`ADENINE`: A Data ExploratioN pipelINE

Development plan

Samuele Fiorini

May 19, 2015

1 ABSTRACT

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