Applied Analytics and Predictive Modeling

Spring 2020

Lecture-3

Lydia Manikonda

manikl@rpi.edu

Rensselaer Polytechnic Institute • 1824 • Rensselaer

Some of the slides adapted from Intro to Data Mining Tan et al. 2nd edition

Today’s agenda

- Data Preprocessing
- Dimensionality Reduction
- Including class exercises
- Case study-1

Overview

What is data?

- Collection of **data objects** and their **attributes**
- According to Tan et al.,

- An **attribute** is a property or characteristic of an object
  - Also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Also known as tuple, record, point, case, sample, etc.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

More views of data

- Data may have parts
- The different parts of data may have relationships
- More generally, data may have structure
- Data can be incomplete

Attribute values

- Attribute values are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: Height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

Types of Attributes

- Nominal
  - Examples: ID numbers, zip codes, eye color
- Ordinal
  - Examples: Rankings (expertise level on a scale of 1-10), grades, height {tall, medium, short}
- Interval
  - Examples: Calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - Examples: Temperature in Kelvin, length, time, counts

Discrete and Continuous attributes

- Discrete Attribute:
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute:
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Types of datasets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Important characteristics of data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Record data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Document data

- Each document becomes a ‘term’ vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

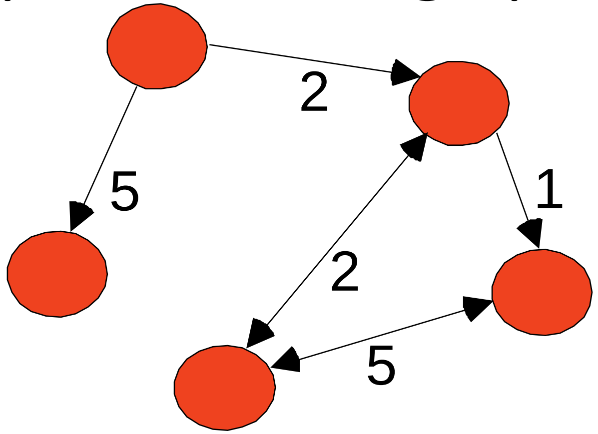
Transaction data

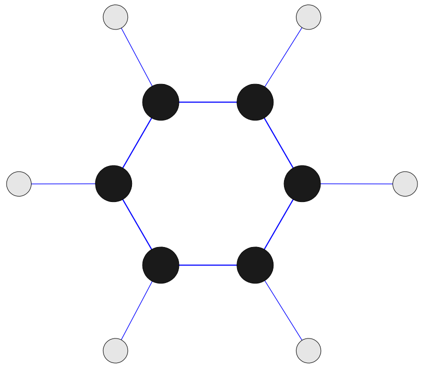
- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

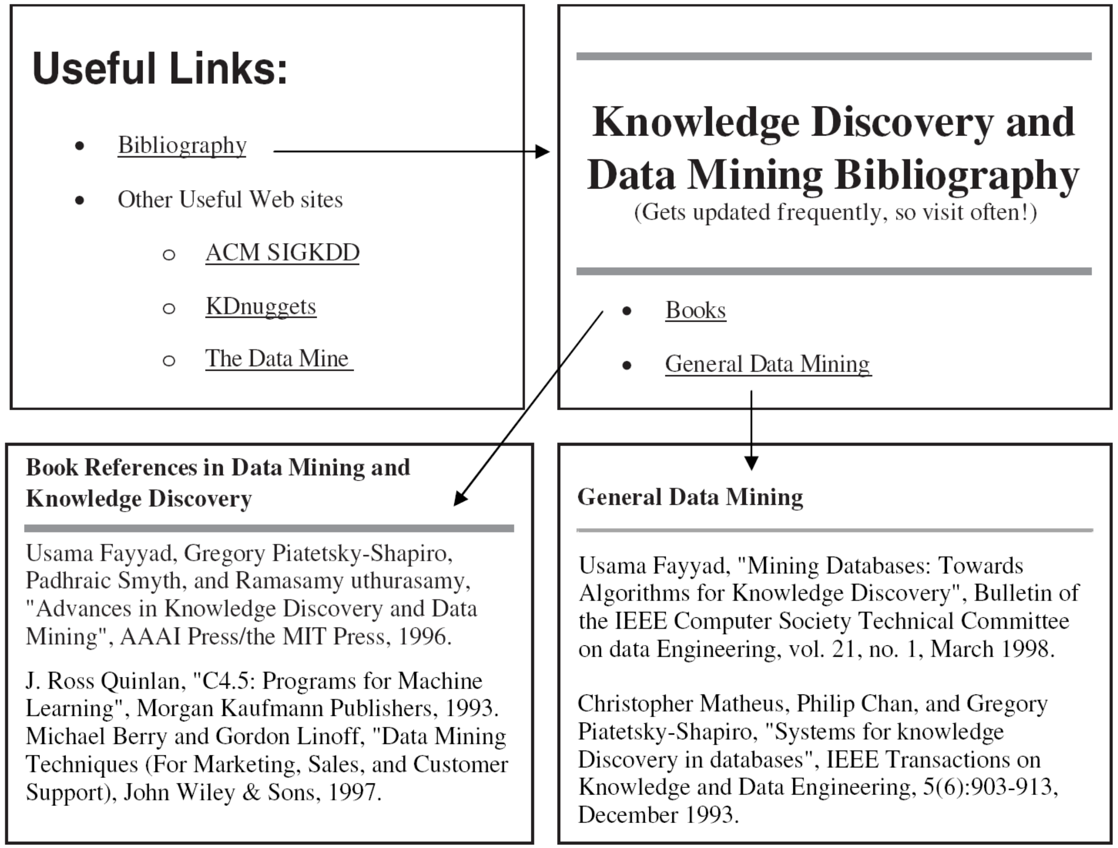
Graph Data

- Examples: Generic graph, a molecule, and webpages



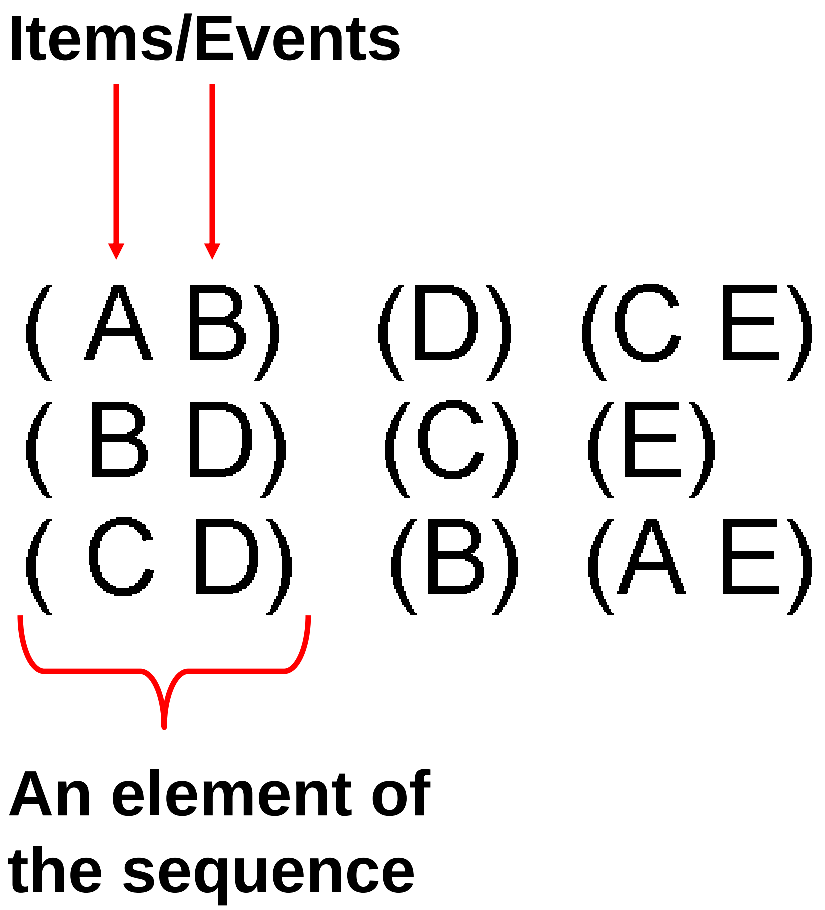


Benzene Molecule: C6H6



Ordered Data

- Sequences of transactions



Ordered Data

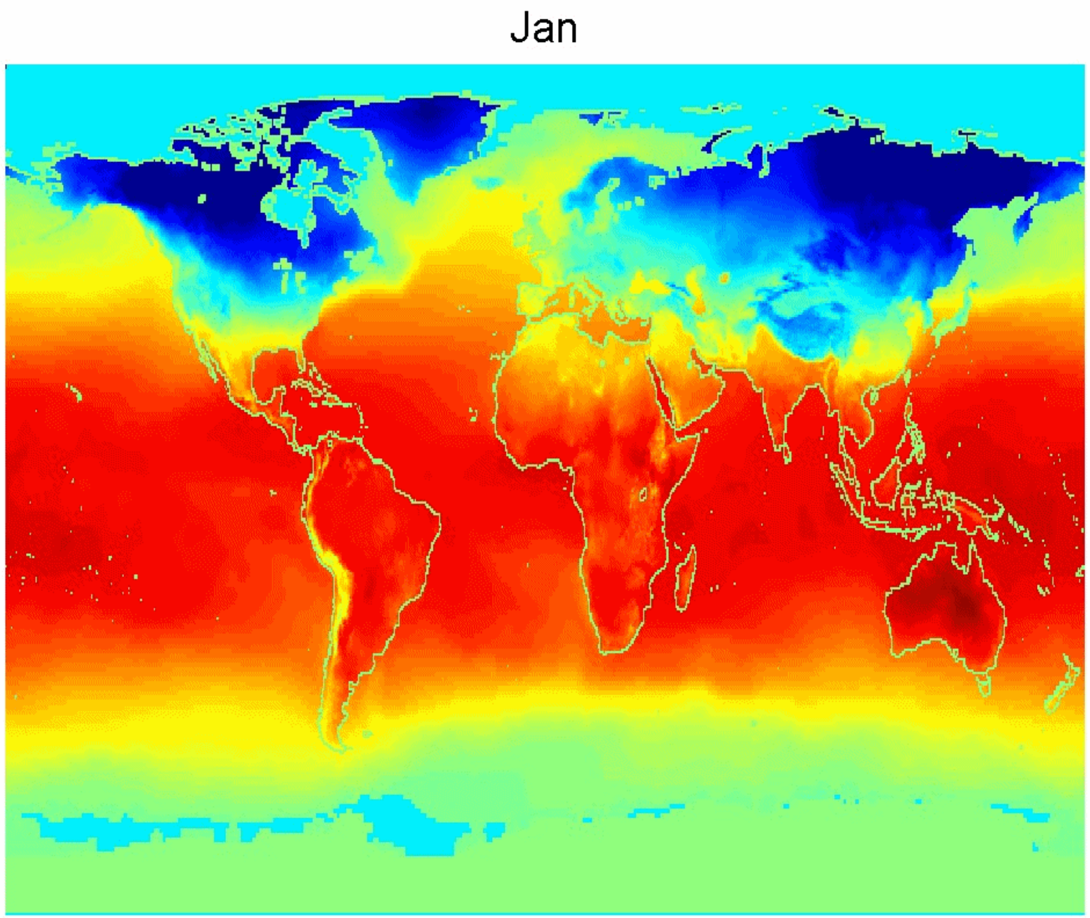
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-temporal data

Average Monthly Temperature of land and ocean



Examples

- ID numbers
  - Nominal, ordinal, or interval?
- Number of cylinders in an automobile engine
  - Nominal, ordinal, or ratio?
- Biased Scale
  - Interval or Ratio

Data Preprocessing

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

Aggregation

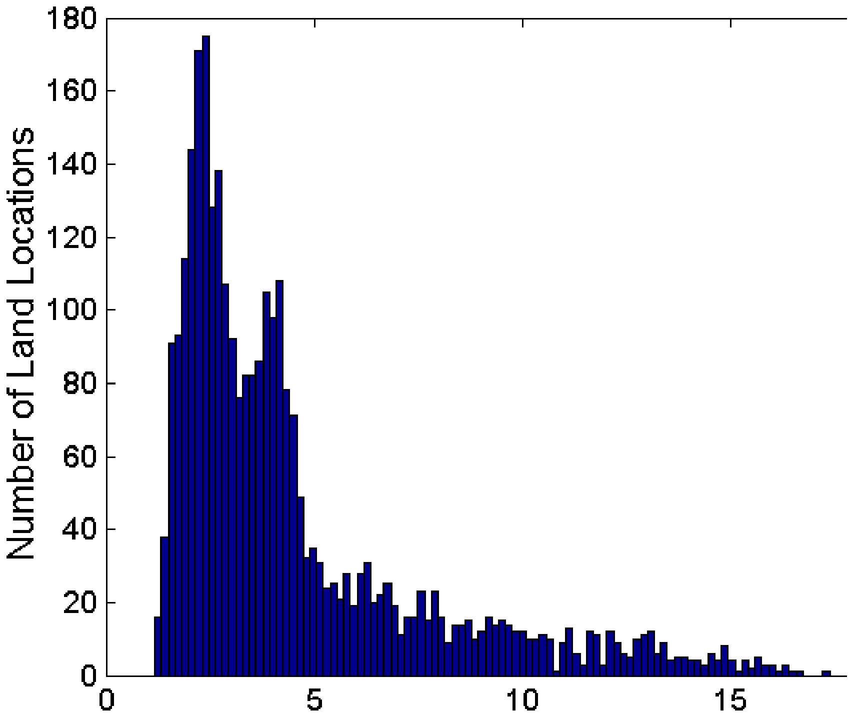
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More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

Example: Precipitation in Australia

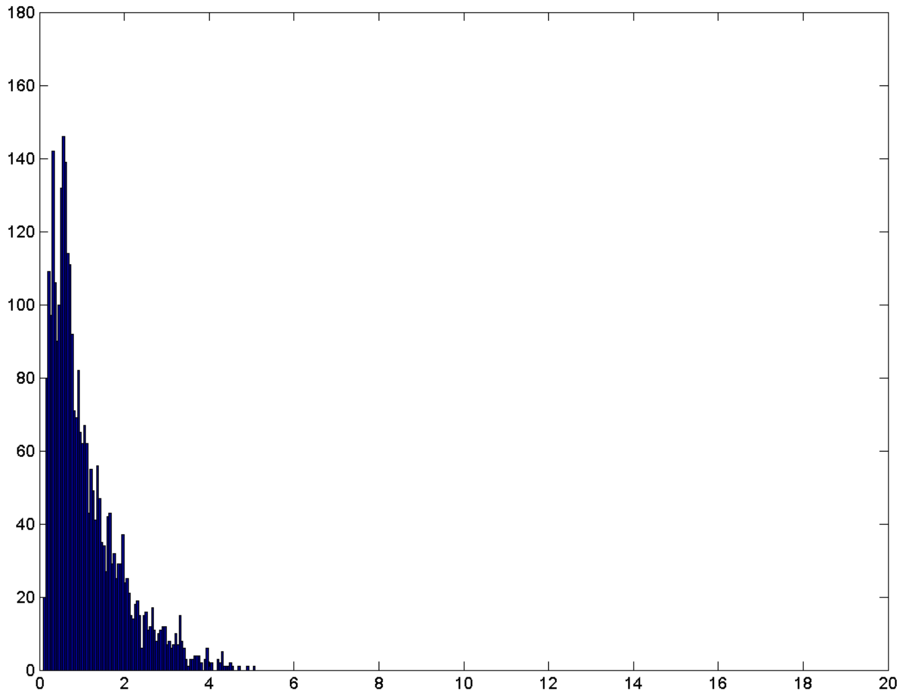
- This example is based on precipitation in Australia from the period 1982 to 1993.
- The next slide shows
  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia..

- Variation of precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

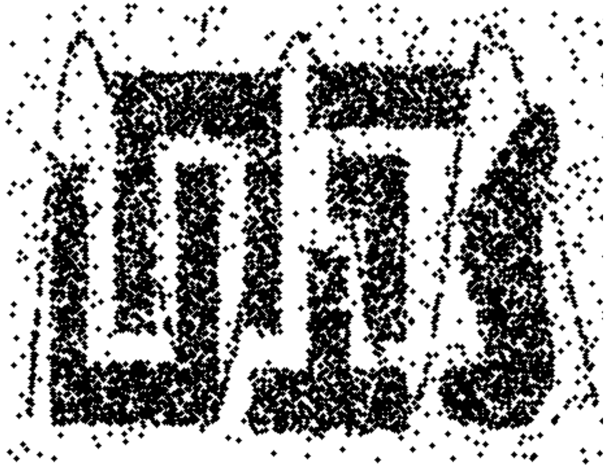
Sampling

- 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.

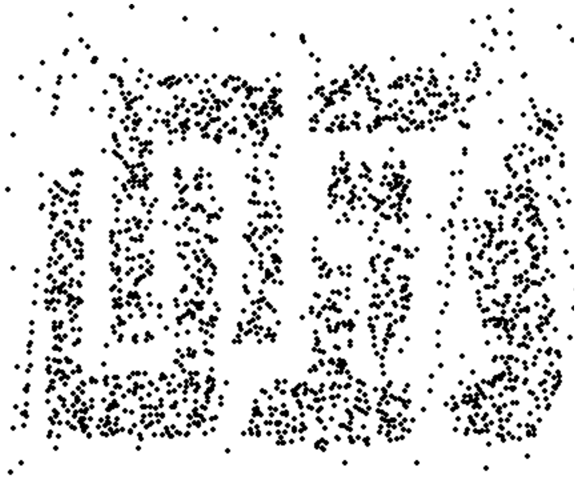
Sampling

- 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

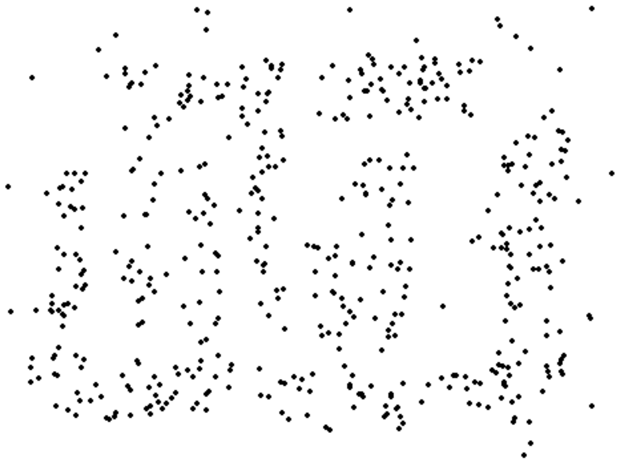
Sample size



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

Curse of dimensionality

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

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

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

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

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}

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

Exercises-1

Principal Component Analysis

Dimensionality Reduction

Eigenvalues and Eigenvectors

- The eigenvector is a vector whose direction will not be affected by a linear transformation.
- Hence eigenvectors represents the direction of largest variance of data while the eigenvalue decides the magnitude of this variance in those directions.

Computing eigenvalues and eigenvectors

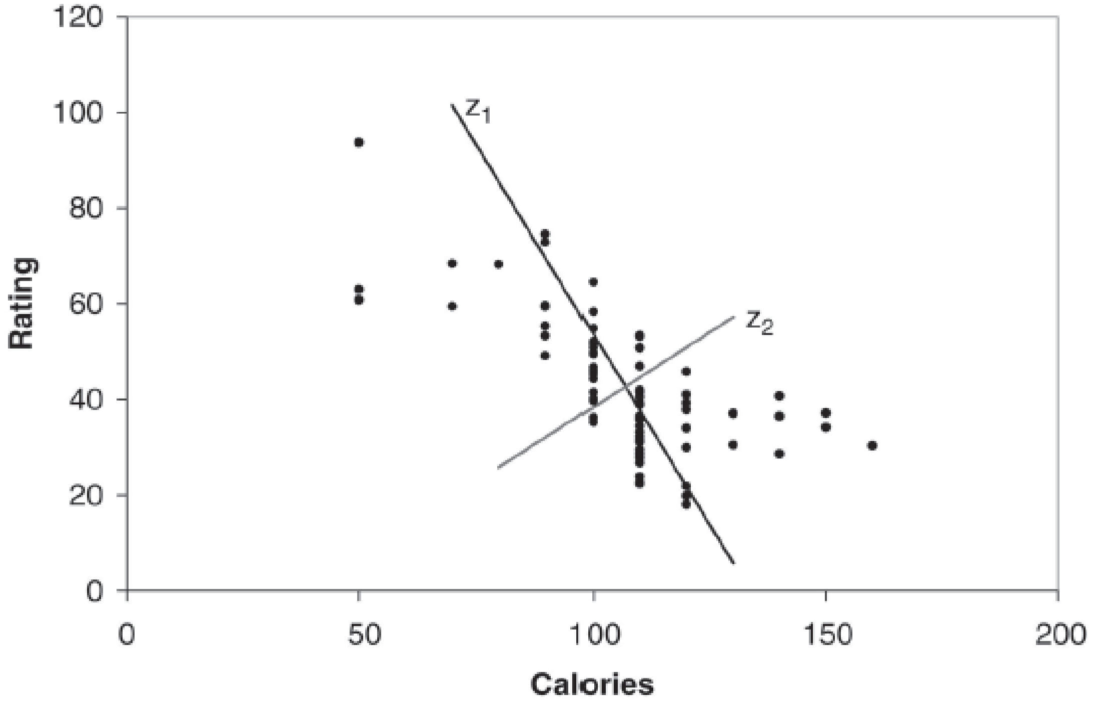
- Let A be a linear transformation represented by a matrix A
  - If there is a vector $X \in R \neq 0$ such that $AX=\lambda X$

LDA

- Step-1: Standardize the data
  - To normalize the variances of data attributes
  - Avoids biased results
  - For any attribute, one way to do is to subtract the mean and divide by standard deviation
- Step-2: Covariance matrix computation
  - How are the attributes related to each other
  - Matrix is symmetric with diagonal values are variances
- Step-3: Compute Eigenvalues and Eigenvectors (that are principal components)
  - principal components represent the directions of the data that explain a **maximal amount of variance**
  - 10-dimensional data gives you 10 principal components
  - 1st component has the maximum information followed by 2nd component and so on.

Using linear combinations to redistribute the variability

- Z1 and Z2 are two linear combinations
- Z1 has the highest variation or spread of values
- Z2 has the lowest variation



Exercises-2

Case Study-1