Università di Pisa	A.A. 2014-2015
Data Mining II	

Project assignment

Entropy and Customer Segmentation

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) and Anna Monreale (anna.monreale@unipi.it).

The dataset

The project will be based on real data describing customers and transactions of a set of department stores, ~~belonging to the category “Supermarket”~~. 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. **Purchases Entropy:** for each customer, the entropy of his/her purchases is computed based on the frequency of purchase of each product sold, normalised with respect to the average frequency of all customers; “product”, in our case, can mean the lowest level of the marketing hierarchy (the value directly reported in the transaction records, i.e. SSET_VEN_CORSODM) or any higher level, at your choice.
3. **Temporal Entropy**: for each customer, the temporal entropy is computed based on the frequency of visits during specific time slots. The two basic approaches consist of monthly time slots (April, May, ...) and week days (Monday to Sunday). Additional definitions of time slots will be welcome (if meaningful).
4. **Spatial Entropy:** for each customer, the spatial entropy is computed based on the frequency of visits in each store.
5. Taking the set of all n-ples <purchase_entropy, temporal_month_entropy, temporal_week_entropy, …, spatial_entropy> generated – one n-ple for each active customer –, study the correlation between the variables and perform a customer segmentation on the basis of these variables. Try to give an interpretation of the groups found.