# Attribute Transformation

Lesson

AI Academy

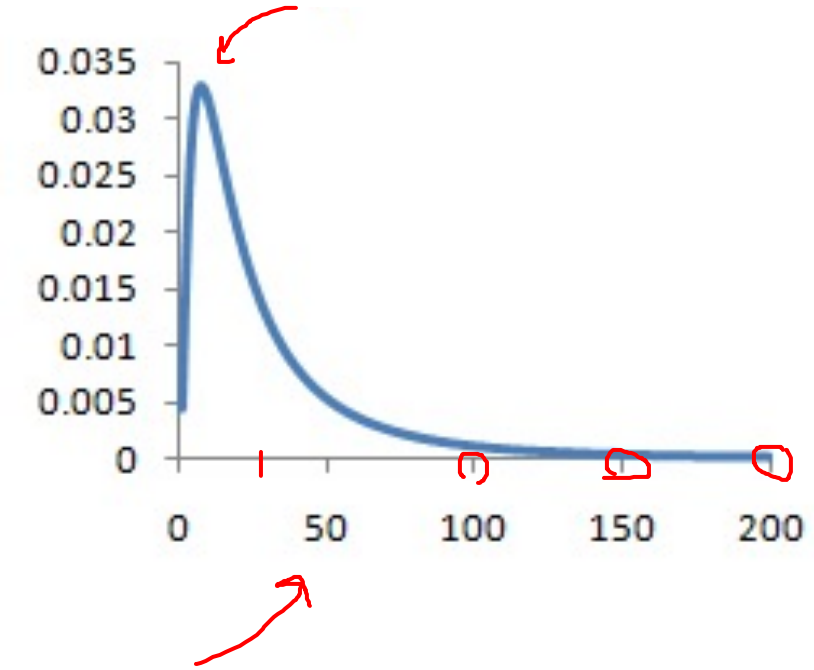
# Data Preprocessing

- Sampling
- Feature subset selection
- Dimensionality Reduction
- Feature creation
- Discretization and binarization
- **Attribute Transformation**

# Motivation: Skewed Data

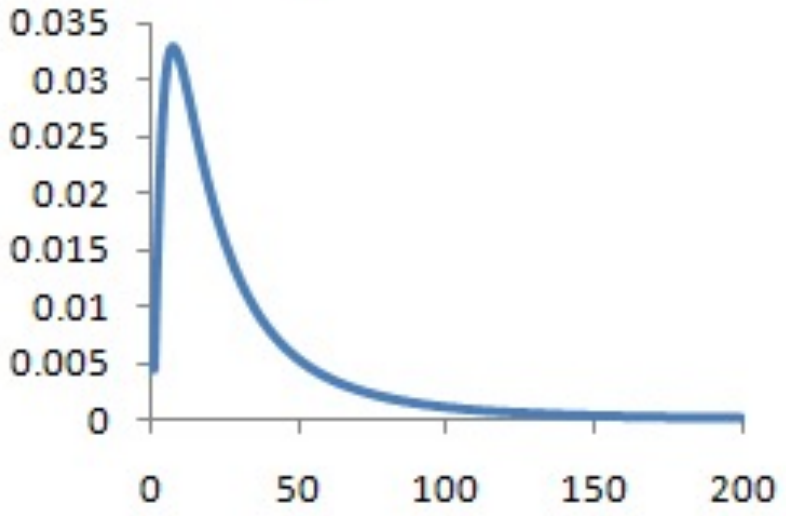
Some learning approaches (e.g. regression) make assumptions about the distribution of data

- However, sometimes data is very *skewed*
- Lots of low values, a few extreme high values

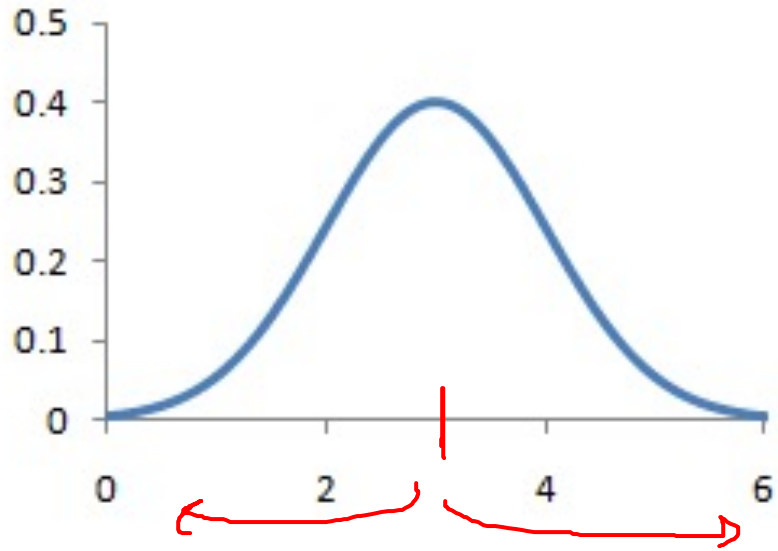


# Transformation Example: Log Scaling

$x' = \log(x)$




x




log(x)

# Attribute Transformation

**Attribute Transformation:** A function that maps all the original attribute values to new values

- 1-to-1 mapping
- Simple functions
  - $x^k$,
  - log(x),
  - $e^x$,
  - |x|
- Standardization and Normalization

# An Example

John and Sara both took a math class under two different professors. They took different exams and wanted to compare their test scores.

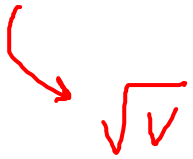
John got an 85 (out of 100) on test A

Sara got an 8.5 (out of 10) on test B.

*Who has the better score?*

# Results

- John's Class
  - Mean = 75
  - SD = 11.81



- Sara's Class
  - Mean = 7.5
  - SD = 1.159

# Attribute Transformation

## Normalization

**Goal:** Ensure all attribute values have a given property

**Example**:
Normalization to [0-1]:
x’ = (x-min) / (max-min)

# Z-score Normalization (Standardization)

- Based on the mean standard deviation
- Centered around 0
- Distance in units of standard deviation

$$z(x_i) = \frac{x_i - \overline{x}}{s}$$

# Results

- John's class
  - Mean = 75
  - SD = 11.81
  - z(85) = ?
- Sara's class
  - Mean = 7.5
  - SD = 1.159
  - z(8.5) = ?

# Results

- John's class
  - Mean = 75
  - SD = 11.81
  - z(85) = (85-75) / 11.81 = 0.847
- Sara's class
  - Mean = 7.5
  - SD = 1.159
  - z(8.5) = (8.5-7.5) / 1.15 = 0.863

# When to use z-score Normalization?

- Before performing PCA.
- Before calculating a distance measure (next Seminar).
- Before using certain learners (e.g. SVM).
- When comparing *across* different attributes

**Not Always**: Outliers can affect z-score normalization.

**Rule of Thumb**: Try it, and see if it helps.

# Learning Objectives: Normalization

## You now should be able to:

Explain what normalization is, when it is used, and what problems it addresses.

AI Academy
NC STATE



# Normalization

## Exercises

AI Academy

| Attribute Level | Transformation | Comments |
|---|---|---|
| Nominal | Any permutation of values | If all employee ID numbers were reassigned, would it make any difference? |
| Ordinal | An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function. | An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10} |
| Interval | *new_value =a * old_value + b* where a and b are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| Ratio | *new_value = a * old_value* | Length can be measured in meters or feet. |

# Practice Exercise

John's class (85)
- Mean = 95
- SD = 10
- z = ?

Sara's class (8.5)
- Mean = 7
- SD = 1
- z = ?

John's class (85)
- Mean = 75
- SD = 10
- z = ?

Sara's class (8.5)
- Mean = 7.5
- SD = 2
- z = ?

# AI Academy

**go.ncsu.edu/aiacademy**

NC STATE UNIVERSITY