

Data Mining – Spring 2020 – Assignment 3

Question 1: Association Analysis

A drug store chain wants to learn more about cosmetics buyers purchase patterns. Specifically, they want to know what items are purchased in conjunction with each other, for purposes of display, point of sale special offers, and to eventually implement a real time recommender system to cross-sell items at time of purchase. The data (in the file *Cosmetics.jmp*) are in the form of a matrix in which each column represents a product group, and each row a customer.

Conduct an association analysis of the data set, and identify five rules that you would deem interesting and/or useful. For the rules that you identify, explain how the reported support, confidence, and lift values are calculated.

Question 2: Prediction using Neural Nets

Car Sales. Consider again the data on used cars (*ToyotaCorolla.jmp*) with 1436 records and details on 38 attributes, including Price, Age, KM, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

- a. Determine which variables to include, and use the neural platform in JMP Pro to fit a model. Create a validation column for validation using a random seed of 4279, and use the default values in the Neural model launch dialog. Record the RMSE for the training data and the validation data, and save the formula for the model to the data table (use the Save Fast Formulas option, which will save the formula as one column in the data table). Repeat the process, changing the number of nodes (and only this) to 5, 10, and 25.
 - i. Using your recorded values, what happens to the RMSE for the training data as the number of nodes increases?
 - ii. What happens to the RMSE for the validation data?
 - iii. Comment on the appropriate number of nodes for the model.
 - iv. Use the Model Comparison platform to compare these four models (use the Validation column as either a By variable or as a Group variable, and focus only on the validation data). Here, RASE is reported rather than RMSE. Compare RASE and AAE (average absolute error) values for these four models. Which model has the lowest “error”?
- b. Conduct a similar experiment to assess the effect of changing the number of layers in the network as well as the activation functions.