Università di Pisa	A.A. 2012-2013
Data Mining II	

Project assignments / Part 1

General information

Objective of this project is to perform the preliminary phases of a data mining process, namely data exploration and preparation, over a real dataset of transactions. 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 department store, belonging to the category "Supermarket". The data cover the purchases performed over 6 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 5 tables, provided as CSV files:

SSET_ART_CORSODM	textual description of the products (in Italian)
SSET_CLIENTE_CORSODM	basic information about customers (in Italian)
SSET_DATA_CORSODM	translation table for date coding
SSET_MKT_CORSODM	marketing hierarchy of products (in Italian)
SSET_VEN_CORSODM	transactions, a line for each product sold

Objectives

Given a specific category "X" of products (level "Settore" of the hierarchy contained in file SSET_MKT_CORSODM), perform the following steps:

1. Data exploration: compute all the basic statistics and distributions considered useful to understand what the data actually contains. That should include also an analysis of which/how many products are contained in each product category (i.e., each "Settore") as well as how much they are actually sold in our transactions. Notice: usually, only a small portion of the products catalog are sold in a shop.
2. Customer profiling: for each customer, compute a set of aggregates that could be useful for a customer segmentation task. Such set should include (at leat) three indicators: (i) monetary volume of purchases made by the customer; (ii) number of transactions (visits) performed; (iii) number of single products bought. These indicators should be computed both considering all products and then focusing only on products of the category "X" mentioned above.
3. Individual events detection: for each customer, compute:
    - (churn analysis) an attribute describing whether (s)he deserted the store – i.e., (s)he is a case of *churn,* and we want to understand when (s)he left the store. The criterion to adopt is the following: for each month compute the number of visits performed by the customer; if (s)he performed none in the last n months with $n>1$, then (s)he is churning;
    - (focusing) an attribute describing whether the purchases of the customer in the last 3 months focused on some particular sub-category of products. In particular, consider only the products of our category "X" mentioned above, and for each sub-category of "X" (level "Reparto" of the hierarchy) compute the percentage of purchases (single products) that belong to it. If this value computed over the last 3 months is more than twice the value obtained over the first 3 months, we will say that the customer is focusing on that sub-category.
4. Time series: choose a (suitable) product of category "X", and compute for each customer a weekly aggregate of the number of purchases of that product. Notice: as an alternative to choose a single product, you can select a "Segmento" (the lower level of the hierarchy, immediately above the single product), which is larger and therefore yields larger numbers in the weekly aggregates.
5. [OPTIONAL] Frequent patterns: start experimenting with tools for extracting frequent itemsets or association rules and sequential patterns.