K-means and Normalizing

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Extensions of K-means

C-means: fuzzy version of K-means

K-means++: chooses initial centroids in a way that it guarantees a certain performance value

K-medians: uses medians instead of means for the centroids

Weighted K-means: assigns different weights to the features involved

Others

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

K-MEANS ALGORITHM

What about feature importance?

- Something to consider in cases of a large number of features in the dataset
- Importance ~ how a feature compares with the other features, in terms of scale
- In clustering, we tend to assume that all features are equally important, unless we are certain otherwise

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NORMALIZING

Why is normalizing necessary in general?

The values of a variable may be all over the place, making it difficult to get a sense of proportion for a data point in it

Variables that are not normalized are very difficult to compare or use together

Scale plays an important role in clustering, so unnormalized variables tend to distort the signal in the dataset

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Linear transformations

Changing a variable can be done in different ways, using mathematical function

- When the function used is linear, we have a linear transformation
  - (e.g. $f(x) = (5x - 2)/12$)
- Very common when normalizing a variable
- Distribution of a variable is preserved

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Non-linear transformations

Any transformation that uses a non-linear function

- e.g. $f(x) = x^2 + 4$, or $f(x) = e^x - x$

- Useful in certain models
- Distribution of a variable changes (oftentimes significantly)
- When applying a non-linear transformation to a feature, we usually store the result as a separate feature

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Max–min normalization:

Sets the min and max values:

- x' = (x - min(x)) / (max(x) - min(x))
- Yields values between 0 and 1 (inclusive)
- Center point: (max(x) + min(x))/2 → 0.5

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Standardization (or Z-Normalized)

Sets the mean value μ to 0.

- x' = (x - μ) / σ, where μ = arithmetic mean, σ = standard deviation
- Yields values approximately between -3 and 3 (exact border values depend on outliers in the variable)
- Center point: μ → 0.0

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

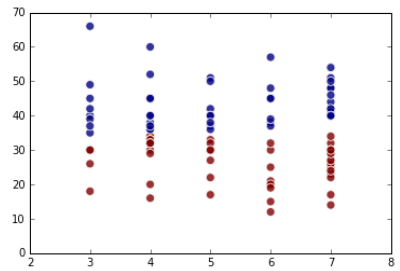
Keep in mind when normalizing:

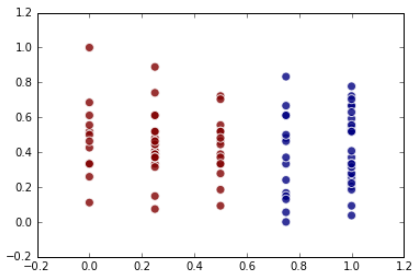
- Best to **remove outliers before** normalizing, since they have a strong effect in all linear transformations
- When normalizing many variables, it is good practice to use the **same normalization method for each** one of the variables
- Variables having no variance (i.e. they have the same value throughout the dataset) may yield errors when you normalize them. Best to remove them altogether.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NORMALIZING AND CLUSTERING

Example of a clustered dataset before and after normalizing its variables





PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Remove outliers before normalizing

>Normalize before clustering

>Choose the normalization method carefully

–Max-min keeps values between 0 and 1

–Z-norm keeps values between about -3 and 3

K-Means

Unsupervised Learning Algorithm

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON