Data Transformation and Distributions

Lecture 5.1
29/07/2020

Recap

- Similarity dissimilarity measures
- Distance measures
- Cosine similarity

Today's plan

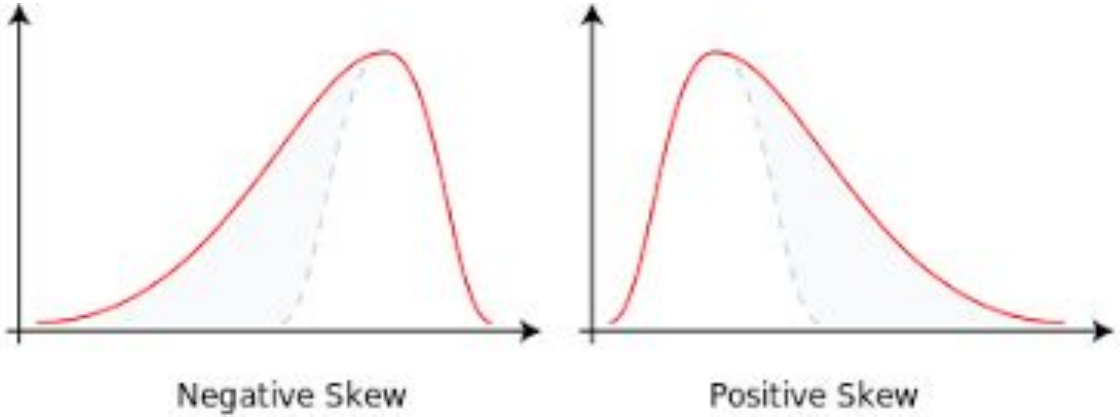
Introduction to

- Data transformation, Tasks
- Distributions

Intuition

Most of the dataset in real world are skewed

That means coming with noise, outliers and missing values

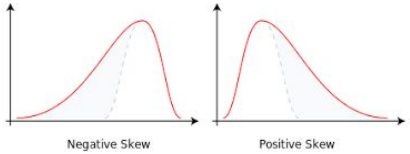


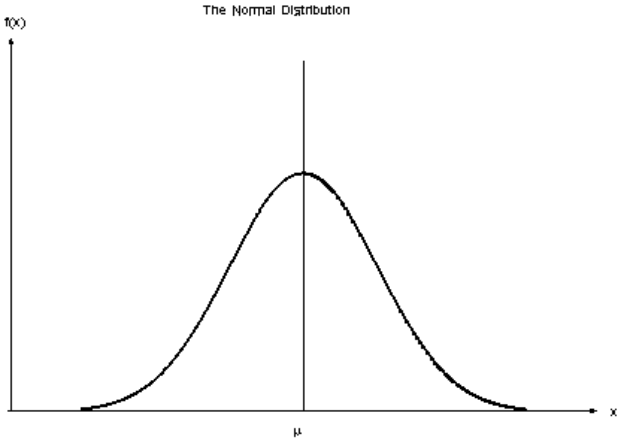
Intuition

Most of the dataset in real world are skewed

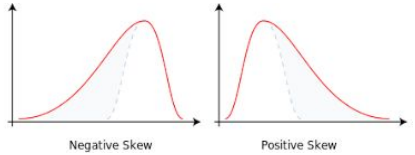
That means coming with noise, outliers and missing values

But data mining algorithms assume data is **“normally”** distributed



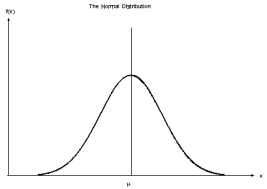


Intuition



Most of the dataset in real world are skewed

That means coming with noise, outliers and missing values



But data mining algorithms assume data is **“normally”** distributed

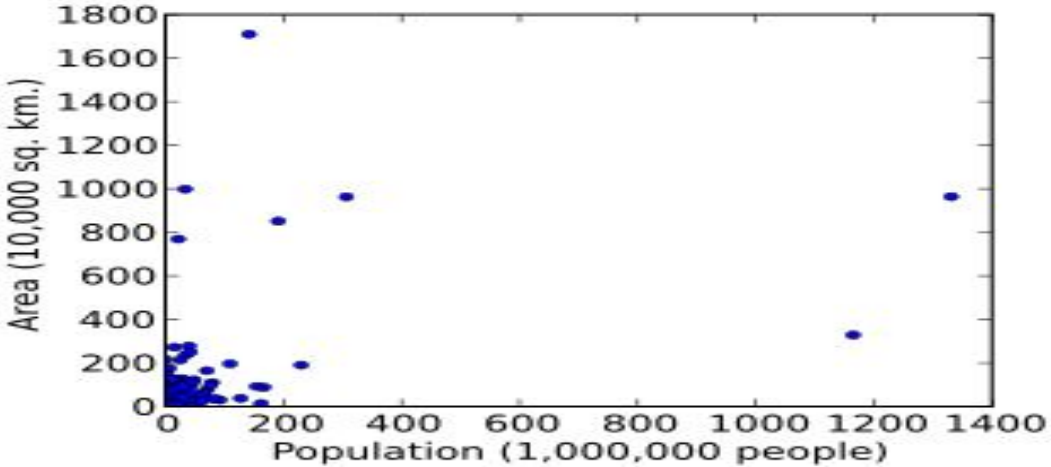
Once skewness is identified, attempt to convert the data into some other form, so that it can “fit” the normal distribution

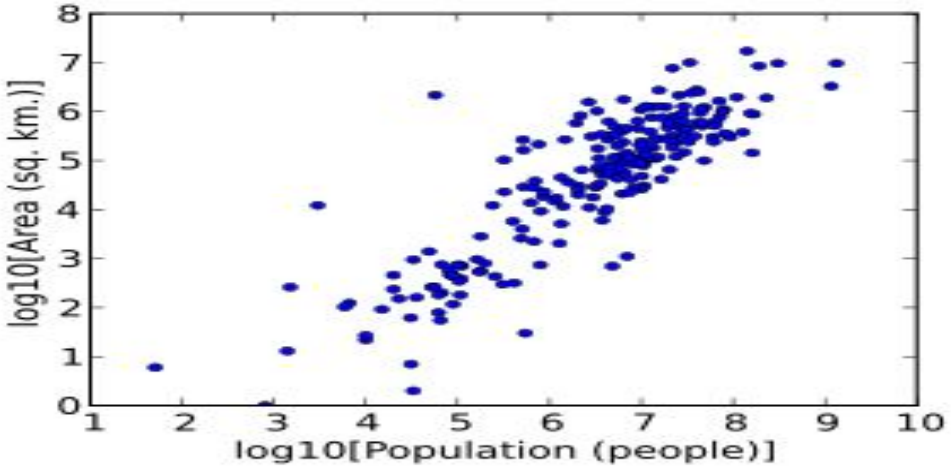
This attempt is called **“ Transformation “**

One more example

Transformed using log10

Example from Wikipedia page on data transformation





Let's define

Distribution

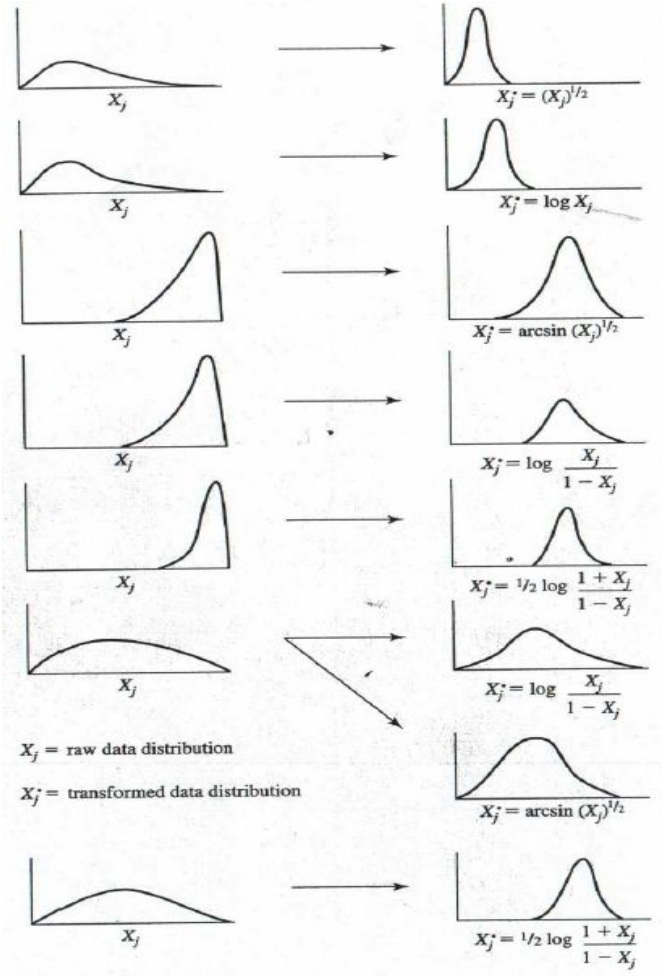
- The **pattern of values** obtained when a variable is measured in large number of individuals is called a distribution

Transformation

- The application of a **deterministic mathematical function** to each point in a data set
- That is, each data point z_i is replaced with the transformed value $y_i = f(z_i)$

Transformation

Transforming Data for Normality



Data Transformation- Tasks

- **Smoothing** is a process of removing noise from the data.
- **Normalization** scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0
- **Attribute construction** new attributes are constructed from the given set of attributes
- **Aggregation:** summary or aggregation operations are applied to the data Daily sales aggregated to monthly sale
- **Discretization:** Dividing the range of a continuous attribute into intervals
- **Generalization:** low level data may be replaced by high level concepts (eg: city replaced by state of country)

Normalization

- An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0.
- Normalization is particularly useful for classification algorithms involving
  - Neural networks
  - Measuring similarity
- For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income) from out-weighing attributes with initially smaller ranges (e.g., binary attributes).
-

Employee Name	Salary	Year of Experience	Expected Position Level
Hitesh Khurana	100000	10	2
Ajeish Khanna	78000	7	4

Normalization Methods

- Min-Max normalization
- Z-score normalization
- Normalization by decimal scaling
- Log-normalization

Min-Max normalization

- Perform a linear transformation on the original data
- Preserves the relationship between original data values
- Suppose min(A), max(A) are minimum and maximum values in the feature attribute A
- Then the value v is transformed in to v’ using

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}(new\_max(A) - new\_min(A)) + new\_min(A)$$

- new_max(A), new_min(A) are the range of the transformation required

Example

Let income range $12,000 to $98,000 normalized to [0.0, 1.0].

Then $73,000 is mapped to

(7300- 12000)/(98000-12000) = 0.76

Let exam marks in range of 10 to 100, normalized to [1, 5].

Then mark 56 is mapped to?

( )/( ) = ?

Z-score normalization

- In z-score normalization(or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
- A value, v, of attribute A is normalized to v’ by computing
-

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score normalization

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

marks	x-mean(x)	(x-mean(x))^2	Z-score
8			
10			
15			
20			
Mean=			
	$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$		

Decimal Scaling

- It normalizes the values of an attribute by changing the position of their decimal points
- The number of decimal points moved depends on the maximum absolute value of A

$$v' = \frac{v}{10^j}$$

- where j is the smallest integer such that Max(|v'|) < 1

Example

Suppose that the recorded values of A range from -986 to 917

The maximum absolute value of A is 986.

To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., j = 3)

So value 500 will be transformed into 500/1000 = 0.5

Tr

The scores of a game for a player is given below:

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

Normalize the values

a. min-max normalization (between 0 and 1)

b. z-score normalization