Data Engineering
Lecture 9: Data Preprocessing V

Nada Sharaf

The German International University

Distance Equations: Euclidean Distance

$$Distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Distance Equations: Euclidean Distance

$$Distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- Similar to finding the shortest distance between two points

Distance Equations: Euclidean Distance

$$Distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- Similar to finding the shortest distance between two points
- n is the number of dimensions
- p_i and q_i are data points

Distance Equations: Euclidean Distance

$$Distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- Similar to finding the shortest distance between two points
- n is the number of dimensions
- p_i and q_i are data points
- Usually used with Interval, Ratio

Distance Equations: Manhattan Distance

$$Distance = \sum_{i=1}^{n} |p_i - q_i|$$

Distance Equations: Manhattan Distance

$$Distance = \sum_{i=1}^{n} |p_i - q_i|$$

- Summing the absolute differences between points among the different dimensions
- Usually used with Ordinal values

Distance Equations: Jaccard Distance

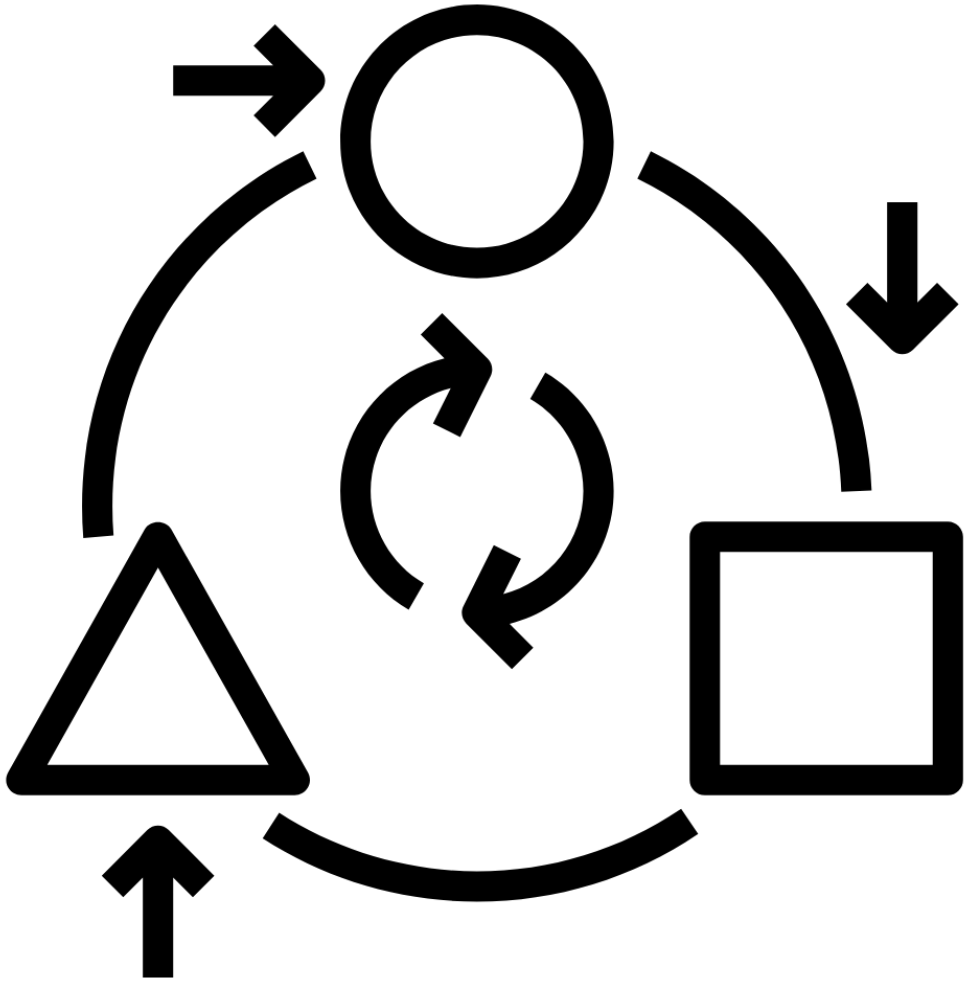
$$Distance = 1 - \frac{P \cap Q}{P \cup Q}$$

Distance Equations: Jaccard Distance

$$Distance = 1 - \frac{P \cap Q}{P \cup Q}$$

- Similarity is $\frac{P \cap Q}{P \cup Q}$
- Used with Categorical values

Let us **Transform** the data



Data Transformation

- Convert the raw data into a suitable structure

Data Transformation

- Convert the raw data into a suitable structure
- Helps data mining techniques to retrieve the needed information efficiently and easily

Data Transformation

- Convert the raw data into a suitable structure
- Helps data mining techniques to retrieve the needed information efficiently and easily
- Usually needed for data integration

Data Transformation Strategies: Smoothing

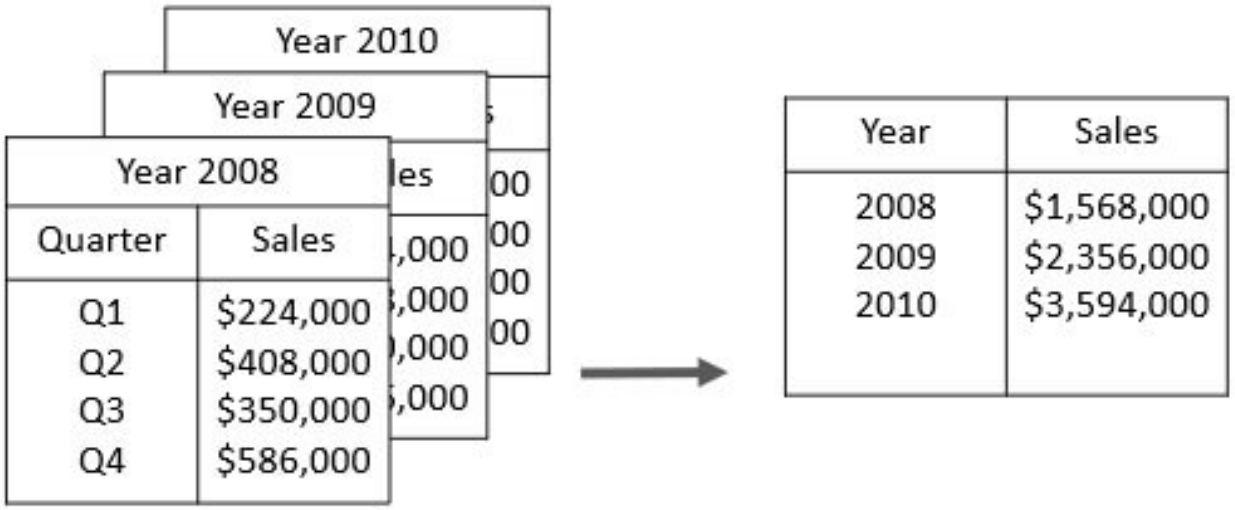
- Discussed in data cleaning
- Removes noise
- **Binning**, **Regression**, **Clustering**

Data Transformation Strategies: Data Aggregation

- Converts a large set into a smaller volume
- Grouping and summarization, reduction methods

Data Transformation Strategies: Data Aggregation

- Converts a large set into a smaller volume
- Grouping and summarization, reduction methods



© https://binaryterms.com/data-transformation.html

Data Transformation Strategies: Attribute Construction

- New attributes constructed using the existing attributes
- To have data set that helps with mining

Data Transformation Strategies: Attribute Construction

- New attributes constructed using the existing attributes
- To have data set that helps with mining
- For example: adding the attribute **area**

Data Transformation Strategies: Database Normalization

- Ensuring having as many normal forms as possible
- Getting rid of insertion, update and deletion anomalies

Data Transformation Strategies: Data Normalization (Attribute Normalization)

- Scaling data to a smaller range
- e.g. [-1,1] or [0,1]

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Min-Max Normalization

- Linear Transformation for the original data

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Min-Max Normalization

- Linear Transformation for the original data
- Assume having values min_A as the minimum value and max_A as the maximum value for a specific attribute A

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Min-Max Normalization

- Linear Transformation for the original data
- Assume having values min_A as the minimum value and max_A as the maximum value for a specific attribute A
- Min-Max Normalization would change the range of the attribute A to [new_minA, new_maxA].

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Min-Max Normalization

- Linear Transformation for the original data
- Assume having values min_A as the minimum value and max_A as the maximum value for a specific attribute A
- Min-Max Normalization would change the range of the attribute A to [new_minA, new_maxA].
- Assume you want to map a value v_i to a new value v_i'
- The formula is:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Example

Assume having an attribute income with a current minimum and maximum values of 12000 and 98000 respectively. We want to normalize the attribute to be in the range [0.0, 1.0]. What would an income of 73600 map to?

Use sli.do code 284942

The formula is:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Z-Score Normalization

- Normalizes the value using mean and standard deviation
- The formula is:

$$v_i' = \frac{v_i - \overline{A}}{\alpha_A}$$

- such that $\overline{A}$ is the mean and α_A is the standard deviation

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Z-Score Normalization

- Normalizes the value using mean and standard deviation
- The formula is:

$$v_i' = \frac{v_i - \overline{A}}{\alpha_A}$$

- such that $\overline{A}$ is the mean and α_A is the standard deviation
- For the same example, the mean is 54000 and the standard deviation is 16000. The Z-Score for the income 73600 is:

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Z-Score Normalization

- Normalizes the value using mean and standard deviation
- The formula is:

$$v_i' = \frac{v_i - \overline{A}}{\alpha_A}$$

- such that $\overline{A}$ is the mean and α_A is the standard deviation
- For the same example, the mean is 54000 and the standard deviation is 16000. The Z-Score for the income 73600 is: $\frac{73600-54000}{16000} = 1.225$

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Decimal Scaling

- Moving the decimal point
- $$v_i' = \frac{v_i}{10^j}$$

  such that j is the smallest integer such that $max(|v_i'|) < 1$

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Decimal Scaling

- Moving the decimal point
- $$v_i' = \frac{v_i}{10^j}$$

  such that j is the smallest integer such that $max(|v_i'|) < 1$
- For example if the range of an attribute A is -986 to 917 then j is 3. (Note that 2 is too little and 4 is too much)
- Thus, every value is divided by 1000.

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Discretization

- Replacing the values of numeric data by the labels

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Discretization

- Replacing the values of numeric data by the labels
- Age could be divided into ranges (0-10, 11-20 ...) or (kid, youth, adult, senior) or encodings (0, 1, 2, ...).

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Discretization

- Replacing the values of numeric data by the labels
- Age could be divided into ranges (0-10, 11-20 ...) or (kid, youth, adult, senior) or encodings (0, 1, 2, ...).
- Supervised discretization: the class information is used
- Unsupervised discretization: based on the direction of the process
  - Top-down splitting
  - Bottom-up merging

Label Encoding

- Convert labels to numeric forms
- Machine-readable form
- Unique number for every class

Label Encoding

- Convert labels to numeric forms
- Machine-readable form
- Unique number for every class

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

ID	Country	Population
1	0	127185332
2	1	326766748
3	2	1354051854
4	3	1415045928
5	1	326766748
6	2	1354051854

Label Encoding

- Convert labels to numeric forms
- Machine-readable form
- Unique number for every class

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

ID	Country	Population
1	0	127185332
2	1	326766748
3	2	1354051854
4	3	1415045928
5	1	326766748
6	2	1354051854

- But,

Label Encoding

- Convert labels to numeric forms
- Machine-readable form
- Unique number for every class

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

ID	Country	Population
1	0	127185332
2	1	326766748
3	2	1354051854
4	3	1415045928
5	1	326766748
6	2	1354051854

- But, does it add bias? would a label with a higher value be given more priority?

One-Hot Encoding

- A number of columns
- Each cell can have a zero or a one.
- For each category, only one of the cells have a 1

	Emploee_ID	Remarks_Good	Remarks_Great	Remarks_Nice	Gender_Female	Gender_Male
0	45	0	0	1	0	1
1	78	1	0	0	1	0
2	56	0	1	0	1	0
3	12	0	1	0	0	1
4	7	0	0	1	1	0
5	68	0	1	0	1	0
6	23	1	0	0	0	1
7	45	0	0	1	1	0
8	89	0	1	0	0	1
9	75	0	0	1	1	0
10	47	1	0	0	1	0
11	62	0	0	1	0	1

© https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Concept Hierarchy Generation

- Nominal attributes form the concept hierarchy by incorporating a group of attributes.
- Replacing low-level concepts with higher-level concepts
- For example, street, city, state, country all together can generate concept hierarchy

Data Transformation Strategies: Data Normalization (Attribute Normalization) Using Concept Hierarchy Generation

- Nominal attributes form the concept hierarchy by incorporating a group of attributes.
- Replacing low-level concepts with higher-level concepts
- For example, street, city, state, country all together can generate concept hierarchy
- Concept hierarchies could be specified by experts of the domain
- They can also be automatically formed (using discretization for example).

Data Reduction

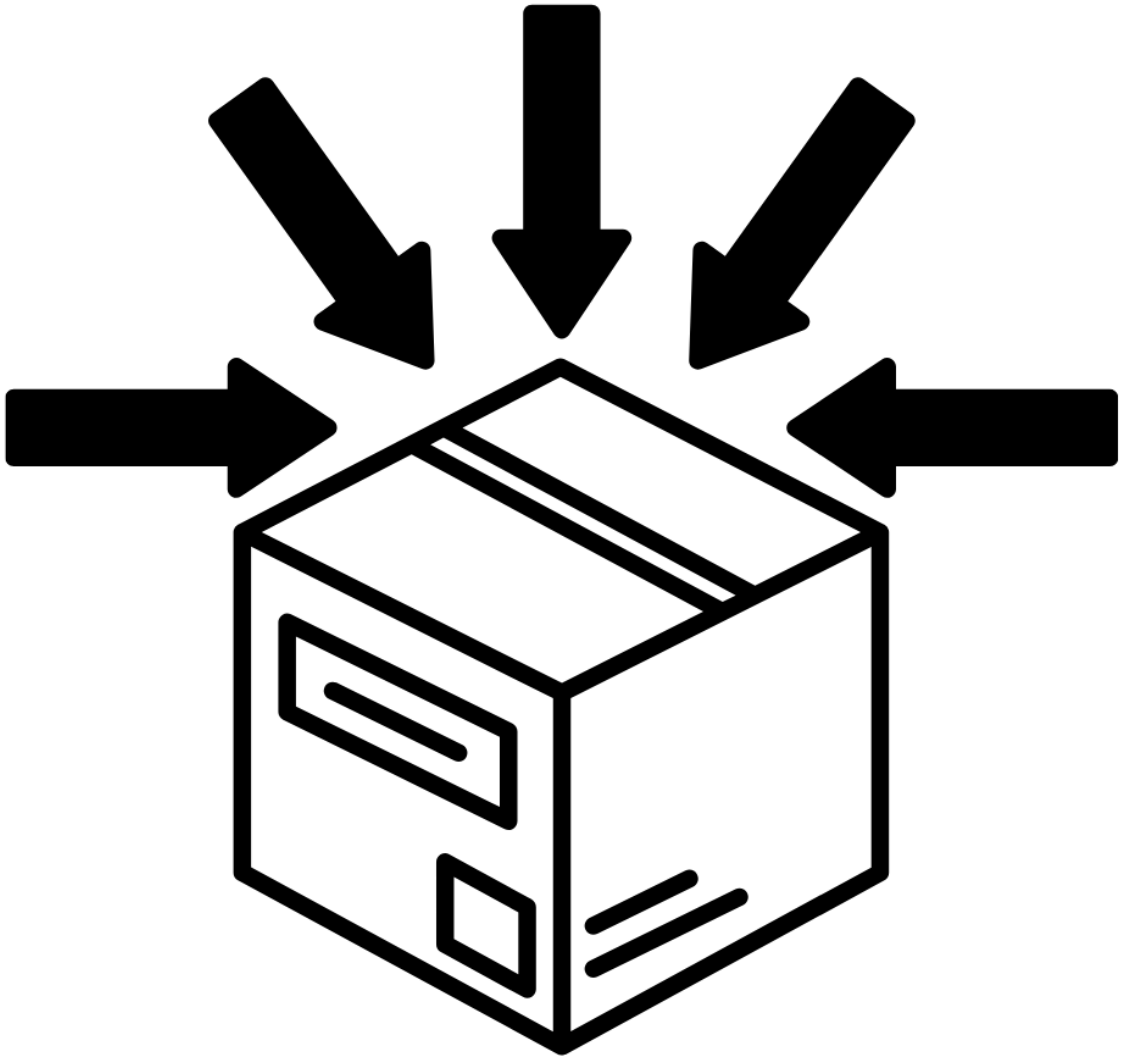
Dimensionality Reduction

- Feature selection select a subset of the original features
- Feature extraction: Transform the high-dimensional data into fewer-dimensional data

Numerosity Reduction

- Replace the original data by smaller data
- Parametric:
  - Assuming a model into which the data fits
  - Only the model parameters are saved
  - e.g. Regression
- Nonparametric
  - Does not assume a model
  - e.g. clustering, histograms, sampling
  - You can check `https://www.jigsawacademy.com/blogs/data-science/data-reduction` for more details

Aggregation



Flipped Classroom

In sampling, what is

- Simple Random Sample Without Replacement of sizes
- Simple Random Sample with Replacement of sizes
- Cluster Sample
- Stratified Sample

Thank you :)