MAS Data Science and Engineering – DSE 220

The IRI data is set available on the AWS instance. Several additional documents describing the data set and pointing to additional Bibliography are made available (including the white paper: Bronnenberg, Bart J., Kruger, Michael W, and Mela, Carl F. The IRI Marketing Dataset.). Getting to know this data set is time well spent as it will be very applicable towards the Final project. The data set is of medium size and complexity – however understanding the data and how file/folders/datasets related to each other will be of crucial importance for creating an appropriate training data set.
The data set has been purchase by UCSD, it is only for internal use and comes with restrictions. Please read and follow the NDA specified at the very end of the "IRI_database_technical_appendix" document provided with the data set.

Assignment #4

1. **Problem #1 (Can be submitted individually or as part of the Final project team)**

   Using any one or several clustering techniques we have covered in class – determine the groups of "similar" panelist. You will choose what set of attributes to use for determining similarity-explain why you chose them. Provide the evaluation and validation of the produced model(s). Describe in details how you prepared the data, created the training data set, designed features and trained the model (what options, parameters, number of clusters, types of clustering techniques(s) used, etc.)

Final Assignment #4

1. **Problem #1**

Using the IRI marketing data set choose a subset of data that is related to chosen product(s) (or family of products). You can answer several different questions regarding the data from different points of view including the store sales, marketing, advertising, manufacturer or the consumer.

There are several options for this assignment:

1. Create a model of consumer demand to forecast future (on a daily bases) sales of particular product(s) for a subset of stores. Show how these might vary over different (types of) stores. Compare how this prediction might change under (deep) discount/promotion period.

2. Choose one of the groups (clusters from Assignment #3) of panelists and train a model predicting a number of products (and types if data available for those products) purchased that can predict purchase patterns for the (next) week for that group at particular stores. Compare how this prediction might change under discount/promotion period.
3. Create a model that take provides insight into how outside factors such as demographics and marketing promotions affect probability that a consumer (some subset of panelist in the IRI data set that have shopped at a particular grocery store) will purchase a particular product(s) (some subset of consumer good of your choice) based on consumer purchase history.

   Each one of the proposed projects can be adjusted. Each team has the freedom of creativity and flexibility providing an opportunity to alter the specifics (product, panelist, etc. data size and variety) as well as the specific questions asked. Each team needs to ensure data availability for the products chosen for a particular analysis. For example salty snacks and beer provide more information than some other products. These assumptions and decisions made should be documented and reported. This is not a Big Data class – so taking (sufficient) subsets of data is acceptable. Start small with a representative sample and grow the training data set bigger once you get to know the data better.

   Several different algorithms should be used and compared. Please ensure that the team utilizes and compares the predictive power of at least one of the models covered in lecture 5 (ANN, SVM or deep learning) as one of the methods used individually or as part of an ensemble.

   The final assignment has 3 equally important deliverables:
   - **a.** 20 min power point presentation to the class
   - **b.** Written report – following the CRISP-DM methodology
   - **c.** Models created, evaluated and validated (10 fold-cross validation is acceptable, for some analysis using a different 'year' for testing might be an option)

   If you would like to create and use an account on one of our supercomputers please create the XSEDE portal usernames (https://portal.xsede.org/#/guest) and email Nicole Wolter – nickel@sdsc.edu with your username. She will add you to our class allocation and you can use the resources.