Data mining - Assignment proposal

Selene Baez Santamaria (2572529)
Andrea Jemmett (2573223)
Dimitris Alivanistos (2578740)

April 4, 2016

1 Kaggle

We would like to work with one of the following datasets provided by the platform *kaggle*. After some research we selected two possible datasets that are readily available and do not violate any Intellectual Property regulations.

1.1 Flavours of Physics

The goal of this competition is to find undiscovered phenomena given data from the Large Hadron Collider (LHC). It is a classification task for the variable *signal*, which can have values 0 or 1, given the values for other 50 variables.

The competition provides a Starter Kit, training and test sets, a specific submission format and a couple check files to evaluate the submitted data. The link for these assets is here This competition was closed only five months ago, and it is still relevant in the field.

1.2 Predict Closed Questions on Stack Overflow

The goal of this competition is to determine whether a question in StackOverflow will be closed, given the question details only. It is a classification task.

The competition provides large train and test sets for questions at different points in time, which can be sampled to become more manageable. The link for these assets is here

2 Independent dataset

Alternatively, we propose to use an independent dataset to branch from existing projects.

2.1 StackOverflow

We propose the use of data mining techniques over the StackOverflow database, in order to identify the best answers to explicit questions, thus creating an organized inventory of coding knowledge. We believe that we could use clustering algorithms to create clusters of knowledge based on different features. For example, create clusters according to various programming languages or clusters based on the nature of the question itself (i.e. String Manipulation, Sorting, Debugging techniques).

The dataset is provided as a csv file or a mat file, where each row represents an answer. The variable *tag* corresponds to the primary features we would like to mine. The link to the available databases is provided here.