Lab Assignment 1

CIS 660 Data Mining

Sunnie Chung

The Marketing department of Adventure Works Cycles wants to increase sales by targeting specific customers for a mailing campaign. The company's database contains a list of past customers and a list of potential new customers. By investigating the attributes of previous bike buyers, the company hopes to discover patterns that they can then apply to potential customers. They hope to use the discovered patterns to predict which potential customers are most likely to purchase a bike from Adventure Works Cycles.

From the data in a view `vTargetMail` from AdventureWork data warehouse, you can see all the customer data in the view by `select * from vTargetMail;`
You may directly use the input file given on the class webpage.

From the vTargetMail, Select a set of all the attributes that would affect to predict future bike buyers and all the necessary info to construct an email list for the future bike buyers. Create a new view VTargetBuyerMailList with all the necessary data of your selections.

Part 1: Feature Selection, Cleaning, and Preprocessing to Construct an Input from Data Source

1. Examine the values of each attribute and Select a set of attributes only that would affect to predict future bike buyers to create your input for data mining algorithms. Remove all the unnecessary attributes from the view (file) VTargetBuyerMailList.
2. Create a new view (or file) VTargetBuyers with the selected attributes only.

Export all the data in the view **VTargetBuyers** to an Excel file and Create a CSV file and a txt file. (If you don't know how to do this, see below)

http://sqlbak.com/blog/export-sql-table-to-excel/
https://msdn.microsoft.com/en-us/library/ms140052.aspx

3. Determine a Data value type (Discrete, or Continuous, then Nominal, Ordinal, Interval, Ratio) of each attribute in your selection to identify preprocessing tasks to create input for your data mining.

Part 2: Data Preprocessing and Transformation

For each selected features (attributes) in Part1,

1. Determine which data preprocessing methods and transformation techniques listed below should be done in a **sequence** depending on the attribute value type that you identified in Part 1.
2. Perform the data preprocessing/transforming tasks for each feature.
3. Transform each objects from your preprocessed data using Banarization (One Hot Encoding) to calculate dissimilarity/similarity distance in Part 3.

Use all the data rows (~= 18000 rows) with the selected features as input to apply all the tasks below, do not perform each task on the smaller data set that you got from your random sampling result.

- Handling Null values
- Random Sampling
- Mean/Variance/Standard Deviation for Ordinal Numeric attributes
- Normalization
- Standardization
- Discretization (Binning/Histogram) on Continuous attributes or Categorical Attributes with too many different values
- Get Median of the Grouped data (those attributes that were transformed into Histograms)
- Binarization (One Hot Encoding)

Write a program with your choice of any language or tools to preprocess the data to create your output files in xls or CSV, or .txt file.

You can use any data preprocessing tool of your choice. The common data preprocessing tools for are Python, R, Weka, or MatLab. See a list of common data analytic tools available at the end of this lab specification.

You can also program in your choice of any language.

Make your lab report in .doc file explaining your data processing platforms, your data preprocessing steps with each output in Screenshots (your system should be shown in the screenshot) with your source scripts (codes) and all the outputs.

Part 3: Calculating Proximity of Two Binary Object Vectors
With Simple Matching, Jaccard Similarity, Cosine Similarity

1. Calculate Similarity in Hamming Distance (Simple Matching), Jaccard Similarity, and Cosine Similarity between two objects - CustomerKey: 11000 and CustomerKey: 11001 of your transformed input data from Part2-3 One Hot Encoding.

2. Calculate Similarity in Hamming Distance(Simple Matching), Jaccard Similarity, and Cosine Similarity between two objects – CustomerKey: 11000 and CustomerKey: 11012 of your transformed input data from Part2-3 One Hot Encoding.

Extra Credit:

Part 4: Correlation Analysis by Building a Correlation Matrix

1. Build a Correlation Matrix for Every Pair of the Features for the entire Data Set
2. Divide your data set into two data sets: One with Bike Buyer = 1 and the other set with Bike Buyer = 0
3. Then build a Correlation Matrix for Every pair of the Features for the record set with Bike Buyer = 1 and the other record set with Bike Buyer = 0 respectively.
4. From three correlation matrix built from 1 and 3 above, Compare the Correlation values between two features Age and Yearly Income with the Correlation between two features Commute Distance and Yearly Income.
5. Compare and Discuss which two features are correlated more strongly than the others for each data set from 1 and 3.

Report:

Show each step for Part 1 and Part 2 in screen captures of the process and the output in each step) and explain BRIEFLY (and clearly) about what you did and why.

For Part 3 and 4: Make outputs with each Screenshot of your scripts (codes) and output in . doc file and explain each step briefly. You will lose points if you don't show and explain each step properly.

Submission:

1) Submit your zip files with all your input and output files, source codes and your report (in .doc file) on Blackboard
2) Turn your printout of your report explaining each data processing steps with your source codes, output in class.

Statistics and Data Analytic Tools

This lab is to practice data preprocessing and transformation with the data from real life in your customized way. Lab1 is also to make you explore different tools and think over the options. You may use your choice of any common data analytic tools available, Or you can write scripts or programming for each task.

R:
http://www.rdatamining.com/
https://www.rstudio.com/products/rstudio/download/#download
http://web.cs.ucla.edu/~gulzar/rstudio/basic-tutorial.html
http://www.inside-r.org/

Documentation
https://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

Python scikit- learn data Preprocessing :

http://scikit-learn.org/stable/modules/preprocessing.html#preprocessing

Weka : Data Mining Software in Java

http://www.cs.waikato.ac.nz/~ml/weka/
https://www.cs.waikato.ac.nz/ml/weka/downloading.html

MATLAB:
Download
https://www.mathworks.com/campaigns/products/ppc/google/matlab-b.html?s_eid=ppc_29850095842-matlab-b&q=matlab

Documentation
http://www.mathworks.com/help/matlab/

Quick Tutorial
http://www.mathworks.com/support/learn-with-matlab-tutorials.html?requestedDomain=www.mathworks.com

Excel:

How to make Add In Analysis ToolPak in Excel: (It may vary slightly different depends on your version and product. Find out in MS Support sites as below)

File -> Open -> Click the file -> Property -> Click Add-In in left pane -> in the Manage group drop down box in the bottom-> Excel Add-in, then go -> check Analysis ToolPak and Add-In available
https://support.microsoft.com/en-us/kb/214269

Excel has a lot of Analysis add in functions to do all the basic data transformation. However, the functions provided there are limited. Sometimes you need to process data in your way that is not available in Excel. For example, an equal frequency histogram or random number generation with the same sample sequence again, ... so on.