Feb – Jun 2023
IT258

# Data Preprocessing

## Dealing with Noisy Data

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

## Data Cleaning

**Tasks –**

1. Dealing with missing values
2. **Identifying outliers and smoothing out noisy data**
3. Correcting inconsistent data



## Noisy data

- **Noise**: random error or variance in a measured variable.
- Incorrect attribute values may due to -
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - inconsistency in naming convention



## How to handle noisy data?

- Simple Discretization Methods
- Grouping and Aggregation
- Clustering
  - Categorize data as per similarity/dissimilarity



## Simple Discretization

- Process of converting continuous features into discrete features.
  - Reduces the degrees of freedom of data.
- creating a set of contiguous intervals (or bins) to represent the desired variable's values.



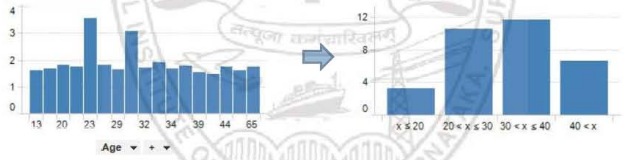
## Simple Discretization

- Process of converting continuous features into discrete features.
  - Reduces the degrees of freedom of data.
- creating a set of contiguous intervals (or bins) to represent the desired variable's values.

* **Continuous data** is *measured*, while **Discrete data** is *counted*.

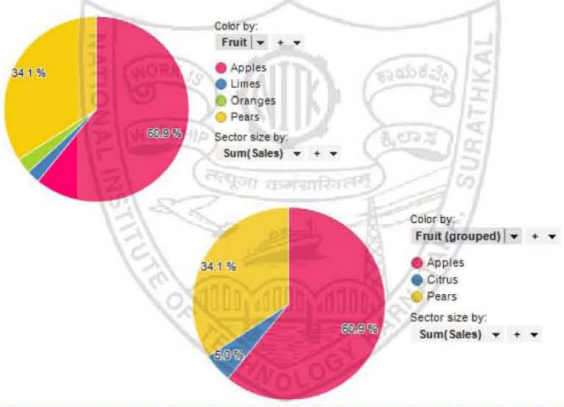
## Simple Discretization - example

- E.g.: data about a group of people is grouped as per their age intervals *(into a smaller number of intervals)*.






## Simple Discretization - example






## Types of Discretization methods.

- Ways to transform numerical variables into categorical variables.
  - **Unsupervised** → do not use the target (class) information
    - Equal-width binning
    - Equal-frequency binning
  - **Supervised** → refer to the target (class) information when selecting discretization cut points.
    - Entropy based discretization



## Unsupervised Discretization Methods

**1. Equal-width partitioning :** *(also called equal-distance partitioning)*

- Process:
  - Divide the range into N intervals of width **w**
  - Width of intervals $w = (max - min)/N$
    - *min and max are the lowest and highest values of the attribute*
  - Interval boundaries - $min+w, min+2w, \ldots, min+(N-1)w$



- **Equal-width partitioning Example**

Data: 5, 10, 11, 13, 15, 35, 50 ,55, 72, 92, 204, 222, 230

→ Decide on N, then calculate w,

Let N=3; Width w = 75

**bin1: 5,10,11,13,15,35,50,55,72** *(i.e. all values between 3 &75 )*
**bin2: 92** *(i.e. all values between 75 & 150)*
**bin3: 204,230** *( i.e. all values between 150 & 230)*

## Unsupervised Discretization Methods

**2. Equal-depth partitioning :**

- Also called equal-frequency partitioning
- divides the range into N intervals, each containing approximately same number of samples.

- **Equal-depth (frequency) partitioning - Example**

Data: 5, 10, 11, 13, 15, 35, 50 ,55, 72, 92, 204, 222, 230

**bin1: 5,10,11,13,15**
**bin2: 35,50,55, 72, 92**
**bin3: 204,222, 230**

## Binning methods for Data Smoothing

- First use Equal-width or Equal-depth partitioning on the data.
- Then, use various smoothing strategies like –
  - Smooth by Bin Mean
  - Smooth by Bin Median
  - Smooth by Bin Boundaries etc..



## Binning methods for Data Smoothing

**Example:** Consider the sorted data for cost (in dollars):
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

First, partition into (equi-depth) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

***Method 1:*** **Smoothing by Bin Mean**
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29



## Binning methods for Data Smoothing

**Example:** Consider the sorted data for cost (in dollars):
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

First, partition into (equi-depth) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

***Method 2:*** **Smoothing by Bin Median**
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29



## Binning methods for Data Smoothing

**Example:** Consider the sorted data for cost (in dollars):
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

First, partition into (equi-depth) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

***Method 3:*** **Smoothing by Bin Boundaries**
- Bin 1: 4, 4, 15, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 34, 34



## Supervised Discretization Methods

- **Entropy based partitioning**
  - Goal: find the best split so that the bins are as pure as possible
    - *i.e. the majority of the values in a bin correspond to the same class label.*
  - characterized by finding the split with the maximal information gain.
    - entropy (or the *goodness of information content*)

## Supervised Discretization Methods

- **Entropy based discretization - Process**
  - *Step 1:* Calculate "Entropy" for the target variable.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Where, $p_i$ = probability associated with observation *i*

  - *Step 2:* Calculate "Entropy" for the target variable, given a bin.

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

Where, $S_v$ = observation at target *v*
*S = total observations*
$E(S_v)$ = *Entropy of observation* $S_v$

  - *Step 3:* Calculate "Information Gain" for given a bin.

$$\textit{Information Gain} = E(S) - E(S,A)$$



## Entropy based discretization

- Process:
  - Step 1: Calculate the "entropy" of the target variable
  - Step 2: Calculate the "entropy" of the target variable for a random "bin"
  - Step 3: Calculate the "Information Gain" for the chosen "bin"
  - Step 4: Repeat Step1-3 for other random bin values and find IG for each bin*
    **Typically at least 2 more iterations with different bin values are to be performed.*
  - Step 5: Compare the information gain values obtained for different bin values.
  - **Decision:** The best bin interval for the dependent variable is the one which achieves the **highest IG value w.r.t target variable.**



## Class Exercise

- Data w.r.t age of a vehicle (in months) and mechanical failure cases is given. For the given observations, derive the best possible discretization interval for the variable *Age of vehicle*, using supervised discretization method.

| Age of vehicle | Mechanical Failure |
|---|---|
| 53 | Y |
| 56 | Y |
| 57 | Y |
| 63 | N |
| 66 | N |
| 67 | N |
| 67 | N |
| 67 | N |
| 68 | N |
| 69 | N |
| 70 | N |
| 70 | Y |
| 70 | Y |
| 70 | Y |
| 72 | N |
| 73 | N |
| 75 | N |
| 75 | Y |
| 76 | N |
| 76 | N |
| 78 | N |
| 79 | N |
| 80 | N |
| 81 | N |



## Step 1: Calculate the "entropy" of the target variable

| Mechanical Failure | |
|---|---|
| Y | N |
| 7 | 17 |

"entropy" probability of individual observations of the target variable

**E(Y)** = 7 / (7+17) = 0.29

**E(N)** = 17/(7+17) = 0.71

## Step 1: Calculate the "entropy" of the target variable

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Where, $p_i$ = probability associated with observation *i*

| Mechanical Failure | |
|---|---|
| Y | N |
| 7 | 17 |

E(Y) = 7 / (7+17) = 0.29

E(N) = 17/(7+17) = 0.71

E (Failure) = E(7, 17) = E(0.29, .71)

= -0.29 x $\log_2$(0.29) - 0.71 x $\log_2$(0.71)

= **0.871**

## Step 2: Calculate the "entropy" of the target variable for a random "bin"

| | | Mechanical Failure | |
|---|---|---|---|
| | | Y | N |
| Age of vehicle (in months) | <=60 | 3 | 0 |
| | >60 | 4 | 17 |

| Age of vehicle | Mechanical Failure |
|---|---|
| 53 | Y |
| 56 | Y |
| 57 | Y |
| 63 | N |
| 66 | N |
| 67 | N |
| 67 | N |
| 67 | N |
| 68 | N |
| 69 | N |
| 70 | N |
| 70 | Y |
| 70 | Y |
| 70 | Y |
| 72 | N |
| 73 | N |
| 75 | N |
| 75 | Y |
| 76 | N |
| 76 | N |
| 78 | N |
| 79 | N |
| 80 | N |
| 81 | N |

## Step 2: Calculate the "entropy" of the target variable for a random "bin"

| | | Mechanical Failure | |
|---|---|---|---|
| | | Y | N |
| Age of vehicle (in months) | <=60 | 3 | 0 |
| | >60 | 4 | 17 |

"entropy" probability of individual observations of the target variable

$E_1(Y,N) = E(3,0) = E(3/(3+0), 0/(3+0)) = E(1, 0)$

$E_2(Y,N) = E(4,17) = E(4/(4+17), 17/(4+17)) = E(0.19, 0.81)$

## Step 2: Calculate the "entropy" of the target variable for a random "bin"

| | | Mechanical Failure | |
|---|---|---|---|
| | | Y | N |
| Age of vehicle (in months) | <=60 | 3 | 0 |
| | >60 | 4 | 17 |

"entropy" probability of individual observations of the target variable

$$E(S_v) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

➔ $E(1,0) = -1*\log_2(1) - 0 = 0$

➔ $E(0.19, 0.81) = -0.19*\log_2(0.19) - 0.81*\log_2(0.81) = 0.7$

## Step 2: Calculate the "entropy" of the target variable for a random "bin"

| | | Mechanical Failure | |
|---|---|---|---|
| | | Y | N |
| Age of vehicle (in months) | <=60 | 3 | 0 |
| | >60 | 4 | 17 |

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

E (Failure, Age of Vehicle) = P(<=60) x E(3,0) + P(>60) x E(4,17)

= 3/24 x 0 + 21/24 x 0.7

= **0.615**

## Step 3: Calculate the "Information Gain" for the chosen "bin"

| | | Mechanical Failure | |
|---|---|---|---|
| | | Y | N |
| Age of vehicle (in months) | <=60 | 3 | 0 |
| | >60 | 4 | 17 |

*Information Gain* = $E(S) - E(S,A)$

*Information Gain* (Failure, Age of Vehicle) = 0.871 - 0.615

= 0.256

## Step 4: Repeat Steps 1-3 for some other random bin values (E.g. 65, 70, 75) and find IG for each bin*

**Typically at least 2 more iterations with different bin values are to be performed.*

## Step 5: Compare the information gain values obtained for the different bin values

**Decision:** The best bin interval for "Age of Vehicle" is the one which achieves the highest IG value w.r.t "Mechanical failure", thus achieving the required discretization interval.

## Homework

- The various observations Derive the best possible discretization interval for relevant variable using entropy based discretization.

| | Student_id | Age | Grade | Employed |
|---|---|---|---|---|
| 0 | 1 | 19 | 1st Class | yes |
| 1 | 2 | 20 | 2nd Class | no |
| 2 | 3 | 18 | 1st Class | no |
| 3 | 4 | 21 | 2nd Class | no |
| 4 | 5 | 19 | 1st Class | no |
| 5 | 6 | 20 | 2nd Class | yes |
| 6 | 7 | 19 | 3rd Class | yes |
| 7 | 8 | 21 | 3rd Class | yes |
| 8 | 9 | 22 | 3rd Class | yes |
| 9 | 10 | 21 | 1st Class | no |

# Aggregation based Data Smoothing

## Aggregation based Data smoothing

- Process of combining two or more attributes *(or objects)* into a single attribute *(or object)*.
- *Purpose* -
  - Data reduction:
    - reduce the no.of attributes or objects.
  - Change of Scale:
    - providing a high level view of data instead of a low level view
  - More 'stable' data
    - Aggregated data tends to have less variability.



## Aggregation based Data smoothing

- Consists of essentially two steps –
  - Data grouping
    - Identifies one or more data groups based on values in selected features
  - Data Aggregation
    - Puts together (aggregates) the values in one or more selected attributes for each group.



## Aggregation - Example

- Requirement:
  - A researcher wants to analyse the opinion of a group of people w.r.t to certain trending topics. The preprocessed dataset contains the opinion of both men and women on the topic. Perform grouping and aggregation tasks for analysing the data further.



## Aggregation - Example

- **Step 1: Data grouping**

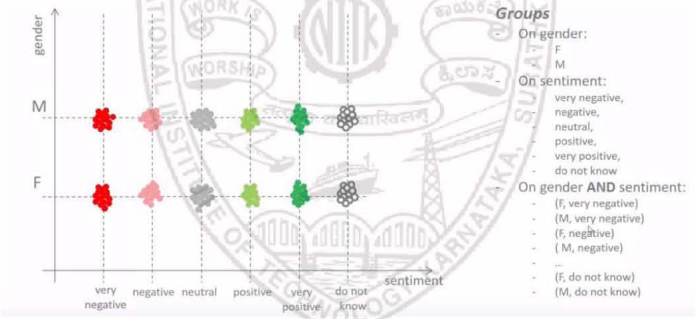



## Aggregation - Example

- **Step 1: Data grouping**

## Aggregation - Example

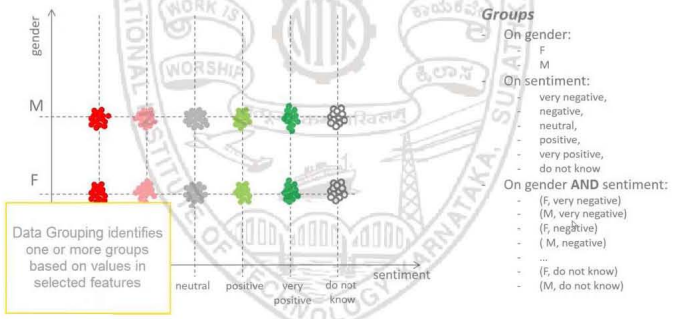
### Step 1: Data grouping






## Aggregation - Example

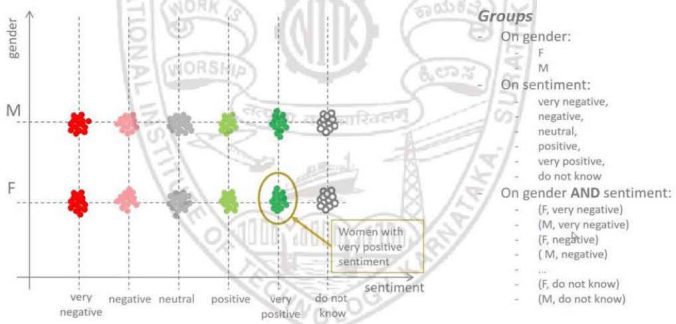
### Step 1: Data grouping






## Aggregation - Example

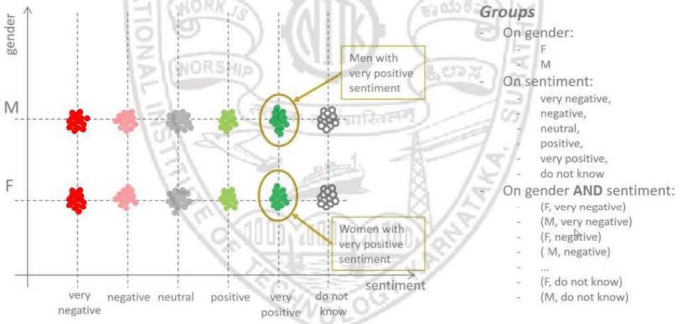
### Step 1: Data grouping






## Aggregation - Example

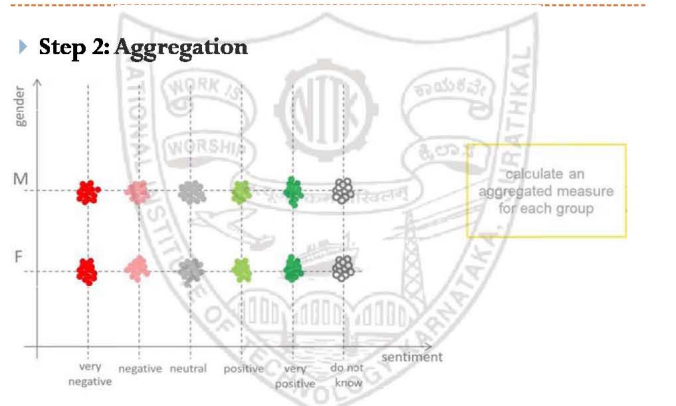
### Step 1: Data grouping






## Aggregation - Example

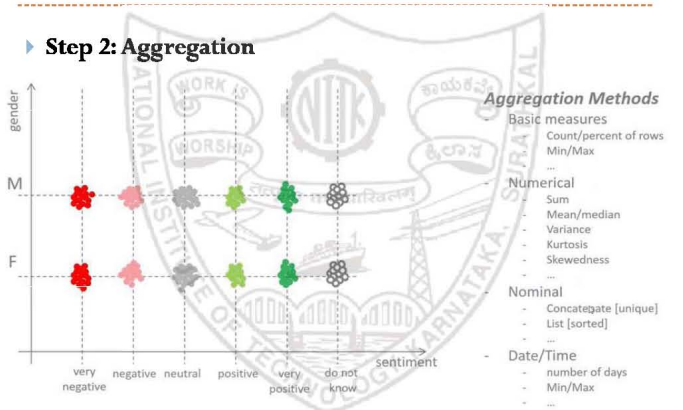
### Step 2: Aggregation





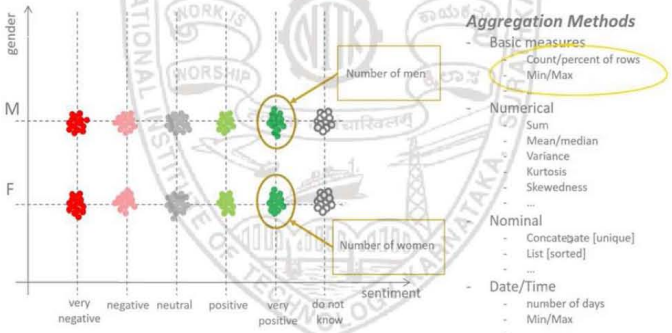

## Aggregation - Example

### Step 2: Aggregation



## Aggregation - Example

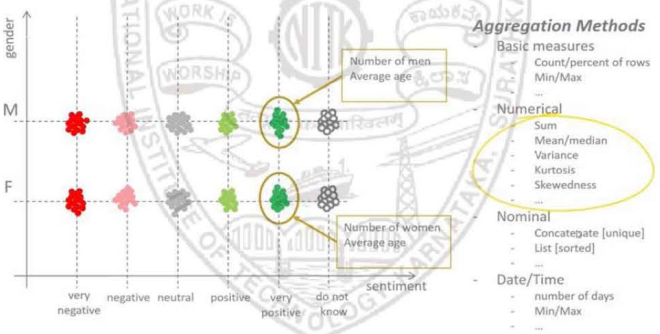
- **Step 2: Aggregation**






## Aggregation - Example

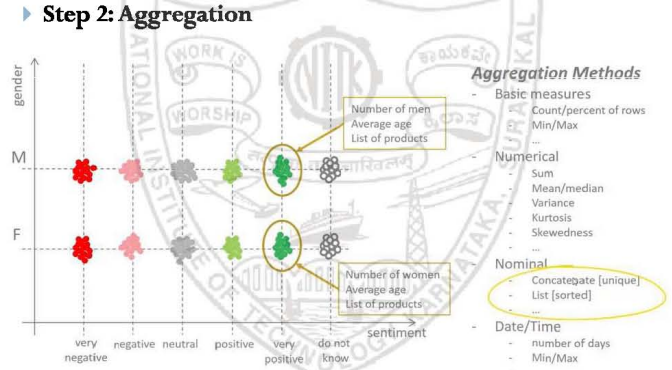
- **Step 2: Aggregation**






## Aggregation - Example

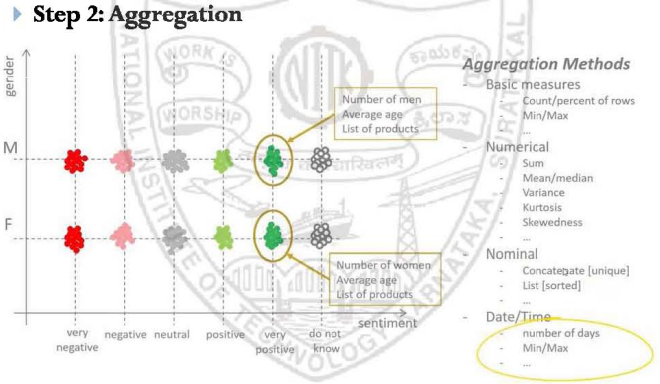
- **Step 2: Aggregation**






## Aggregation - Example

- **Step 2: Aggregation**






## Clustering for Data Discretization

## Clustering for Data Discretization

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group.
  - dissimilar (or unrelated) to the objects in other groups.

## Clustering for Data Discretization

- **Cluster:** A collection of data objects
  - similar (or related) to one another within the same group.
  - dissimilar (or unrelated) to the objects in other groups.
- **Cluster analysis** (or *clustering, data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
    - Data is discretized!

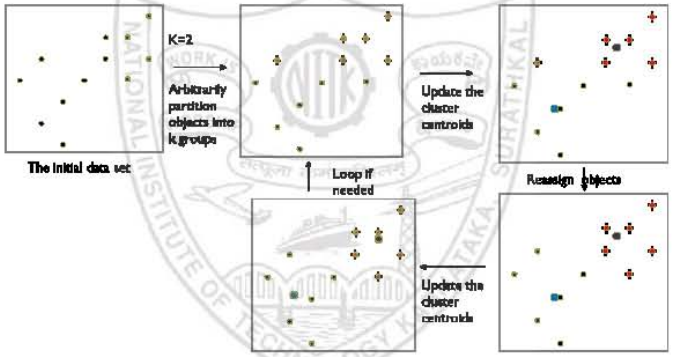


## K-Means Clustering

- Given $k$, the k-means algorithm is implemented in four steps:
  - Partition objects into $k$ nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when the assignment does not change.



## K-Means Clustering - illustration






## Data smoothing - summary

- Discretized dataset is more compact, easy to be understand/use.
- Improves efficiency.
  - Continuous feature spaces drastically slow learning algorithms.
- Many algorithms discretize as part of the learning process.
  - E.g. clustering, pattern analysis, data mining



## References


- Berka, P., & Bruha, I. (1998). Discretization and grouping: Preprocessing steps for data mining. In *Principles of Data Mining and Knowledge Discovery: Second European Symposium, PKDD'98 Nantes, France, September 23–26, 1998 Proceedings 2* (pp. 239-245). Springer Berlin Heidelberg.
- Yang, Ying, Geoffrey I. Webb, and Xindong Wu. "Discretization methods." *Data mining and knowledge discovery handbook*: 101-116.
- Dougherty, James, Ron Kohavi, and Mehran Sahami. "Supervised and unsupervised discretization of continuous features." In *Machine learning proceedings 1995*, pp. 194-202. Morgan Kaufmann, 1995.