ADENINE: A Data ExploratioN pipelINE

development plan

Samuele Fiorini

May 19, 2015

1 Introduction and Motivation

A question that arises at the beginning of almost every new data analysis is the following: *are my data relevant with the problem I'm trying to solve*?

The final goal of adenine is to help its user to have a glimpse of the answer of this tedious problem.

In order to reach this goal adenine will make wide use of machine learning ad data mining techniques. The final pipeline will be essentially consist of three steps:

1. **Preprocessing**: have you ever wondered what would have been changed in your problem if only you had centered your data? Or if you would have scaled down all the measures in a certain interval? adenine will present several classical preprocessing procedures, such as: data centering, Min-Max scaling, standardization, normalization, and so on.
2. **Dimensionality reduction**
3. **Clustering**

The final output of adenine will be an as compact as possible visual and textual representation of the results obtained from the pipelines made with each possible combination of the algorithms available for each step. For instance, referring to a pipeline built as:

data normalization + PCA + K-Means

you can find something like:

- an output file containing the norm of the original variables (which has been divided to in order to coerce all the features in $[0, 1]$),
- a 2-D or 3-D scatter plot of the data projected along the 2 or 3 principal components estimated by means of PCA and the percentage of explained variance associated with each component,
- a pictorial representation of the clustering results of the data obtained with the optimum number of cluster (learned from the data).

1.1 Material for PhD progress

The study behind the implementation of adenine will be useful in terms of four PhD courses of my first-year work plan:

1. *A Machine Learning Crash Course* [DIBRIS] (Odone, Rosasco): `adenine` will cover a fair number of (mainly unsupervised) machine learning techniques. Hence, this course has been fundamental to acquire the statistical learning background needed to become aware of the underlying mechanisms of the algorithms.

2. *Programming Concepts in Python* [DIBRIS] (Tacchella): I plan to implement `adenine` in `Python`. Hence, most of the implementation choices will be made on the basis of the material covered in the course.

3. *Programming Complex Heterogeneous Parallel Systems* [IMATI] (Clematis, D'Agostino, Danovaro, Galizia) and the *24th Summer School on Parallel Computing* [CINECA] (Erbacci): `adenine` will present several *embarrassingly parallel workload* as well as several *isolate GPU accelerable* computations. The former PhD course and the latter school will allow me to develop the parallel computing attitude I need to implement `adenine` in an as optimized as possible way.

2 Covered Topics

2.1 Preprocessing

2.2 Dimensionality reduction

2.3 Clustering