Normalizing and Binning Continuous Variables

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NORMALIZATION

Overview

- Also referred to as “scaling” a variable
- Applies to numeric variables only (usually continuous)
- Essential as part of data engineering
- Various ways of performing normalization

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NORMALIZATION

Min-max normalization method

- Often called feature scaling (https://en.wikipedia.org/wiki/Feature_scaling)
- Involves rescaling the variable from 0 and 1
- Is often favored because the range is always the same.
- Is strongly affected by outliers

NORMALIZATION

Z-normalization method

- Also referred to as standardization
- Ideal for variables following the normal distribution
- Involves changing the variable so that its mean is equal to 0.0 and its standard deviation equal to 1.0
- Outliers affect the overall normalization to a lesser extent

NORMALIZATION

Useful considerations when normalizing a variable

- Combining (linear) normalization methods is unnecessary, since it's just the final normalization that matters
- Binary variables can be normalized too, but in the case of min-max normalization it's unnecessary
- Variable values become comparable if one uses the same normalization method for all normalizations in a dataset
- When normalizing based on a sample, it is best to use the same values of min/max or μ/σ when you normalize the rest of the values of the variable
- Normalization can be reversed, if you have kept the parameters used for it

BINNING

Overview

- Involves grouping values of a numeric variable together and substituting them with a single value, usually a category
  - Groups = bins
- Loses part of the signal in the original variable
- Useful for replacing a continuous numeric variable with a categorical variable
  - Boundaries of each bin can be predefined or selected automatically

BINNING

Standard binning method (Equal-width binning)

1. Define the number of bins (N)
2. Find the bin width: W = (max(x) - min(x)) / N
3. For each bin:
    1. Calculate the boundaries low, high
    2. Find all the data points in x belonging to [low, high]
    3. Assign a unique bin label to these points

BINNING

Binning and histograms

Histograms are great for depicting what a variable's distribution looks like:

- Oftentimes, a variable's histogram may help setting binning limits
- The numpy histogram function can be used to determine boundaries: Try: plt.hist(x)

BINNING

Useful considerations when binning a variable

- Selecting an appropriate number of bins is very useful for meaningful results
- Usually various scenarios are tried before committing to a single one
- Binning is not reversible as a process

SIMPLE STATISTICS

Python functions and classes

- Normalizing: *sklearn* package, *preprocessing* class, *StandardScaler* and *MinMaxScaler* functions
- Binning: *numpy* package, *histogram* function
- Comparison of various normalization methods in Python: http://bit.ly/2hty6M4

Summary

- Normalization
  - Numeric to Numeric
  - Shifts and sets the scale
  - Reversible
- Binning
  - Numeric to Categorical
  - Sets a categorical label
  - Irreversible