Main Data Mining Problems

UNIVERSITY OF LIVERPOOL

Four fundamental problems

- Association pattern mining

- Classification

- Clustering

- Outlier detection

Association Pattern Mining

- Special case: **Frequent Pattern Mining** (binary data sets)
  - Given $n \times d$ data matrix, identify all subsets of columns (**features**) such that at least a fraction s of rows (**objects**) in the matrix have all the features enabled (i.e., the features take on the value of 1).

Example (let $s = 0.65$)

Transaction	Milk	Butter	Bread	Mashrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	1	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

{Milk, Butter, Bread} are frequently bought together

Classification

- The goal is to use **training data** to learn relationships between a fixed feature (called **class label**) and the remaining features in the data.

- The resulting **learned model** may then be used to estimate (predict) values of the class label for records, where the value is not known.

- The objects whose class label is unknown are **test objects** (**test data**).

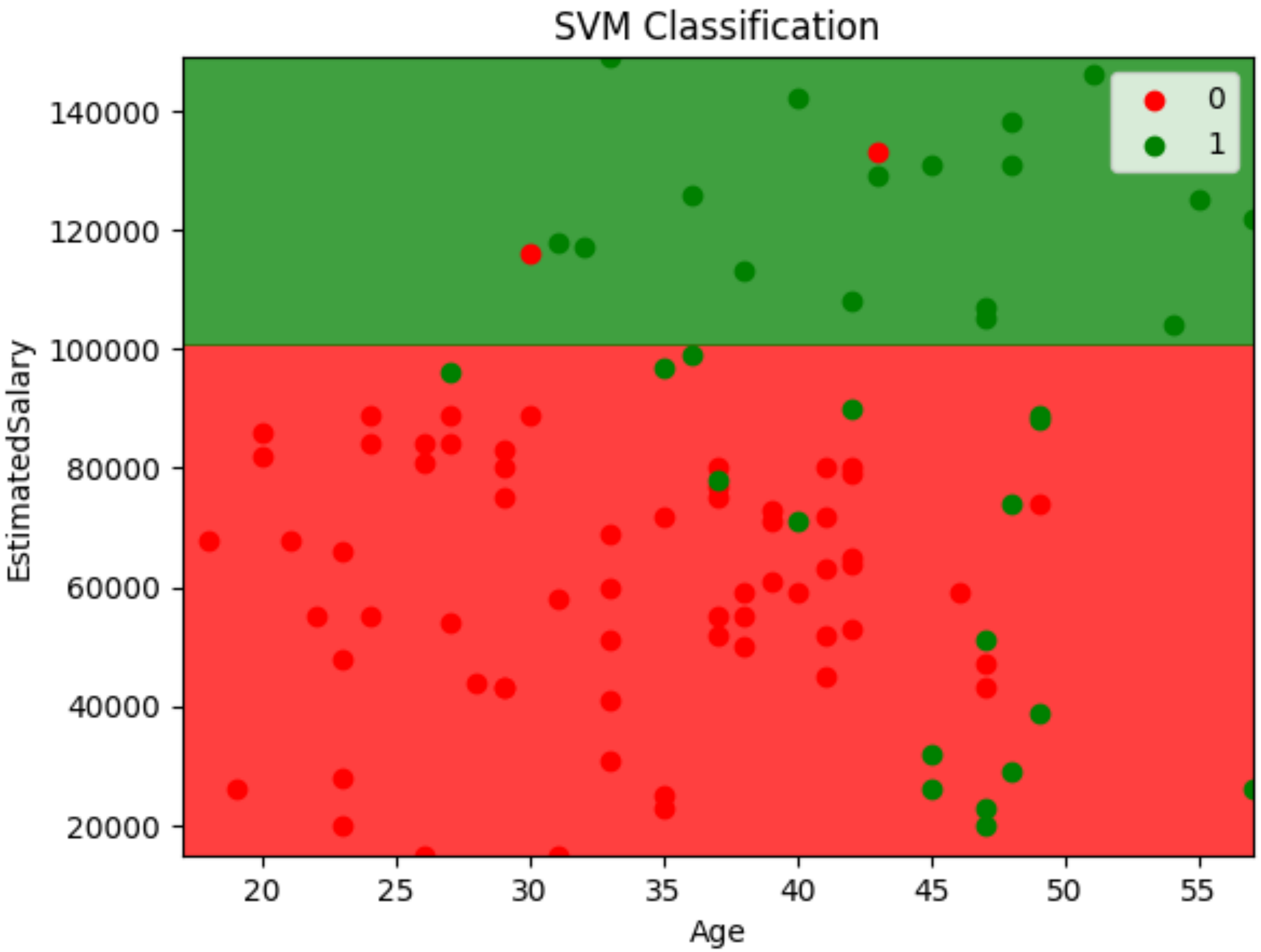
- Supervised learning.

- Example Algorithms: Decision Tree, Naive Bayes

Examples

- Targeted marketing
- Text recognition

Illustration of a Simple Classification Problem



Binary Classification of Salary with Respect to Age

Clustering

- Given a data set (data matrix), partition its objects (rows) into sets (clusters) $C_1, C_2, \ldots, C_k$ such that the objects in each cluster are "**similar**" to one another.

- Specific definitions depend on how the notion of **similarity** is defined.

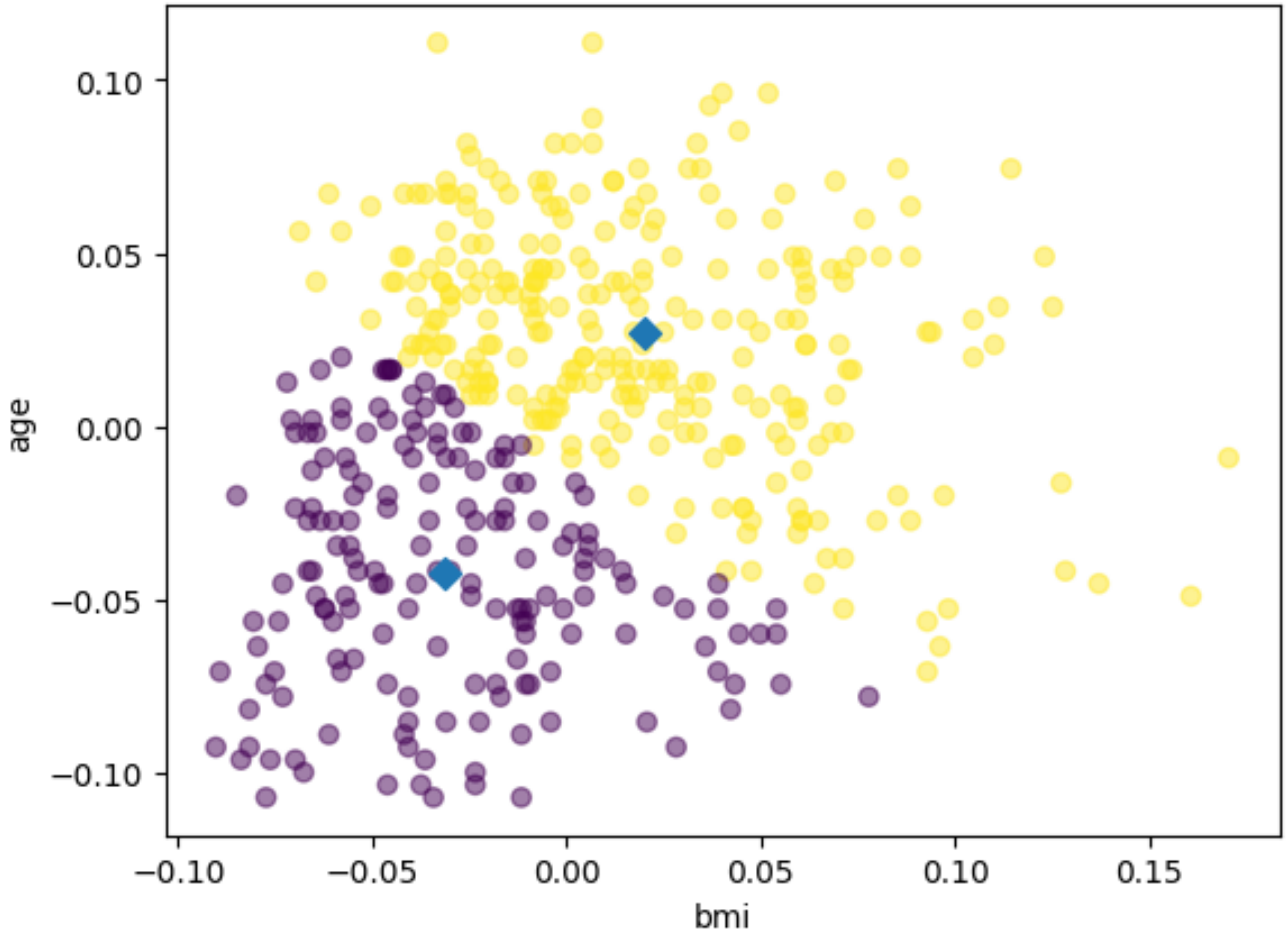
- Can be seen as an unsupervised version of classification.

- Primary objective is to increase intra class similarity and minimise inter-class similarity.

Examples

- Customer segmentation (identify similar customers for targeted product promotion)
- Data summarisation (cluster can be used to create a summary of the data)

Illustration of a Clustering Problem



Clusstring based on BMI and Age of a group of Users

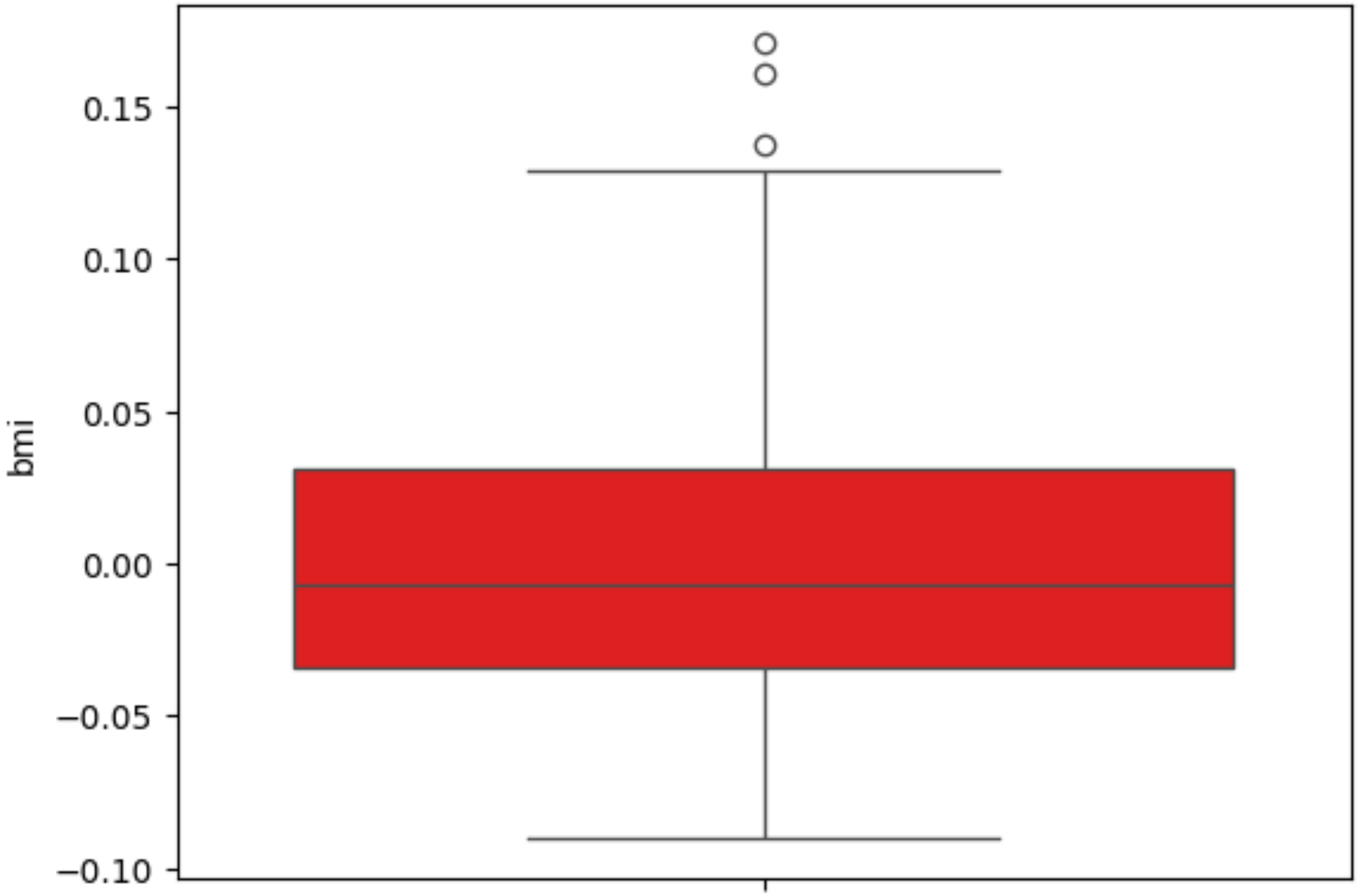
Outlier Detection

- Given a data set, determine the **outliers**, i.e. the objects that are significantly different from the remaining objects.
- It can be noise or exception.

Examples

- Credit card fraud
- Detecting sensor events
- Medical diagnosis
- Earth science

Illustration of Outlier Detection



Outlier Value in BMI