DATASCI 510 Lesson 6

Feature Engineering: Normalization

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Reflections

Scientists study the world as it is; engineers create the world that has never been.

Theodore von Karman, Hungarian-American mathematician, aerospace engineer, and physicist

Presentation of the National Medal of Science to Theodore von Kármán by President Kennedy
Photo by Abbie Rowe - JFK Library

Lesson 6 Agenda

- Lesson_06_Slides.pdf
  - Review
  - Normalization in Clustering
- Jupyter Lab (Lesson_06_Normalization.ipynb)
  - Linear Normalization Methods
  - Min-max-normalization *vs.* Z-normalization
  - Use scikit-learn (sklearn) for Normalization
  - Compounding Linear Normalizations
  - De-normalize
  - Apply normalization to other data
- Normalization in Python and Assignment (Lesson_06_student.ipynb)
- Interview question

Normalizing Continuous Variables

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

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

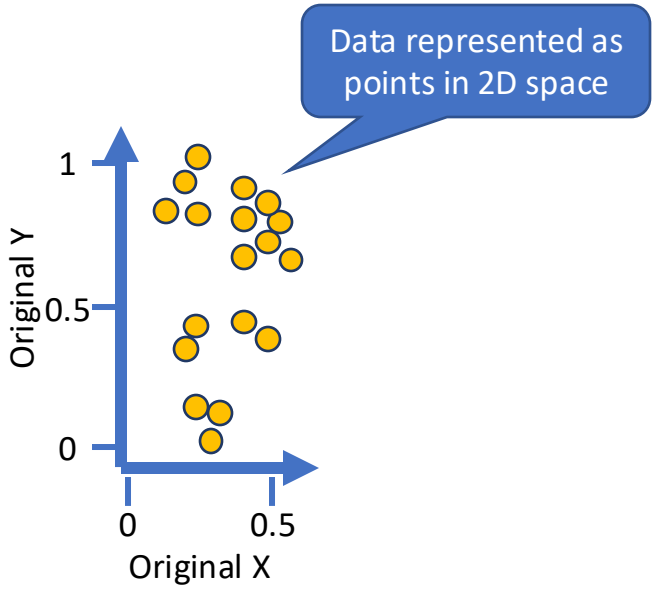
Summary

- Normalization
  - Numeric to Numeric
  - Shifts and sets the scale
  - Reversible

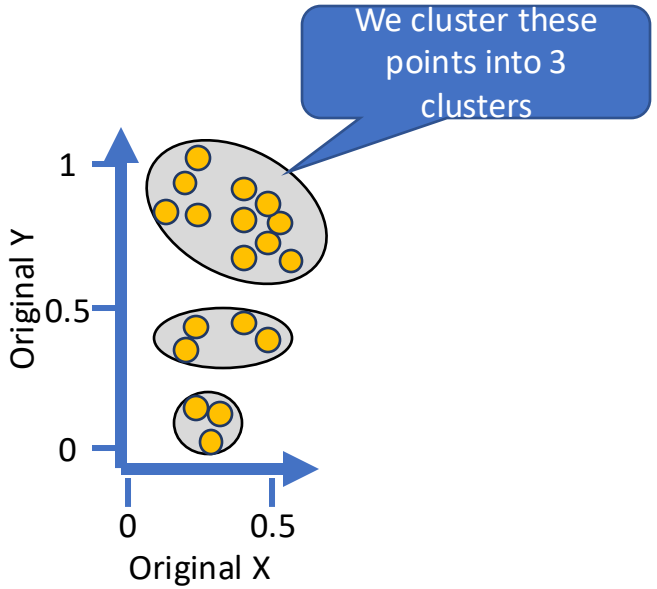
- Python functions and classes
  - Normalizing: *sklearn* package, *preprocessing* class, *StandardScaler* and *MinMaxScaler* functions
  - Comparison of various normalization methods in Python: http://bit.ly/2hty6M4

Normalization in Clustering

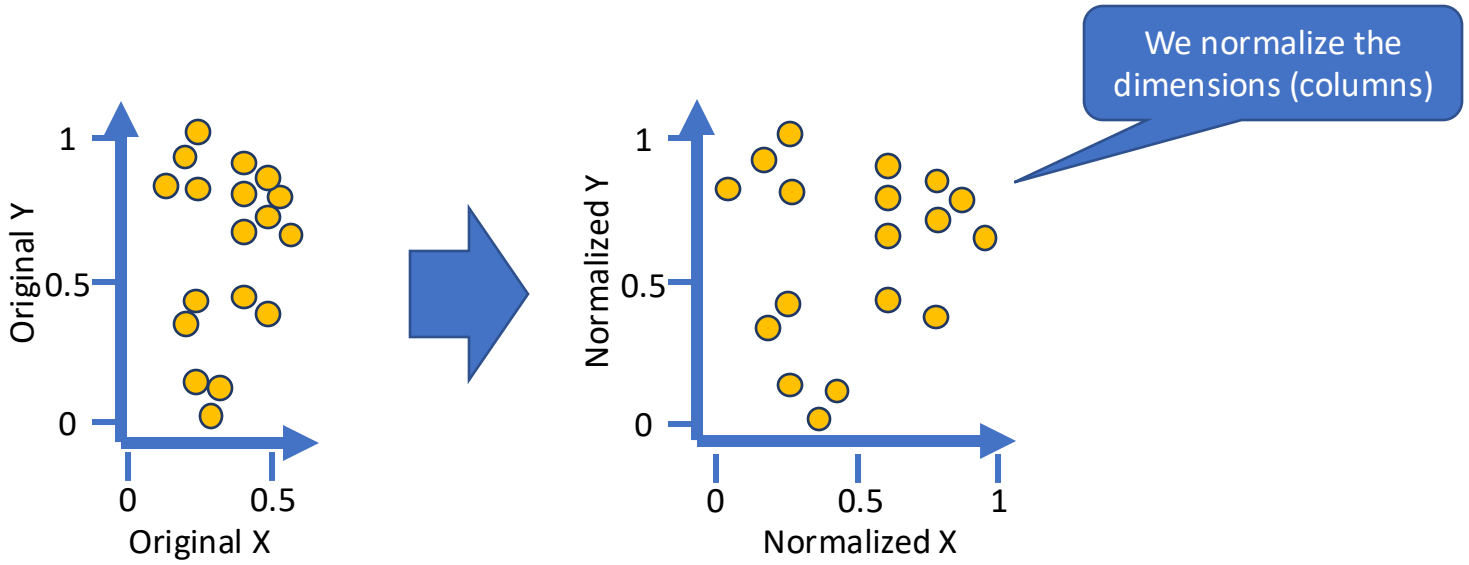
Effect of Normalization



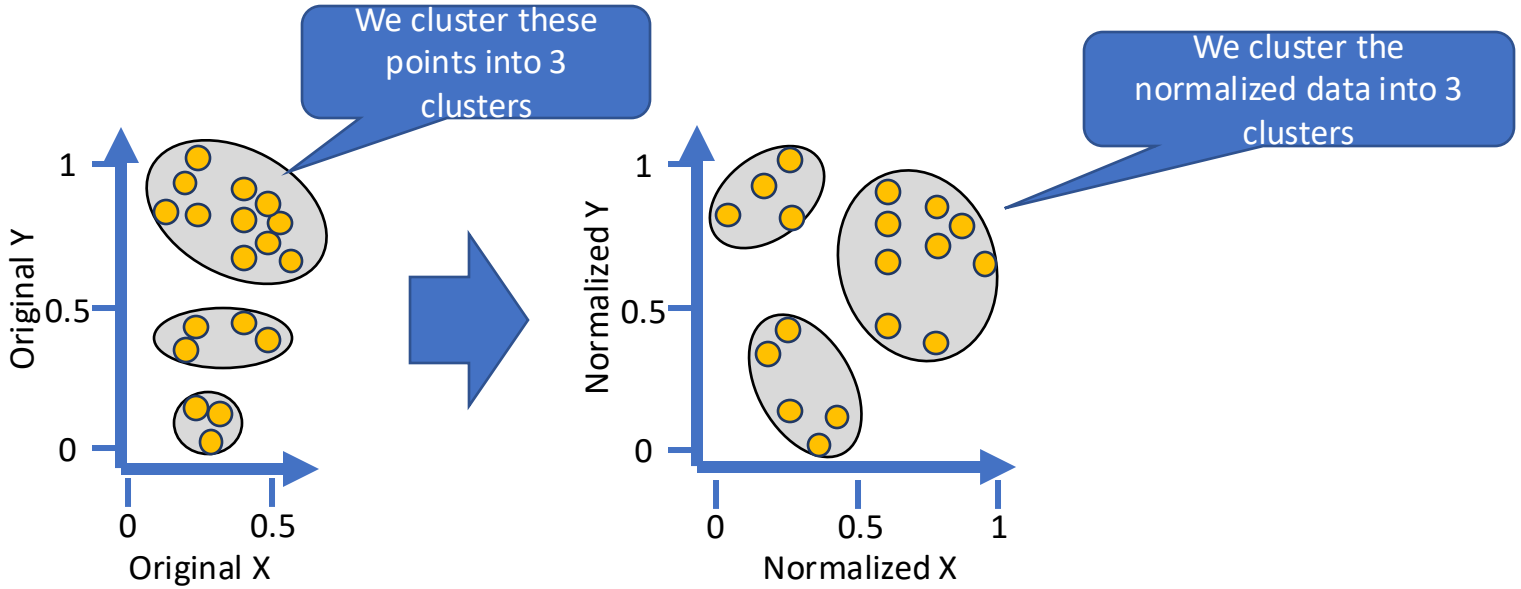
Effect of Normalization



Effect of Normalization



Effect of Normalization



Interview question (The Two-Envelope Problem)

- Imagine you are presented with two indistinguishable envelopes. You are told that each envelope contains a sum of money, and one envelope contains twice as much money as the other. However, you don't know which envelope holds the larger amount.
- You are allowed to choose one envelope and open it to reveal the amount of money inside. After seeing this amount, you have the option to either keep that envelope or switch to the other one.
- **Question:** Which strategy would maximize your expected gain: sticking with your initial choice or switching to the other envelope?