Lab Three: Clustering, Association Rules, or Recommenders

CRISP-DM Capstone: Association Rule Mining, Clustering, or Collaborative Filtering

In the final assignment for this course, you will be using one of three different analysis methods:

- Option A: Use transaction data for mining associations rules
- Option B: Use clustering on an unlabeled dataset to provide insight or features
- Option C: Use collaborative filtering to build a custom recommendation system

Your choice of dataset will largely determine the task that you are trying to achieve. Though the dataset does not need to change from your previous tasks. For example, you might choose to use clustering on your data as a preprocessing step that extracts different features. Then you can use those features to build a classifier and analyze its performance in terms of accuracy (precision, recall) and speed. Alternatively, you might choose a completely different dataset and perform rule mining or build a recommendation system.

Dataset Selection and Toolkits

As before, you need to choose a dataset that is not small. It might be massive in terms of the number of attributes (or transactions), classes (or items, users, etc.) or whatever is appropriate for the task you are performing. Note that scikit-learn can be used for clustering analysis, but not for Association Rule Mining (you should use R) or collaborative filtering (you should use graphlab-create from Dato). Both can be run using iPython notebooks as shown in lecture.

- One example of a recommendation dataset is the movie lens rating data: http://grouplens.org/datasets/movielens/
- Some examples of association rule mining datasets: http://fimi.ua.ac.be/data/

Write a report covering in detail all the steps of the project. The results need to be reproducible using only this report. Describe all assumptions you make and include all code you use in the iPython notebook or as supplemental functions. Follow the CRISP-DM framework in your analysis (you are performing all of the CRISP-DM outline). This report is worth XX% of the final grade.

Grading Rubric

Business Understanding (10 points total).

- [**10 points**] Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). How will you measure the effectiveness of a good algorithm? Why does your chosen validation method make sense for this specific dataset and the stakeholders needs?

Data Understanding (20 points total)

- [**10 points**] Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file. Verify data quality: Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems?
- [**10 points**] Visualize the any important attributes appropriately. Important: Provide an interpretation for any charts or graphs.

Modeling and Evaluation (50 points total)

Different tasks will require different evaluation methods. Be as thorough as possible when analyzing the data you have chosen and use visualizations of the results to explain the performance and expected outcomes whenever possible. Guide the reader through your analysis with plenty of discussion of the results.

- **Option A: Cluster Analysis**
    - Perform cluster analysis using several clustering methods
    - How did you determine a suitable number of clusters for each method?
    - Use internal and/or external validation measures to describe and compare the clusterings and the clusters (some visual methods would be good).
    - Describe your results. What findings are the most interesting and why?
- **Option B: Association Rule Mining**
    - Create frequent itemsets and association rules.
    - Use tables/visualization to discuss the found results.
    - Use several measure for evaluating how interesting different rules are.
    - Describe your results. What findings are the most compelling and why?
- **Option C: Collaborative Filtering**
    - Create user-item matrices or item-item matrices using collaborative filtering
    - Determine performance of the recommendations using different performance measures and explain what each measure
    - Use tables/visualization to discuss the found results. Explain each visualization in detail.
    - Describe your results. What findings are the most compelling and why?

Deployment (10 points total)

- Be critical of your performance and tell the reader how you current model might be usable by other parties. Did you achieve your goals? If not, can you reign in the utility of your modeling?
    - How useful is your model for interested parties (i.e., the companies or organizations that might want to use it)?
    - How would your deploy your model for interested parties?
    - What other data should be collected?
    - How often would the model need to be updated, etc.?

Exceptional Work (10 points total)

- You have free reign to provide additional analyses or combine analyses