Otto-von-Guericke-Universität Magdeburg
Fakultät für Mathematik
MathCoRe Compact Course WarnemÃ¼nde

2023

Programming exercise dimension reduction

You can find the relevant data at ????

Exercise 0: Simple descriptive statistics

Load the *Boston* dataset into the workspace of the programming language of your choice, look at some simple descriptive plots and basic summary statistics of the data to get a good first impression of the data.

Exercise 1: PCA and linear regression

We have learned that linear regression has some problems when the explaining variables $x_i = (x_{i1}, \ldots, x_{id}$ are highly correlated. Investigate what happens when we want to predict the variable *medv* from the dataset from the remaining variables, to which we applied PCA beforehand. Investigate the performance for different values principal components.

(i) Are the explaining variables still highly correlated?

(ii) How is the performance of the regression where we used principal components as explaining variables compared to the performance of regression using the given variables?

(iii) Is there a theoretical guarantee that principal components with high variance explain the variable we want to predict well?

Exercise 2: Clustering

Load the *Titanic* dataset into the workspace of the programming language of your choice, look at some simple descriptive plots and basic summary statistics of the data to get a good first impression of the data. Create a new object of the dataset without the *survived* column.

(i) Try to naively use the k-means clustering algorithm on the dataset. Does this approach work?

(ii) Apply a suitable transformation to the dataset such that applying k-means is justifiable.

(iii) Does your k-means clustering fit to the survivors?

(iv) Would you have the same problem with agglomerative clustering and mixture models?