CIS 335 Data Mining

Semester Project

Due: 4/20/17

Description

The semester project is the culmination of everything that you have learn this semester applied to a task. You are encouraged to work in teams of two for this project but it is very important that both students contribute equally.

For this project, we make the following assumptions. You are working as a consultant to an organization that sells 10 products or services. The organization name, nature of their operation and the products/services will not be specifically given. It is not important to the assignment and I do not want you to have preconceived ideas that will cause confusion. The products will be named prod0 through prod9.

There are three data files:

- custs.txt – this is a table of all 3,297 customers. Attributes include id#, org (g=govt, b=business), prod0 through prod9 (quantity sold), sales (total sales) and daysLate (average number of days between shipment and payment).
- trans.txt – a list of transactions for the year. Each line contains a list of the products bought for that transaction (numbered 0 through 9). Customers only buy one of a product for each transaction. For example, the first line is (6,4,3) so that one each of product 6, 4 and 3 are included on this transactions.
- product.txt – a file containing (for each product) the cost of the product, the price of the product and the number sold for the year. You can calculate the profit by subtracting the cost from price.

Your contact is the CEO (chief executive officer) Dot A. Myning. Ms. Myning suspects that using data mining will help her with some questions she wants answered having to do with their marketing strategy. She has called you (and your partner) in to help answer the questions. Specifically they want to know:

1. how should they segment their customers into groups (to focus their marketing efforts)
2. how to identify customers that may pay late (so they can be given an incentive to pay on time). You should provide a classifier and a general idea of what sort of customers pay late.
3. the best way to give discounts to increase sales. They want to have a strategy where they give customers discounts on one product to increase the sales of another product.

Instructions

You and your partner will need to work through the following steps to complete the project:

1. Explore the data. Use the summary statistics and visualization tools we learned about to do some initial exploration. Consider such things as outliers, discretization, creating new columns, conversions.
2. For this project, you are allowed to ask questions of your old professor, Dr. Scripps. He will answer these questions collectively in class but you must submit a list of questions to ask. Any submitting more than 5 questions will risk upsetting the good nature of the old professor.
3. Do the analysis. This can include clustering, classification, and association analysis. You might feel that you want to try several different classifiers to see which provides the best results. Consider using 10-fold cross validation to evaluate the classifiers.
4. Make conclusions and write up the report. It is important to report the results of the analysis as plainly and directly as possible. It is also important to interpret the results for the reader. You can report findings such as "classifier X predicted new instances with a precision of 85%" but what does that mean? Will it lead to an improvement in some process? What is the recall? Which is more important in this study and why? Make the conclusions clear for your reader.

Your report is to be 2 pages, maximum. If you want to provide supporting graphs or tables, make them small and clear.

Rubric:

	points
questions	2
preliminary analysis	3
answer to question 1	4
answer to question 2	4
answer to question 3	4
overall analysis	4
professionalism (clarity, formality, persuasion)	3
on-time	1
total	25