Lecture 7 Homework

Due Date: Thursday Nov. 20th 2014

Data Preparation Assignment

Consider the data collected by a hypothetical video store for 50 regular customers. This data consists of a table which, for each customer, records the following attributes: Gender, Income, Age, Rentals (total number of video rentals in the past year), Avg. per visit (average number of video rentals per visit during the past year), Incidentals (whether the customer tends to buy incidental items such as refreshments when renting a video), and Genre (the customer's preferred movie genre). This data is available as an Excel spreadsheet on our class website.

Perform each of the following data preparation tasks:

a. Use **smoothing by bin means** to smooth the values of the **Age** attribute. Use a **bin depth of 4**.
b. Use **min-max normalization** to transform the values of the **Income** attribute onto the range **[0.0-1.0]**.
c. Use **z-score normalization** to standardize the values of the **Rentals** attribute.
d. **Discretize** the (original) **Income** attribute based on the following categories: High = 60K+; Mid = 25K-59K; Low = less than $25K.
e. Convert the original data (not the results of parts a-d) into the **standard spreadsheet format** (note that this requires that you create, for every categorical attribute, additional attributes corresponding to values of that categorical attribute; numerical attributes in the original data remain unchanged).
f. Using the standardized data set (from part e), perform basic correlation analysis among the attributes. Discuss your results by indicating any strong correlations (positive or negative) among pairs of attributes. You need to construct a complete Correlation Matrix (Please read the brief document Basic Correlation Analysis (see course website) for more detail). Can you observe any "significant" patterns among groups of two or more variables? Explain.
g. Perform a cross-tabulation of the two "gender" variables versus the three "genre" variables. Show this as a 2 x 3 table with entries representing the **total counts**. Then, use a graph or chart that provides the best visualization of the relationships between these sets of variables. Can you draw any significant conclusions?
h. Select all "good" customers with a high value for the Rentals attribute (a "good customer is defined as one with a Rentals value of greater than or equal to 30). Then, create a summary (e.g., using means, medians, and/or other statistics) of the selected data with respect to all other attributes. Can you observe any significant patterns that characterize

this segment of customers? Explain. Note: to know whether your observed patterns in the target group are significant, you need to compare them with the general population using the same metrics.

i. Suppose that because of the high profit margin, the store would like to increase the sales of incidentals. Based on your observations in previous parts discuss how this could be accomplished (e.g., should customers with specific characteristics be targeted? Should certain types of movies be preferred? etc.). Explain your answer based on your analysis of the data.

Use your favorite machine learning/data mining tool, such as WEKA, Excel or scripting to perform the following tasks on the original data set. If using WEKA, load the data into WEKA Explorer (the Preprocessing module). Remove the Customer ID attribute. Review basic statistics for different attributes by clicking on the name of each one in "attribute" panel. Next, use the unsupervised attribute "Discretize" filter to discretize the Age attribute. Finally, use the unsupervised attribute "Normalize" filter to convert all of the remaining numerical attribute into [0…1] scale. Save the resulting data set into an ARFF formatted file and submit with your answers for the above questions.

Note: You can give the final results of parts (a) through (d) as a single table which includes the original data and has an added column for each of the parts (a) through (d). The results of part (e) should be a separate table. For the correlation analysis (part f) give your correlation matrix (rows and columns of the matrix are the attributes, and entries would represent correlation value for a pair of attributes (e.g., "Income" versus "Age"). Your analyses for various parts can be added to the same spreadsheet file, or it could be included in another document (e.g., an MS Word or PDF file). Submit via course website online.

SVM Assignment (optional)

For this exercise, we apply SVM with several different kernels and hyperparameter choices to the veh-prime.arff file. As a first step you will need to modify the .arff file so that the car, noncar classes are re-placed with 1 and -1 respectively. Import this new file into Weka and then select the SMO classifier found under classifiers/function. Use 10 fold cross-validation (this should come up as the default).

You can make kernel and hyperparameter choices by clicking on "SMO ....." appearing next to Choose. You will make 5 runs of the algorithm. Select PolyKernel with exponent options 1, 2 and 4. Then select RBFKernel with gamma set to 0.01 and 1.0. For each run record the number of correctly and incorrectly classified instances.

If some of the choices do not work well, explain why you think this is the case.

public class **SMO**
extends DistributionClassifier
implements OptionHandler
Implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial kernels. Transforms output of SVM into probabilities by applying a standard sigmoid function that is not fitted to the data. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. For more information on the SMO algorithm, see

J. Platt (1998). *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. Advances in Kernel Methods - Support Vector Learning,

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy (2001). *Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 13(3), pp 637-649, 2001.*

Note: for improved speed normalization should be turned off when operating on SparseInstances. Valid options are:

-C num
The complexity constant C. (default 1)

-E num
The exponent for the polynomial kernel. (default 1)

-N
Don't normalize the training instances.

-L
Rescale kernel.

-O
Use lower-order terms.

-A num
Sets the size of the kernel cache. Should be a prime number. (default 1000003)

-T num
Sets the tolerance parameter. (default 1.0e-3)

-P num
Sets the epsilon for round-off error. (default 1.0e-12)