Knowledge and the Web 2015/2016: Exercise session 7

November 18, 2015

In this exercise session, you will get familiar with the basics of data mining. First, download the RapidMiner studio, a tool that makes data analysis easy. Then, go over a short tutorial how to use RapidMiner. Go over the tutorial built-in the `RapidMiner` itself (find it on the home screen). For everything else, you have a user guide available here.

Tasks for this exercise session include:

1. **Exploring the data.** Download the `eda.csv` from `Toledo`. Load it in `RapidMiner`. Use `RapidMiner` visualization tools to **look** at the data. Search for interesting patterns. We will discuss what have you found.

2. **Classification.** Download the `titanic_train.csv`, `titanic_test.csv` and `titanic.names` from `Toledo`. Load it in `RapidMiner`. Find a good classification model that predicts will a passenger survive. Focus on simple and interpretable models. Find out what the **parameters of a model mean**. Modify the parameters to see how they effect the performance. **Think of a way how to find out which features are important for the survival**.

3. **Regression.** Download the `OnlineNewsPopularity.csv` and `OnlineNewsPopularity.names` from `Toledo`. Load it in `RapidMiner`. Find a good regression model that predicts the shares of a news item. Again, focus on simple and interpretable models. Find out what the **parameters of a model mean**. Modify the parameters to see how they effect the performance. **Think of a way how to find out which features are important for the given task**.

4. **Clustering.** Download the `dermatology.csv` and `dermatology.names` from `Toledo`. Load it in `RapidMiner`. Exclude class information from your dataset. Find a good clustering model for this task. **Think of a way how to find the right number of clusters in a dataset**.

We will discuss the results of every task during the session (how to find important features, how to identify the right number of clusters,...).

NOTES:

1. each group has to upload a report by **Sunday November 22, 2015**. **Each group** has to choose one of the problems above (classification/regression/clustering). Explain in detail you approach - which algorithms have you tried; which one was the best; your observations about the parameters; which features are important for each of the tasks.

2. there is an **Linked Open Data** extension in RapidMiner. Be sure to check that, might be useful for your project work.

3. No need to handle anything at the end of the session (check the deadline for the report). If you have questions, feel free to email Sebastijan Dumancic (be sure to put [KaW] in the subject line), or ask during the exercise session.