | high |
| 3 | $> 7$ | f | rural | low |
| 4 | $1 - 2$ | f | rural | high |
| 5 | $> 7$ | m | rural | high |
| 6 | $1 - 2$ | m | rural | high |
| 7 | $2 - 7$ | f | urban | low |
| 8 | $2 - 7$ | m | urban | low |

- Incrementally construct a Hoeffding tree for this example.
  Use information gain and $\delta = 0.2$ and $N_{\text{min}} = 2$.
- Compute the value of $\delta$ at which the tree would still consist of the leaf only.