Data transformation and standardization
UNIT 3

Dr. Praveen Barapatre

Data transformation and standardization

- Data transformation and standardization are fundamental processes in data wrangling, especially for statistical analysis and machine learning. They both aim to improve the quality of your data for further analysis, but in slightly different ways:

Data Transformation

- Broader concept: Encompasses a variety of techniques that modify data to achieve specific goals.
- Goals can include:
  - Normality: Transforming data to resemble a normal distribution, crucial for many statistical tests.
  - Outlier handling: Identifying and addressing extreme values that can skew analysis.

Data Transformation

- Feature engineering: Creating new features from existing ones to improve model performance.
- Encoding categorical data: Converting non-numerical data (like text categories) into a format usable by algorithms.
- Examples: Logarithmic transformation (compressing skewed data), square root transformation, binning (grouping data into ranges), one-hot encoding (creating binary features for categories).

Data Standardization

- Focuses on scaling: Aims to bring all features (data points) within a specific range.
- Common techniques:
  - Min-Max scaling: Scales data between a minimum and maximum value (often 0 and 1).
  - Standardization: Scales data to have a mean of 0 and a standard deviation of 1.

- Benefits:
- Creates a level playing field for features with different scales during analysis (e.g., prevents features with large values from dominating models).
- Improves the convergence of some machine learning algorithms.

analogy:

- Imagine a recipe that calls for ingredients in equal parts. But, if one ingredient comes in tablespoons and another in kilograms, it'll be difficult to measure accurately.
- Data transformation is like converting both ingredients to a standard unit (e.g., grams) for easier mixing.
- Data standardization is like ensuring you have the same amount (e.g., 100 grams) of each ingredient for consistent results.

Box-Cox and power transforms

- Box-Cox transformation and power transforms are closely related concepts in statistics, both used to manipulate data for better analysis.

Power Transform (general concept)

- A general family of functions that modify data using various powers (λ).
- Aims to achieve several goals:
  - Normality: Make data resemble a normal distribution, a crucial assumption for many statistical tests.
  - Variance stabilization: Reduce uneven spread (variance) across the data.
  - Improve association measures: Enhance the validity of statistics like correlation coefficients.
- Achieved through various functions, with the Box-Cox transformation being a specific type.

Box-Cox Transformation (specific type of power transform)

- Developed by statisticians George Box and David Cox.
- A particular function within the power transform family.
- Designed to transform data (especially non-normal dependent variables) towards a normal distribution.
- Works by raising each data point to a power (λ), with the optimal value chosen through statistical methods.
- Handles negative data by including a special case for the logarithm (λ = 0).

- analogy: Imagine stretching or compressing a spring (data) to achieve a desired shape (normal distribution).
- Power transforms provide a toolbox for this stretching/compressing, and the Box-Cox transformation is a specific tool within that toolbox.

Box-Cox Transformations

- Also called power transformations
- These transformations adjust for non-Normality and nonconstant variance
- $Y' = Y^{\lambda}$ or $Y' = (Y^{\lambda} - 1)/\lambda$
- In the second form, the limit as λ approaches zero is the (natural) log

Important Special Cases

- $\lambda = 1$, $Y' = Y^{1}$, no transformation
- $\lambda = .5$, $Y' = Y^{1/2}$, square root
- $\lambda = -.5$, $Y' = Y^{-1/2}$, one over square root
- $\lambda = -1$, $Y' = Y^{-1} = 1/Y$, inverse
- $\lambda = 0$, Y' = (natural) log of Y

Box-Cox Details

- We can estimate λ by including it as a parameter in a non-linear model
- $Y^{\lambda} = \beta_0 + \beta_1 X + e$

  and using the method of maximum likelihood
- Details are in KNNL p 134-137
- SAS code is in boxcox.sas

Box-Cox Solution

- Standardized transformed Y is
  - $K_1(Y^{\lambda} - 1)$ if $\lambda \neq 0$
  - $K_2 \log(Y)$ if $\lambda = 0$

  where $K_2 = (\prod Y_i)^{1/n}$ (the geometric mean)

  and $K_1 = 1/ (\lambda K_2^{\lambda-1})$
- Run regressions with X as explanatory variable
- estimated λ minimizes SSE