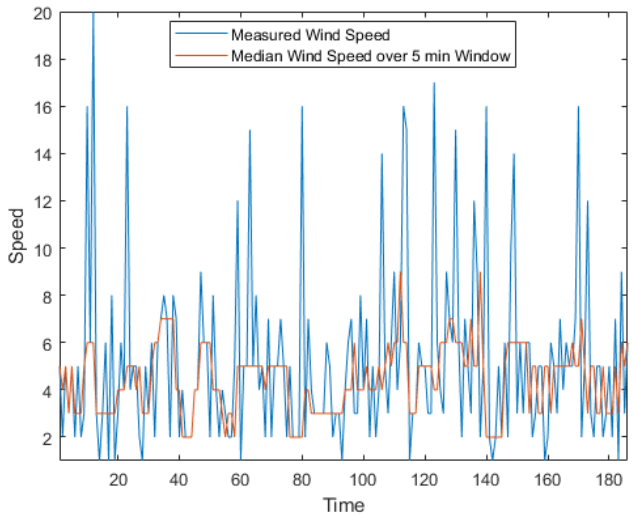
Data Transformation in Data Mining

- **It** is done for combining unstructured data with structured data to analyze it later.
- It is also important when the data is transferred to a new cloud data warehouse.
    - As the homogeneous and well-structured data are easier to analyze and look for patterns.
    - For example, a company has acquired another firm and now has to consolidate all the business data. The smaller company may be using a different database than the parent firm. Also, the data in these databases may have unique IDs, keys and values. All this needs to be formatted so that all the records are similar and can be evaluated.
- This is why data transformation methods are applied. And, they are described below:

Data Smoothing

- This method is used for removing the noise from a dataset.
- Noise is referred to as the distorted and meaningless data within a dataset.
- Smoothing uses algorithms to highlight the special features in the data.
- After removing noise, the process can detect any small changes to the data to detect special patterns.
- When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
    - Any data modification or trend can be identified by this method.



Smoothing using binning

Unsorted data for price in dollars
Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

First of all, sort the data
After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

1. Smoothing the data by equal frequency bins

Bin 1: 8, 9, 15, 16
Bin 2: 21, 21, 24, 26,
Bin 3: 27, 30, 30, 34

- **Smoothing by bin means**
  For Bin 1:
  (8+ 9 + 15 +16 / 4) = 12
  Bin 1 = 12, 12, 12, 12

  For Bin 2
  (21 + 21 + 24 + 26 / 4) = 23
  Bin 2 = 23, 23, 23, 23

  For Bin 3:
  (27 + 30 + 30 + 34 / 4) = 30
  Bin 3 = 30, 30, 30, 30

- **Smoothing by bin boundaries**
  - pick the minimum and maximum value.
  - Put the minimum on the left side and maximum on the right side.
  - Middle values in bin boundaries move to its closest neighbor value with less distance.

Before bin Boundary: Bin 1: 8, 9, 15, 16
Here, 8 is the minimum value and 16 is the maximum value. 9 is near to 8, so 9 will be treated as 8. 15 is more near to 16 and farther away from 8. So, 15 will be treated as 16.

After bin Boundary: Bin 1: 8, 8, 16, 16

Before bin Boundary: Bin 2: 21, 21, 24, 26,
After bin Boundary: Bin 2: 21, 21, 26, 26,

Before bin Boundary: Bin 3: 27, 30, 30, 34
After bin Boundary: Bin 3: 27, 27, 27, 34

Pros of Smoothing :

- Data smoothing clears the understandability of different important hidden patterns in the data set.
- Data smoothing can be used to help predict trends. Prediction is very helpful for getting the right decisions at the right time.

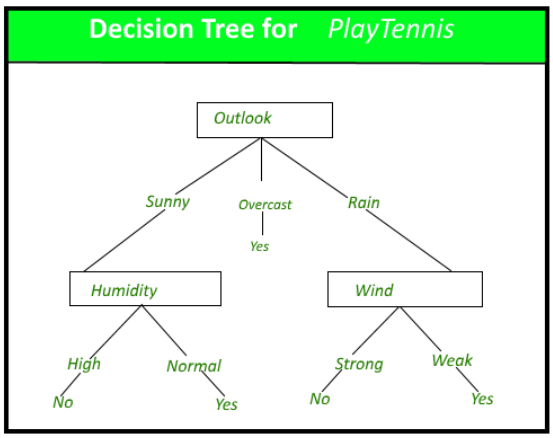
- Data smoothing helps in getting accurate results from the data.

Cons of data smoothing

- Data smoothing doesn’t always provide a clear explanation of the patterns among the data.
- It is possible that certain data points being ignored by focusing the other data points.

Discretization

- This is a process of converting continuous data into a set of data intervals.
- Continuous attribute values are substituted by small interval labels.
- This makes the data easier to study and analyze.
- If a continuous attribute is handled by a data mining task, then its discrete values can be replaced by constant quality attributes.
- This improves the efficiency of the task.
- This method is also called data reduction mechanism as it transforms a large dataset into a set of categorical data.
- Discretization also uses decision tree-based algorithms to produce short, compact and accurate results when using discrete values.



- It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.

- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

    - For **example**, (1-10, 11-20) (age:- young, middle age, senior).

Generalization

- In this process, low-level data attributes are transformed into high-level data attributes using concept hierarchies.
- This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data.
- For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old).
- Data generalization can be divided into two approaches – data cube process (OLAP) and attribute oriented induction approach (AOI).

- It converts low-level data attributes to high-level data attributes using concept hierarchy.

    - For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).
    - For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

Attribute construction

- In the attribute construction method, new attributes are created from an existing set of attributes.
- For example, in a dataset of employee information, the attributes can be employee name, employee ID and address. These attributes can be used to construct another dataset that contains information about the employees who have joined in the year 2019 only.
- This method of reconstruction makes mining more efficient and helps in creating new datasets quickly.
- Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

Normalization

- Also called data pre-processing, this is one of the crucial techniques for **data transformation in data mining.**
- Here, the data is transformed so that it falls under a given range.
- When attributes are on different ranges or scales, data modelling and mining can be difficult. Normalization helps in applying data mining algorithms and extracting data faster.
- Data normalization involves converting all data variable into a given range. Techniques that are used for normalization are:

The popular normalization methods are:

1. Min-max normalization
2. Decimal scaling
3. Z-score normalization