Università di Pisa	A.A. 2015-2016
Data Mining II	

Project assignment

Individual vs. collective purchase behaviours in supermarkets

General information

Objective of this project is to perform a few analyses on a dataset of transactions involving the customers of a supermarket chain. The general guidelines for this assignment are the following:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the processing and analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

The project will be based on real data describing customers and transactions of a set of department stores. The data cover the purchases performed over 12 months, and includes the details of each product sold in each transaction, together with the ID of the customer who performed the transaction (where available). The dataset consists of the following tables, provided as CSV files:

articolo.csv	textual description of the products (in Italian)
cliente.csv	basic information about customers (in Italian)
data.csv	translation table for date coding
marketing.csv	marketing hierarchy of products (in Italian)
venduto.csv	transactions, a line for each product sold

Objectives

The following activities should be performed and reported:

1. **Exploration**: a **short** data exploration phase, aimed at understanding what data can be useful and whether they present any issues or anomalies.
2. For each user, identify his **top 10 most purchased products** on a specific period of the day (e.g. between 17 and 18 everyday) to be decided based on the explorative analysis performed above. Then, compute his distribution of purchases over such products. "Product", here, can stand for single article or any other aggregation level that you decide to adopt, e.g. product category, segment, etc.
3. **Customer segmentation**: based on the purchases distribution of each customer, segment them into homogeneous groups, and try to characterize each segment.
4. **Individual vs. collective**: choose a small number of customers, and find some effective way to compare them to the customer segment they belong to, thus sketching a basic self-awareness service to the user.