

DS 440 Data Mining

Lecture 7: Data quality and preprocessing

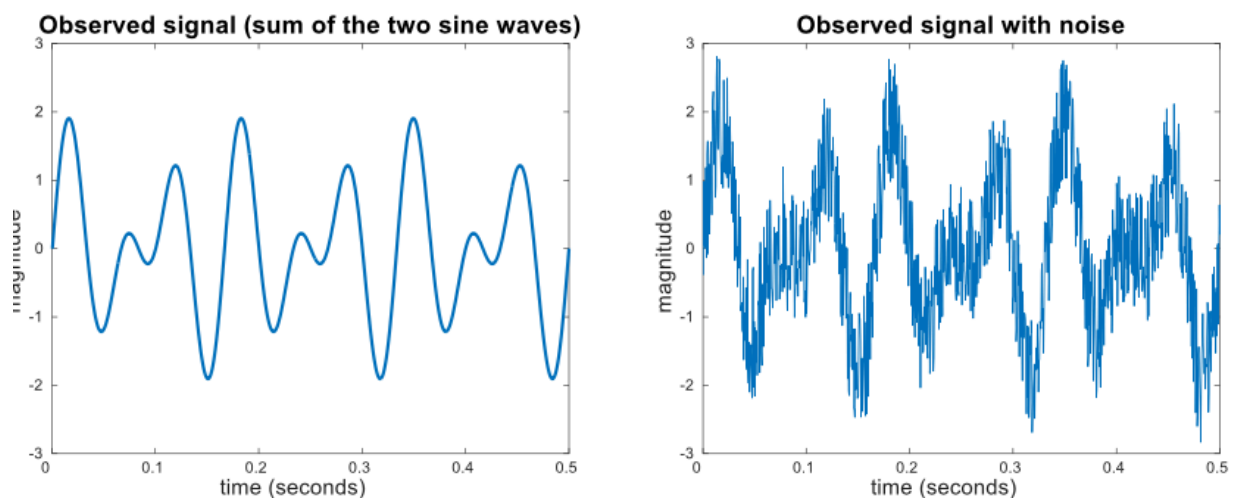
Poor data quality negatively affects many data processing efforts. For example, if a classification model for detecting people who are loan risks is built using poor data, the:

- Some credit-worthy candidates are denied loans
- More loans are given to individuals that default

1. Major data quality problems

1.1 Noise

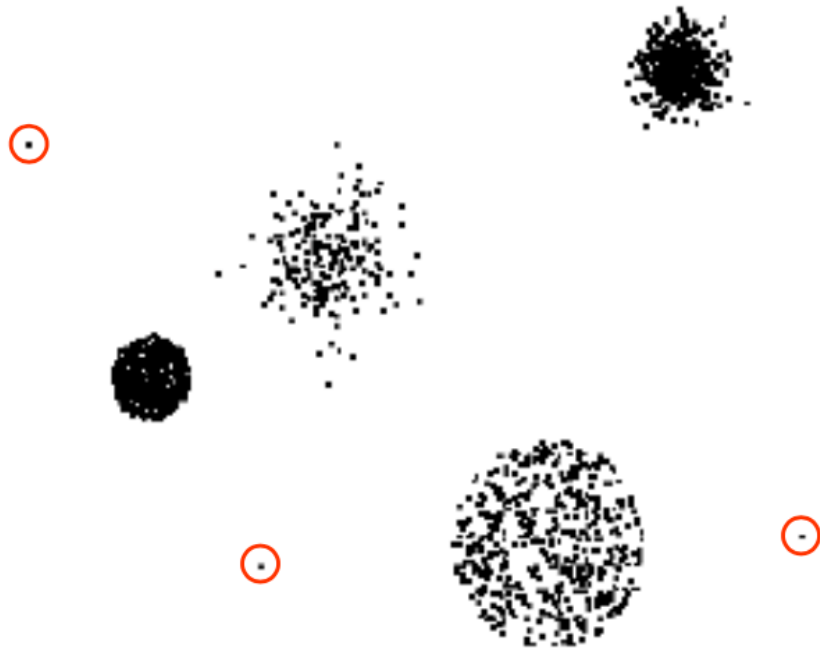
Signal distorted by the presence of noise



1.2 Outliers

Outlier detection can have 2 objectives:

1. Outliers are noise that interferes with data analysis. For example measuring daily temperature in daytona and obtaining 500F
2. Outliers are the goal of our analysis such as in credit card fraud detection



1.3 Wrong data

For example we are recording daily temperature in Fahrenheit, and mistakingly record data in Celsius for a few days

1.4 Fake data

For example filling gaps in a dataset using a machine learning model. If this model is not suited for the dataset, it can negatively effect the entire analysis

1.5 Missing values

For example people refuse to answer some fields in a public survey. Then in the final collected dataset these fields appear as missing values for these people.

1.6 Duplicate data

For example same person has multiple email addresses. **Whether two samples are duplicate or approximately duplicate can be quantified using the distance/similarity measure discussted in the last class.**

2. Data Preprocessing

2.1 Data Aggregation

It can be of 2 types:

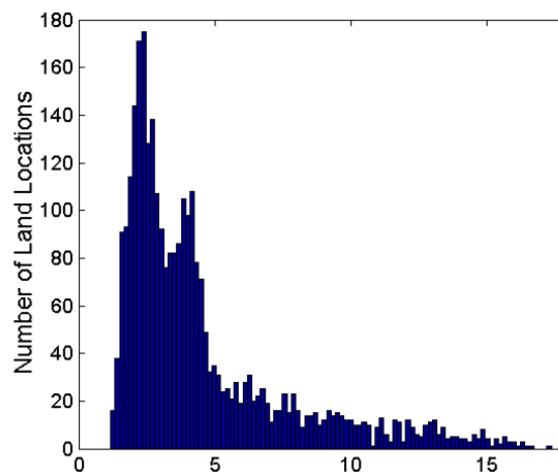
1. Combine multiple samples to form 1 new sample
2. Combine multiple features to form a new feature

The idea of data aggregation is to reduce the size of the dataset which can help the data mining algorithms in several ways including

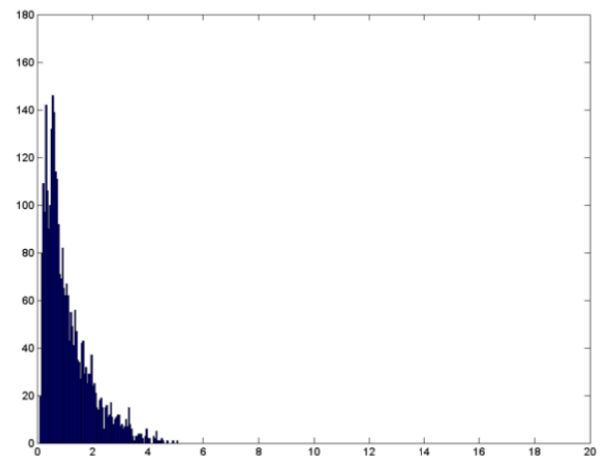
1. Leads to faster computation
2. Leads to more stable algorithms as more data means more noise

However, an obvious downside of data aggregation is that we are losing information. Hence, we should be careful while engaging in this kind of preprocessing

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



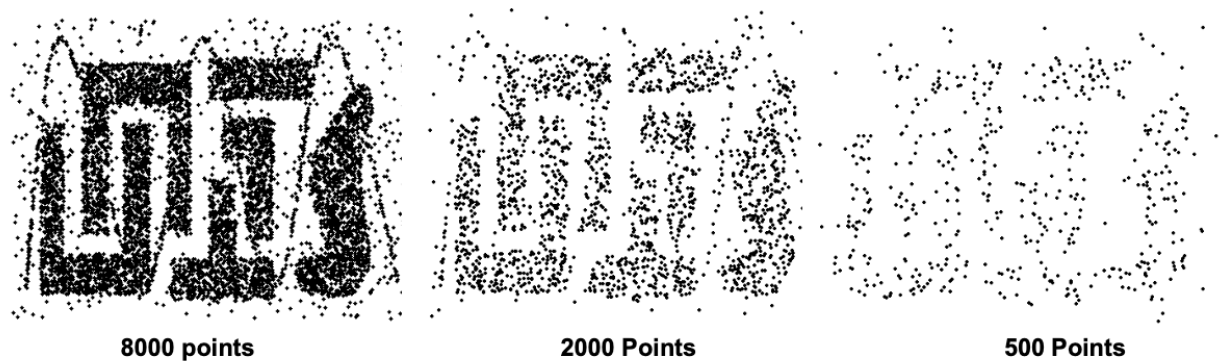
Standard Deviation of Average Yearly Precipitation

2.2 Sampling

1. Sampling is the main technique employed for data reduction.
2. Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

2.2.1 Key principle:

- Using a sample will work almost as well as using the entire data set, if the sample is representative
- A sample is representative if it has approximately the same properties (of interest) as the original set of data



2.2.2 Types of sampling

1. Simple random sampling:

There is an equal probability of selecting any particular sample. It is further divided into 2 categories:

1. **Sampling without replacement:** As an object is selected, it is removed from the population
2. **Sampling with replacement:** Objects are not removed from the population as they are selected to be withdrawn.

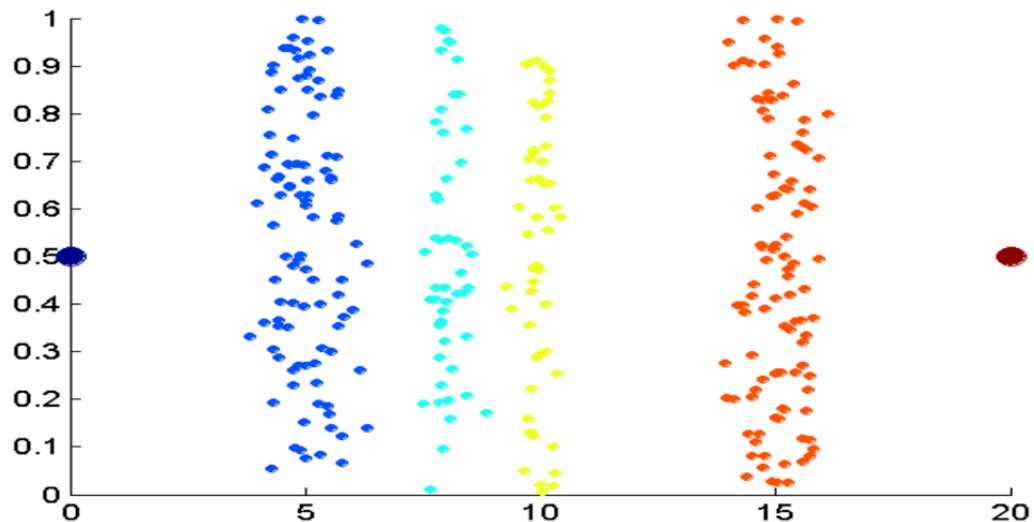
2. Stratified sampling:

Split the data into several partitions; then draw random samples from each partition

2.3 Discretization

Discretization is the process of converting a continuous attribute into an discrete attribute. A potentially infinite number of values are mapped into a small number of categories

Mapping samples with multiple unique x-values into 4 classes with 2 outliers



Exercise In the above figure, look along x-axis and y-axis individually and comment on the direction where you see more well defined clusters. How can we use this ?

2.4 Attribute transformation

An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values. For example: x^2 , e^x etc

$$x^2 : X : [3, 4] \rightarrow X' : [9, 16]$$

Most commonly used transformations:

2.4.1 Normalization:

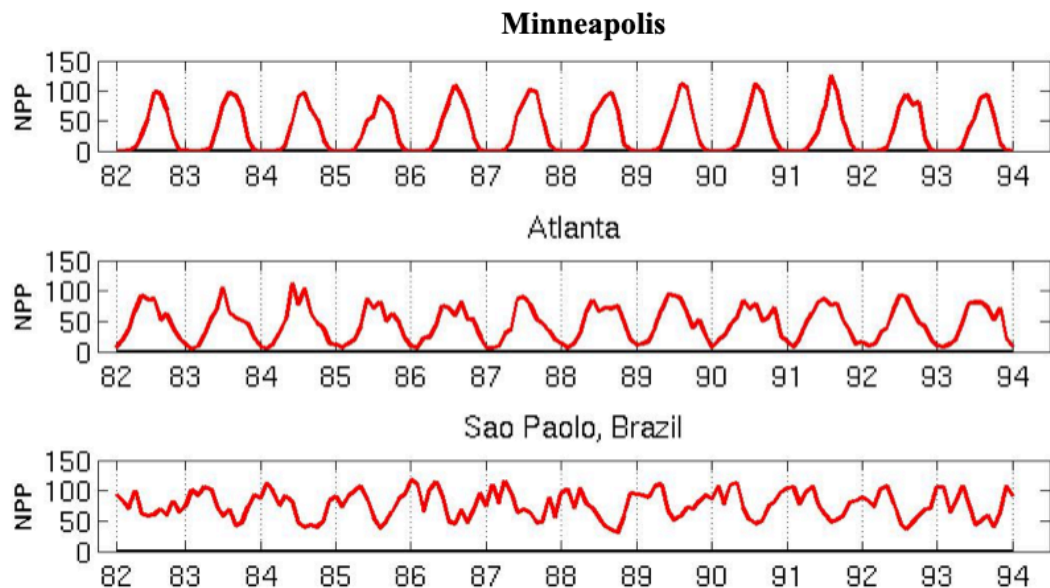
Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range

2.4.2 Standardization:

It refers to subtracting off the means and dividing by the standard deviation

2.4.3 Why Normalization/Standardization is important

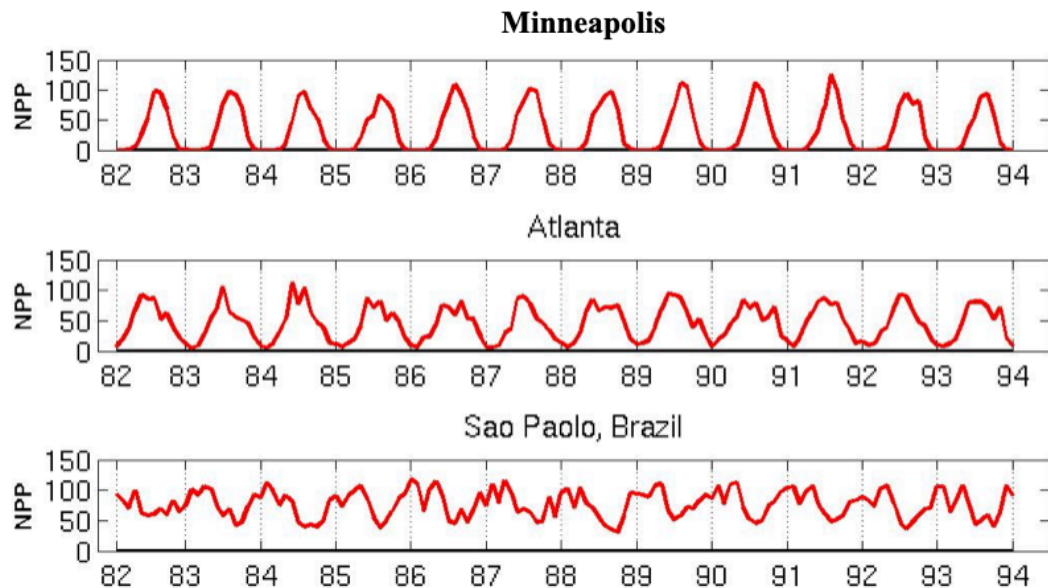
Before Standardization



Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

After Standardization



Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000

In the above example, before standardization, seasonality accounted for much of the correlation in the Net Primary Production (NPP).

2.5 Feature Subset selection

We sometimes need to select useful features from the full set of features due to following reasons:

1. **Redundant features:** For example purchase price of a product and the amount of sales tax paid
2. **Irrelevant features:** For example students' ID is often irrelevant to the task of predicting students' GPA

In []:

References:

1 <https://www-users.cse.umn.edu/~kumar001/dmbook/index.php> (<https://www-users.cse.umn.edu/~kumar001/dmbook/index.php>)

In []: