Univariate Binning with Normal Mixtures

version 1.0

Author: Sam Gardner, JMP/SAS Institute

Date: 22 April 2013

Disclaimer:

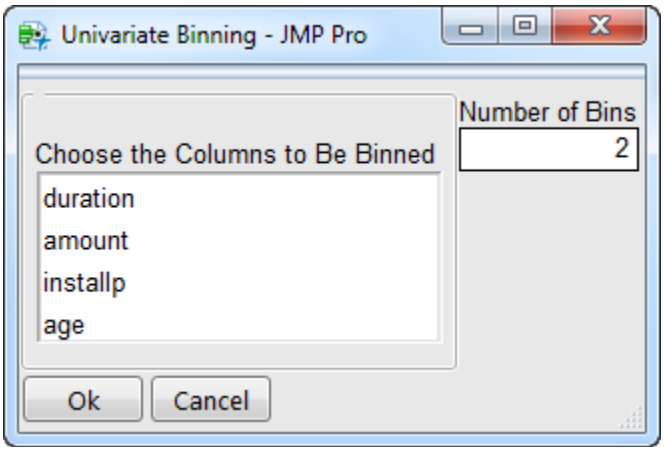
This addin is provided "as is" by SAS Institute Inc. There are no warranties, expressed or implied, as to merchantability or fitness for a particular purpose regarding the materials or code contained herein. The Institute is not responsible for errors in this material as it now exists or will exist, nor does the Institute provide technical support for it.

- Requires JMP 10.0.1 or above
- Creates a binned version of one or more continuous variables
- Use the Normal Mixtures Distribution fitting facilities in the Distribution Platform
- Operationally, the script works as follows:
    - Dialog is presented for the user to choose which columns are to be binned
    - and to select the number (k) of bins.
    - For each column selected
        - Using the Distribution platform, a Normal (k) Mixtures distribution fit is performed (where k is the number of bins entered in the initial dialog), and the parameter estimates for the fitted distribution are scraped from the output report.
        - The fitted distribution parameters are used to build a posterior probability formula for each group/distribution in the normal mixture
        - Rows are assigned to the group with the maximal posterior probability
        - Rows are given a bin number that is ordered with respect to the original variable (that is, a higher bin number is indicative of a higher value on the original data).
        - The bin prediction assignment formula is saved as a new column in the data table. This allows the binned variable to be used in further analysis, but retains the relationship to the original variable, which is useful for model scoring and profiling.

Usage:

1. Launch the Add-in (Addins > Univariate Binning with Normal Mixtures).

   A dialog with a list of the Continuous Variables in the data table will be presented



2. Choose the columns you wish to perform binning on (you may select more than one).
3. Enter the number bins to create.
4. Click Ok.

Methodology:

This addin utilizes the Normal Mixture distribution fitting that is available in the Distribution platform in JMP.

The probability density function for a Normal Mixture distribution has the form

$$f(x) = \sum_{i=i}^{K} \pi_i \phi((x - \mu_i) / \sigma_i)$$

where

$f(x)$ is the probability density function

K is the number of distributions in the mixture

π_i is the weight for the ith normal distribution in the mixture. This can be thought of as being the probability that an observation comes from the ith normal distributin in the mixture.

$\phi(u)$ is the standard normal probability density function $\left(e^{-u^2/2} / \sqrt{2\pi}\right)$

μ_i is the mean for the ith normal distribution in the mixture

σ_i is the standard deviation for the ith normal distribution in the mixture

For a given value of *K*,maximum likelihood estimation is performed to determine estimates $(\hat{\pi}_i, \hat{\mu}_i, \hat{\sigma}_i)$ of (π_i , μ_i , σ_i), and these estimate are used to build a decision formula

$$\text{bin}(x_m) = \arg\max_i \{\hat{\pi}_i \phi((x_m - \hat{\mu}_i)/\hat{\sigma}_i)\}$$

which assigns row m to the group (or bin) which has the maximum probability.

Known limitations:

Fitting normal mixture distributions for large dataset may take a large amount of computational time. One way to speed up the computation is to round the data to a reasonable level of precision, so that the number of unique levels in the column is smaller.

There is a minimum number of observations required to estimate any given normal mixture. Because there are 3 parameters that characterize each distribution in the normal mixture, a Normal K Mixture distribution has 3*K parameters and requires at least 3*K + 1 unique levels in the column. If a highly discretized variable is selected for binning with a large number of bins, the addin may give an error message and halt execution on that variable. You can close the status window for the binning progress and restart the binning on columns that were not binned prior to the halt.