Data Mining Project Description

For our final project we would like to investigate how data mining models can be used to predict the dollar value of the next grocery trip. Grocery chains typically give out incentive coupons (e.g. $5 off your next basket of $70 or more). However, having a reasonable assessment of the size of a customer’s next purchase would considerably improve the targeting – and response – to such initiatives.

Our data catalogs customers’ loyalty card purchases at 12 grocery stores over a period of 2 years. Because most customers are repeat purchasers, but purchase at different, random times, the data set is an unbalanced panel. The variables include the dollar amount spent during each transaction, the number of different kinds of products purchased, and demographics describing the customer’s household. We have already done some work in constructing this data set which now consists of 528 loyalty card numbers, and comprises approximately 20,000 transactions (trips to a given store on a given date). By partitioning this data into training and testing sets, we hope to use different data mining methods to develop several alternate predictions of how large the customers “next” basket will be.

We will begin our analysis by first building a linear regression model to predict the dollar amount spent per “next” transaction. We will then go on to develop several nonparametric models and use them to predict transaction size. The nonparametric methods used will include Support Vector Machine Classification, K-Nearest Neighbors, and Neural Nets. In order to use nonparametric models we will transform the dollars spent per transaction into a discrete variable. The discrete variable will be a dummy variable indicating whether the amount purchased is larger than a certain threshold. At this point, we may also choose to include a logistic model.

The key difficulty will be in determining the optimal location of the threshold amount of dollars spent per transaction. Several plausible thresholds will be analyzed in order to determine the optimal threshold value. Moreover, we expect that some thresholds will be easier to predict than others; however, we will have some discussion of which thresholds are useful and which allow for fairly accurate predictions (i.e. it is of no use to predict which baskets are greater than $150, since this represents a very small proportion of purchase occasions, even if the predictions are very accurate!).

The final results of our analysis will be written up as a paper and presented in class.