Introduction to Data Science, WMCS16002, semester 1b 2016

2 Assignment: Presentation

Material	**due November 29, 2016 12:00:00 (noon) CET**
Presentation	**on November 30 during the lecture (9:00-11:00)**

In contrast to the first exercise, which needed quite some coding, this homework will focus on knowledge acquisition and presentation. We take your feedback seriously and students suggested that we refrain from too general and abstract explanation. Therefore, we adapted the topics to problems arising from increasing dimensionality of data, data cleaning and preprocessing in more detail. Note that this is a unique test for a didactic concept trying to integrate you more interactively.

For this assignment each team prepares a presentation on one of the following topics:

1. Sampling
2. Missing value imputation
3. Feature Selection

presentation date to be changed

The presentations will occur during the lecture of November 30th. As there are many teams, presentations will be done in parallel. You can decide if one member of your team presents on behalf of all of you or if you run a "multiple men show".

Your team's topic: Use the following formula to get your topic:

$$n_{\text{topic}} = \mod(n_{\text{team}}, 3) + 1,$$

where n_{topic} is your topic number and n_{team} is your team number.

Schedule:	09:00 - 09:30	Presentations topic 1
	09:30 - 10:00	Presentations topic 2
	10:00 - 10:15	Short break
	10:15 - 10:45	Presentations topic 3

2.1 What we expect for Topic 1

Explain what phenomena are referred to by the term “curse of dimensionality”. This should include

- Combinatorial explosion: the effect that the volume in space increases so fast that data becomes sparse, which leads to difficulties for statistical analysis.
- Distance concentration: with increasing dimensionality pairwise distances may converge to the same value (lack of contrast). Since many data analysis machine learning techniques base on distances this may be problematic.

Think about how you can present this using illustrative examples and synthetic data.

2.2 What we expect for Topic 2

As introduction to the problem, explain the different types of missing data that people can encounter and the difficulties this generates.

Explain the following proposed techniques to deal with missingness:

- Single imputation: replace the missing value using a similar record or mean imputation. This is usually very easy to implement, but might distort relationships between variables.
- Multiple imputation: one of the most recommended methods for imputing missing values.

Think about how you can present this using illustrative examples and synthetic data where you randomly remove dimensions and compare imputation with the ground truth.

Tip: Chapter 25 from the book “Data Analysis Using Regressiona and Multilevel/Hierarchical Models” by Andrew Gelman and Jennifer Hill is a nice introduction to the topic. There’s a link in the Assignments Wiki.

2.3 What we expect for Topic 3

Variable selection can be useful to 1) improve prediction performance, 2) improve cost-efficiency and 3) to learn about the process that generated the data. Explanation should include:

- Filter Methods: those apply a statistical measure to assign scores for ranking the features to be kept or to be removed from the data set (e.g. Correlation or Information Ranking Criteria).
- Usefulness and the limitations of variable ranking techniques.

Think about how you can present this using illustrative examples and synthetic data.

Tip: The article “An Introduction to Variable and Feature Selection” by Isabelle Guyon and André Elisseeff, JMLR 3 (2003) pp. 1157-1182 constitutes a rich overview paper. There’s a link in the Assignments Wiki.

Submit your presentation material by **creating a pull request**.

3 Tip:

Have a look at assignment 3. You may be able to reuse content if you plan wisely.

Presentation Requirements

- Present for 20 minutes
- Question and Answers for 10 minutes

Presentation Format

You are free to choose any format.

Slides are OK, but remember that there will be about 5-6 presentations in parallel so we cannot use beamers. Interactive presentations (with e.g. Jupyter or RMarkdown) are highly encouraged! If you need additional equipment for you presentation (e.g. an external screen, a poster board, a flipchart, etc.) let us know before Monday November 28th by opening an issue: `https://github.com/RUG-IDS/Assignments/issues`.

Grading

You will be graded based on your **presentation material and presentation style**. We kindly ask you to fill in the questionnaires for the presentations that you have attended. We will provide these to you online.

These are the question that you will be assessed with:

- Material (+50P)
  - The content in the presentation was of good quality
  - The content in the presentation supported the main points
  - The content in the presentation was relevant for the topic

- Style (+50P)
  - The presentation contained clear goals and a common thread
  - The speaker conveyed the information in an engaging and attractive manner
  - The selected presentation format was used in an engaging and attractive manner

Good luck and have fun working on the exercises!