# Week 4 Weka Assignment (75 points)

## **Assignment Due Date and Time**

* Sunday at 11:59 p.m. ET.

## **Purpose**

The purpose of this assignment is to get experience processing data using MS Excel and Weka.

## **Description**

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**.

1. You will use MS Excel (not Weka) to explore the general characteristics of the data, by computing the means and standard deviations of the numerical attributes, as well as the distributions of male and female customers, the preferred movie genres, etc.
2. Perform the following data preparation steps on the data (for each add a new column to the original table for comparison purposes)
   1. Use **smoothing by bin** means to smooth the values of the **Age** attribute. Use a **bin depth of 4**.
   2. Use **min-max normalization** to transform the values of the Income attribute onto the range [0.0-1.0].
   3. Use **z-score normalization** to standardize the values of the Rentals attribute.
   4. **Discretize** the (original, non-normalized) Income attribute based on the following categories: High = 60K+; Mid = 25K-59K; Low = less than $25K.
   5. **Save** your revised spreadsheet at **Step2.xls**.
3. Convert the original table (not the results of Step 2) into the standard spreadsheet format. Note that this requires converting each categorical attribute into multiple attributes (one for each values of the categorical attribute) and assigning binary values corresponding to the presence or not presence of the attribute value in the original record). For example, the Gender attribute will be transformed into two attributes, "Genre=M" and "Genre=F." The numerical attributes will remain unchanged. This process should result in a new table with 12 attributes (one for Customer ID, two for Gender, one for each of Income, Age, Rentals, Avg. Per Visit, two for Incidentals, and three for Genre). **Save** your spreadsheet as **Step3.xls**.
4. Using the standardized data set (from Step 3), perform a basic correlation analysis among the attributes. If you are unsure how to do this, refer to the Basic Correlation Analysis Guide. You may need to install the **MS Excel Data Analysis Toolpak**. Analyze your results by indicating any strong correlations (positive or negative) among pairs of attributes. You need to construct a complete **Correlation Matrix**. Be sure to first remove the Customer ID column before creating the correlation matrix. **Save** your spreadsheet as **Step4.xls**.
5. Use **WEKA** to perform the following tasks on the comma separated, data set, “**Video\_Store.csv**.”

## Load the data into WEKA Explorer (the Preprocessing module).

## **Remove** the Customer ID attribute. Select the attribute “Cust ID” and click on the “Remove” button that will show below the list in the “attribute” panel.

## Review basic statistics for different attributes by clicking on the name of each one in "attribute" panel.

## Use the unsupervised attribute "**Discretize**" **filter** to discretize the Age attribute.

## 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.

## **Submission Instructions**

Upload your **Step2.xls**, **Step3.xls**, **Step4.xls** and Weka **ARFF** files into the Week 4 Weka Assignment submission area. Note: you need to submit all four files for grading.