Introduction to Data Science, WMCS16002, semester 1a 2017

2 Assignment: Presentation

Material	**due September 17, 2017 23:59:59 (midnight) CET**
Presentation	**on September 18 during the lecture (11:00-13:00)**

In contrast to the first exercise, which needed quite some coding, this homework will focus on knowledge acquisition and presentation. We will provide at least one source of information giving a general overview. It is expected from you that you expand the material by searching for additional relevant literature (following literature referenced in the provided source material to get deeper insight into subparts of the topics counts as well).

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

1. Sampling
2. Missing values
3. Feature Selection

The presentations will occur during the lecture of September 18th. You can decide if one member of your team presents on behalf of all of you or if you run a "multiple men show". Also the style of the presentation is up to you and your inspiration, you can combine different resources, having slides, simulations, videos, etc. and/or involve your audience, e.g. in a Quizz.

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:

11:00 - 11:30	Presentations topic 1
11:30 - 12:00	Presentations topic 2
12:00 - 12:15	Short break
12:15 - 12:45	Presentations topic 3

2.1 What we expect for Topic 1

Explain some basic concepts related to the idea of sampling. This should include

- Traditional statistical sampling (just the basics, it should be a reminder)
- General-purpose sampling strategies (Simple random, stratified, density biased sampling)
- Sampling for Data science (Motivation: mainly 2 uses, embedded by the techniques or running separately before as preprocessing)

  - Bagging and boosting (since this is often used!)
  - Extracting representative subsets of the data to solve certain objective: data summarization via coresets (the basic idea is sufficient, do not get lost in all its beauty). This has been shown to work very well for objectives like clustering and dictionary learning (I see this as one example of an embedded technique. Collect some references so interested people can dive in deeper if they like).

Think about how you can present this using e.g. illustrative examples, synthetic data, simulations etc..

Tip: Chapter 2 from the book "Introduction to data mining" by Tan at al has a very quick overview, while the following is an elaborate survey: F. H. Gaohua Gu and H. Liu. Sampling and Its Application in Data Mining: A Survey. Technical Report TRA6/00, National University of Singapore, Singapore, 2000. Information about corsets can be found in P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximations via coresets. Combinatorial and Computational Geometry - MSRI Publications, 52:130, 2005. There's a link in the Assignments Wiki.

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.
- 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 experiments where you show the problem and differences in the methods, their advantages and disadvantages.

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:

- Explain the phenomena summarized by "curse of dimensionality":
  - Combinatorial explosion: the effect that the volume in space increases so fast that data becomes sparse, which leads to difficulties for statistical analysis and sampling.
  - 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.
- Explain the three standard approaches:

  - 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).
  - Wrapper approaches
  - Embedded approaches (the main idea, advantages and disadvantages, not going into details of the algorithms, that comes later).
- 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.

2.4 Appendix

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

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.

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 Thursday September 14th 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!