Data Mining

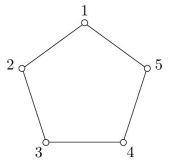
Lecture 6

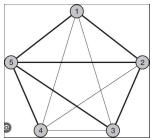
Ananya Jana
CS360

Fall 2024

Density

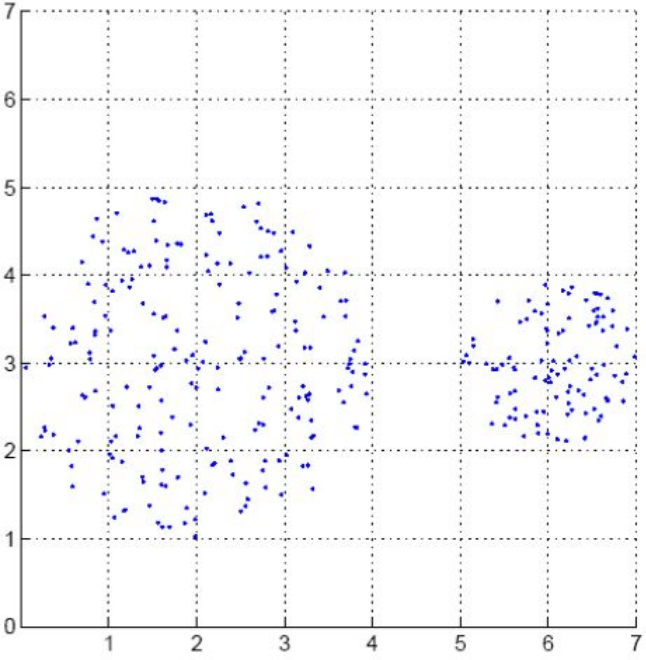
- Measures the degree to which data objects are close to each other in a specified area
- The notion of density is closely related to that of proximity
- Concept of density is typically used for clustering and anomaly detection
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
    - Estimate what the distribution of the data looks like
  - Graph-based density
    - Connectivity





Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

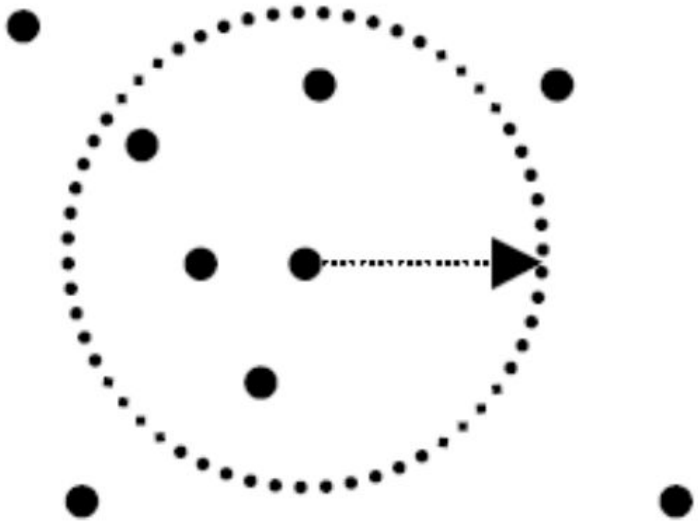


Illustration of center-based density.

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Consider a data set consisting of transactions (data objects) recording the daily sales of products in various store locations (Minneapolis, Chicago, Paris, . . .) for different days over the course of a year(See Table). One way to aggregate transactions for this data set is to replace all the transactions of a single store with a single storewide Transaction.

From this viewpoint, aggregation is the process of eliminating attributes, such as the type of item, or reducing the number of values for a particular attribute; e.g., reducing the possible values for date from 365 days to 12 months.

Sampling

Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed.

The motivations for sampling in statistics and data mining are often different.

Sampling

- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

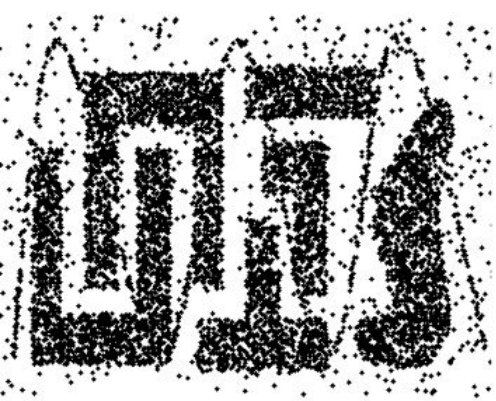
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

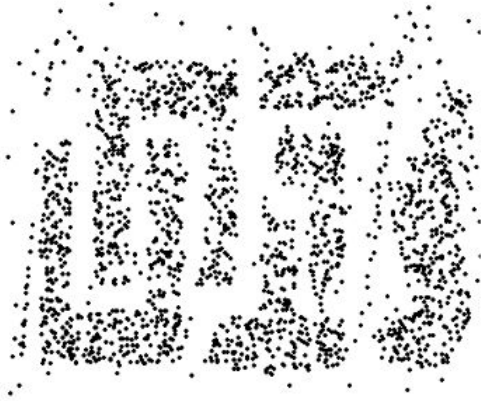
Sampling …

- The key principle for effective sampling is the following:
  - 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

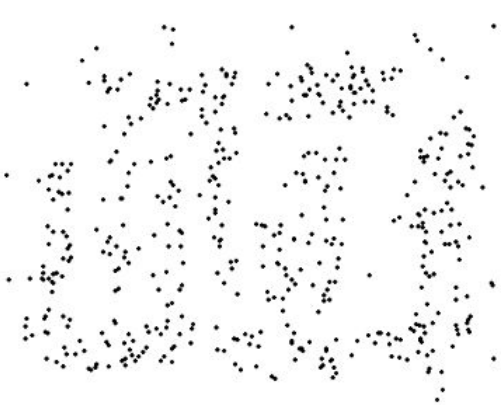
Sample Size



8000 points



2000 Points



500 Points

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Types of Sampling

Stratified Sampling: When the population consists of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all object types.

For example, when building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample

Stratified sampling can be done in two ways:

a) equal numbers of objects are drawn from each group even though the groups are of different sizes.

b) the number of objects drawn from each group is proportional to the size of that group.

Curse of Dimensionality

The curse of dimensionality refers to the phenomenon that many types of data analysis become significantly harder as the dimensionality of the data Increases.

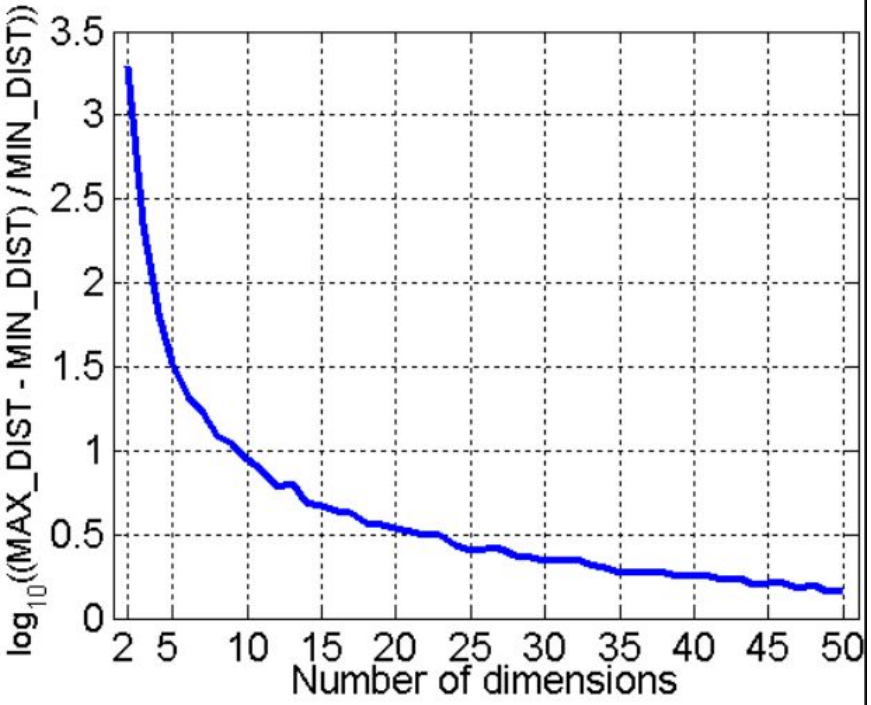
For classification, this can mean that there are not enough data objects to allow the creation of a model that reliably assigns a class to all possible objects.

For clustering, the definitions of density and the distance between points, which are critical for clustering, become less meaningful.

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

Curse of Dimensionality

Python code for showing curse of dimensionality (ipynb file on Blackboard)

```
import numpy as np
import matplotlib.pyplot as plt
import os
import math
```

```
values = []
for N in range(2,50):
    # Generate 1000 random points in N dimensions.
    P = [np.random.randint(-100, 100, N) for _ in range(1000)]
    # Generate 1 random point P2 in N dimensions.
    P2 = np.random.randint(-100,100,N)
    # calculate the difference between the set of points P and the random point P2
    diffs = [np.linalg.norm(p-P2) for p in P]
    max_d = max(diffs)
    min_d = min(diffs)
    value = math.log10(max_d-min_d)/min_d
    values.append( value )
```

```
plt.plot(range(2,50),values)
plt.xlabel('Number of dimensions')
plt.ylabel('Values')
plt.show()
```

Dimensionality Reduction

Data sets can have a large number of features. Consider a set of documents, where each document is represented by a vector whose components are the frequencies with which each word occurs in the document.

In such cases,there are typically thousands or tens of thousands of attributes(components), one for each word in the vocabulary.

The term dimensionality reduction is often reserved for those techniques that reduce the dimensionality of a data set by creating new attributes that are a combination of the old attributes. The reduction of dimensionality by selecting new attributes that are a subset of the old is known as feature subset selection or feature selection.

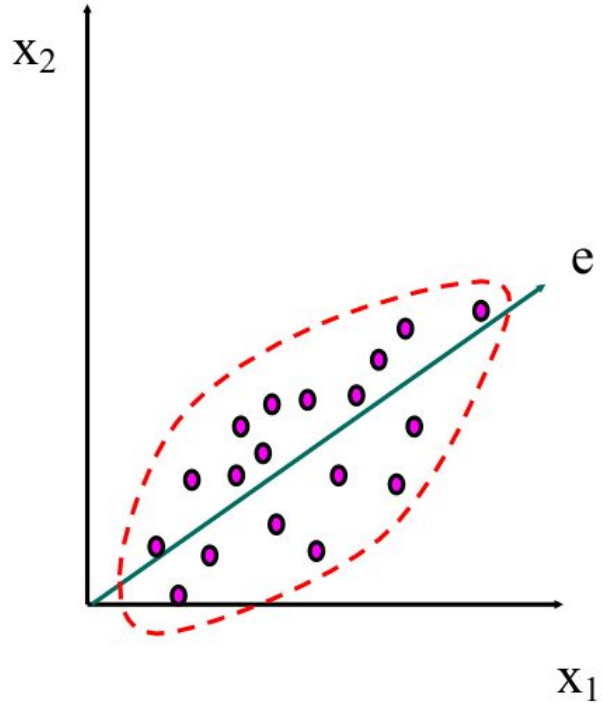
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

Difference between PCA and Linear Regression:

PCA is a method of dimensionality reduction. The goal is to transform a high dimensional data into a low dimensional data by retaining as much variance as possible. This method finds new axes in the directions where variance is the most with the first principal component accounting for the most variance and subsequent components explaining less.

Linear Regression is a predictive modelling technique which is used to model the relationship between one or more independent variables(features) and a dependent variable(target). It estimates how the dependent variable changes as the independent variable changes , by fitting a linear equation to the observed data.

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

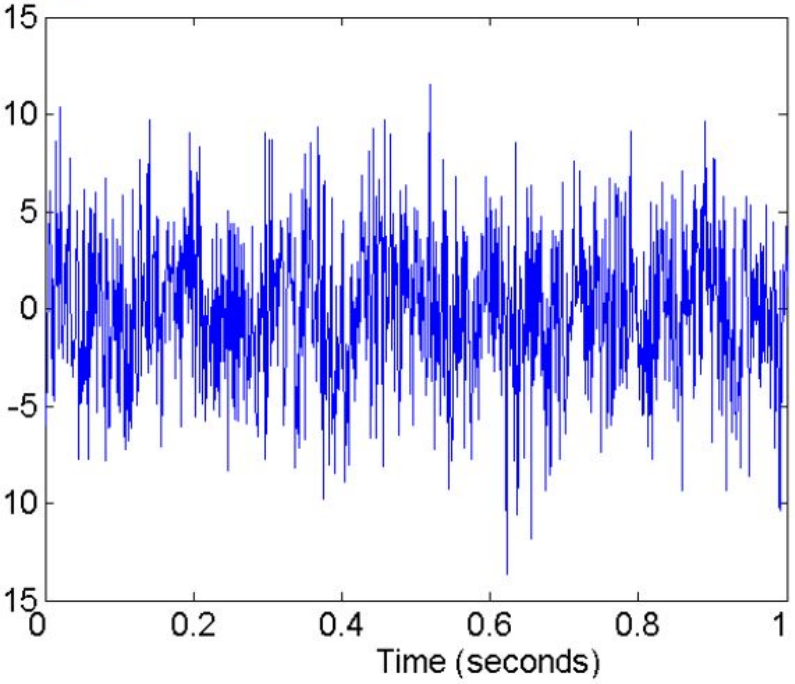
Feature Creation

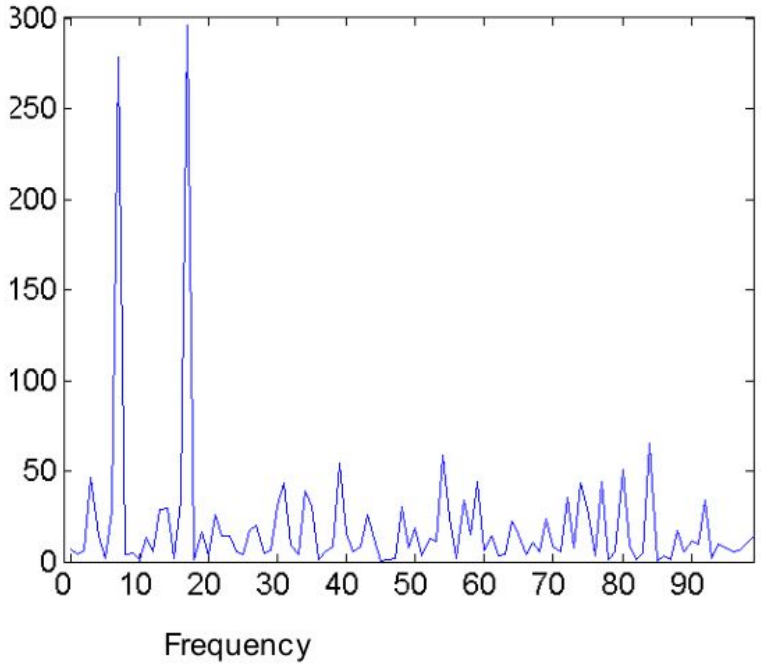
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier and wavelet transform





Two Sine Waves + Noise **Frequency**

Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values
  - We give an illustration of the usefulness of discretization using the Iris data set

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Binarization

Conversion of a categorical attribute to three binary attributes

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

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
  - Simple functions: x^k, log(x), e^x, |x|
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation