Data Analytics 1

Seminar Projects

Prof. Dr. Heike Trautmann

Dipl.-Inf. Jakob Bossek

MSc. Pascal Kerschke

Winter Term 2015 / 2016

Organization

Now it is your turn!

Case Studies

- Data analysis based on suitable methods
- **max. 6** students per group
- **Your work:** 10.12.2015 - 28.01.2016
- **Presentations:** 4.2.2016, 9.2.2016, 11.2.2016, $\approx$ 6 slots (**we expect you to be present at each presentation**)
- **Documented R-Code** until 29.1.2016
- **Final Grade:** case study (40 %), exam (60 %)

Organization

What we expect from you

- Apply suitable data analytics methods to the provided data.
- Use at least one method/algorithm, that was not taught in the course, e.g., use another normality test besides Shapiro-Wilk or an alternative clustering method.
- Perform the data analysis with the R programming language.
  - Document your code!
  - Write down which parts of the R code were ellaborated by whom[a].
- Present your results in a (visually) appealing way.
  - Presentation should take 30 minutes plus $\approx$ 10-15 minutes discussion.
  - Each group member must actively participate in the presentation.

[a] This way we can contact the corresponding person directly if questions come up.

1. LVM isurance data

Background

- Dataset provided by the LVM insurance (`https://www.lvm.de/`).
- LVM insurance is interested in the analysis of the users of their webportal.
- **Your task:**
  - Dataset of 25000 observations with 18 variables (characteristics of the users and product groups).
  - Try to find structures and relationships within the data.
  - Apply as many appropriate techniques learned in DA1 as possible.

2. Travelling Salesperson Problem (TSP)

Background

- The *Travelling Salesperson Problem* (TSP) is one of the most prominent combinatorial optimization problems. Given a set of cities with distances, we aim to find a minimum-costs roundtrip.
- In the area of algorithm selection we try to predict the algorithm which will most likely operate best on a given problem instance.
- Characterization of problem instances by a set of computationally cheap *instance features*.
- **Your task:**
  - Give an introduction to the TSP.
  - Analyze a given set of instance features with the methods taught in the lecture.

Pihera, J., Musliu, N., Application of Machine Learning to Algorithm Selection for TSP, in Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on , vol., no., pp.47-54, 10-12 Nov. 2014

3. Flacco - Expensive Features

Background

- In the context of continuous optimization, researchers try to find the optimum of a problem instance (i.e., a mathematical function) by using only a small amount of function evaluations. Therefore, they use so-called *exploratory landscape analysis (ELA)* features to characterize the problem in order to select the best optimization algorithm for that instance.
- The R package `flacco`[a] provides numerous feature sets measuring different aspects of an instance.
- **Your task:**
  - This project group will analyze a data set containing mainly expensive feature sets (i.e., they need additional function evaluations) of problem instances, which have been created using a random instance generator.
  - Perform a detailed data analysis of the underlying feature data set.

[a] Further information can be found on `https://github.com/kerschke/flacco`.

Mersmann, O., Bischl, B., Trautmann, H., Preuss, M., Weihs, C., Rudolph, G., *Exploratory Landscape Analysis* in Proceedings of GECCO 2011, pp. 829–836.
Kerschke, P., Preuss, M., Hernández, C., Schütze, O., Sun, J.-Q., Grimme, C., Rudolph, G., Bischl, B., Trautmann, H., *Cell Mapping Techniques for Exploratory Landscape Analysis* in EVOLVE 2014, pp. 115–131.

4. Flacco - Cheap Features

Background

- In the context of continuous optimization, researchers try to find the optimum of a problem instance (i.e., a mathematical function) by using only a small amount of function evaluations. Therefore, they use so-called *exploratory landscape analysis (ELA)* features to characterize the problem in order to select the best optimization algorithm for that instance.
- The R package `flacco`[a] provides numerous feature sets measuring different aspects of an instance.
- **Your task:**
  - This project group will analyze a data set containing very promising cheap landscape features of problem instances, which have been created using a random instance generator.
  - Perform a detailed data analysis of the underlying feature data set.

[a] Further information can be found on `https://github.com/kerschke/flacco`.

Mersmann, O., Bischl, B., Trautmann, H., Preuss, M., Weihs, C., Rudolph, G., *Exploratory Landscape Analysis* in Proceedings of GECCO 2011, pp. 829–836.
Kerschke, P., Preuss, M., Wessing, S., Trautmann, H., *Detecting Funnel Structures by Means of Exploratory Landscape Analysis* in Proceedings of GECCO 2015, pp. 265–272.

5. TripAdvisor - Denver

Background

- Nowadays, various hotel search engines – such as TripAdvisor, Booking.com, Expedia, etc. – exist.
- Each of those portals provides a lot of information for each of the hotels for a given location and travel time.
- **Your task:**
  - This project aims at exploring a data set, which contains information on hotels in Denver, Colorado at the end of July, 2016.
  - Analyze the given data set, i.e. detect outliers, analyze the distributions of the features, cluster the hotels, think about dimensionality reduction etc.
  - Note that the data sets might contain missing values. You do not have to worry about imputing those values. Instead, you are allowed to work with reasonable subsamples of the data set.

6. TripAdvisor - Stanford

Background

- Nowadays, various hotel search engines – such as TripAdvisor, Booking.com, Expedia, etc. – exist.
- Each of those portals provides a lot of information for each of the hotels for a given location and travel time.
- **Your task:**
  - This project aims at exploring a data set, which contains information on hotels in Stanford, California around the end of June / beginning of July, 2016.
  - Analyze the given data set, i.e. detect outliers, analyze the distributions of the features, cluster the hotels, think about dimensionality reduction etc.
  - Note that the data sets might contain missing values. You do not have to worry about imputing those values. Instead, you are allowed to work with reasonable subsamples of the data set.

7. TripAdvisor - Münster

Background

- Nowadays, various hotel search engines – such as TripAdvisor, Booking.com, Expedia, etc. – exist.
- Each of those portals provides a lot of information for each of the hotels for a given location and travel time.
- **Your task:**
  - This project aims at exploring a data set, which contains information on hotels in Münster within the first half of March, 2016.
  - Analyze the given data set, i.e. detect outliers, analyze the distributions of the features, cluster the hotels, think about dimensionality reduction etc.
  - Note that the data sets might contain missing values. You do not have to worry about imputing those values. Instead, you are allowed to work with reasonable subsamples of the data set.

8. Travelling Salesperson Problem (TSP)

Background

- The *Travelling Salesperson Problem* (TSP) is one of the most prominent combinatorial optimization problems. Given a set of cities with distances, we aim to find a minimum-costs roundtrip.
- In the area of algorithm selection we try to predict the algorithm which will most likely operate best on a given problem instance.
- Characterization of problem instances by a set of computationally cheap *instance features*.
- **Your task:**
  - Give an brief overview of the state-of-the-art in TSP-solving. Do not lose yourself in technical details.
  - Analyze a given set of instance features with the methods taught in the lecture.

Pihera, J., Musliu, N., Application of Machine Learning to Algorithm Selection for TSP, in Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on , vol., no., pp.47-54, 10-12 Nov. 2014