Module 2

Know Your Data

Getting to Know Your Data

- *Data Objects and Attribute Types*
- *Basic Statistical Descriptions of Data*
- *Measuring Data Similarity and Dissimilarity*

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - **sales database: customers, store items, sales**
  - **medical database: patients, treatments**
  - **university database: students, professors, courses**
- Also called samples , examples, instances, data points, objects, tuples.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): A data field, representing a characteristic or feature of a data object.
  - **e.g., customer _ID, name, address**
- **Observations:** observed values for a given attribute
- **Attribute vector/feature vector:** A set of attributes that define an object.
- univariate.vs Bivariate Distribution vs Multivariate
- **Types:**
  - **Nominal**
  - **Ordinal**
  - **Interval-scaled**
  - **Ratio-scaled**

Attribute Types – Categorical/Qualitative

1. **Nominal:** categories, states, or "names of things"

   - **Hair_color = {auburn, black, blond, brown, grey, red, white}**
   - **marital status, occupation, ID numbers, zip codes**

- In the cases of nominal attributes with numeric values e.g. Cust_ID, the numbers are not intended to be used quantitatively.

- Also in case of numeric nominal attributes , values do not have any meaningful order about them.

Attribute Types – Categorical/Qualitative

2. Binary attributes

- **Nominal attribute with only 2 categories/states (0 or 1)**
  - 0: attribute is absent
  - 1: attribute is present
- **Symmetric binary: both outcomes equally important**
  - e.g., gender
- **Asymmetric binary: outcomes not equally important.**
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

- If two states are True and False ,then called as Boolean Attribute

Attribute Types – Categorical/Qualitative

3. Ordinal Attributes:

Values have a meaningful order or a ranking among them but magnitude between successive values is not known.

Ex: Size = {small, medium, large}, grades, professor rankings

- Useful for registering subjective assessments of qualities that cannot be measured objectively; thus often used in surveys for ratings.
  E.g Customer satisfaction survey

- We can compute mean and median but not mode for the ordinal attributes .

- Note:nominal, binary, and ordinal attributes are qualitative attributes.

Numeric/Quantitative Attribute

4. Numeric Attributes /Quantitative Attributes

- Represents measurable quantity in integer or real values

  4.1 Interval scaled attributes

  - Measured on a scale of **equal-sized units**
  - Values have order and can be positive, 0, or negative
- E.g., temperature in C˚or F˚, calendar dates
  - We can obtain a ranking of objects by ordering the values.
  - Also allow us to compare and quantify the difference between values.
  - No true zero-point -We can not speak of values in terms of ratio.

    e.g without a true zero point, we can't say that 10 C˚ is twice as warm as 5C˚.
- Mean ,Median and Mode

Numeric Attribute Types

4.2 Ratio scaled attributes

- Inherent **zero-point**
- The values are ordered, and we can also compute the difference between values, as well as the mean, median,and mode
- Examples: Count attributes such as years of experience and number of words attribute for a document
- attributes to measure age, weight, height and monetary quantities (e.g., you are 100 times richer with $100 than with $1).

Discrete vs. Continuous Attributes

- <u>**Discrete Attribute**</u>
  - **Has only a finite or countably infinite set of values which may or may not be represented as integers.**
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - **Sometimes, represented as integer variables**
  - **Note: Binary attributes are a special case of discrete attributes**

Discrete vs. Continuous Attributes

- **Continuous Attribute**
  - **Has real numbers as attribute values**
    - E.g., 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**

Basic Statistical Descriptions of Data

- Used to identify properties and identify which data values can be treated as noise or outliers.

- 3 areas of statistical descriptions:
  - **Measuring central tendencies**
  - **Measuring dispersion of data(How data spread out?)**
  - **Graphic display of basic statistical description**

Measuring the Central Tendency:Mean

Mean (algebraic measure) (sample vs. population):

- If we were to plot the observations for attribute, where would most of the values fall?
- The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) mean.
- Let x_1, x_2, x_3, x_4, ... x_N be a set of N values or observations, such as for some numeric attribute X, like salary.

The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Weighted mean is :

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Problem with mean is its sensitivity to extreme values.
- For skewed (asymmetric) data, a better measure of the center of data is the **median**

Measuring the Central Tendency:Median

- ***Median:***
  - ***Middle value if odd number of values, or average of the middle two values otherwise***
  - ***The median is expensive to compute when we have a large number of observations.***
  - ***Estimated by interpolation (for grouped data):***

$$\text{Median} = l + \left[\frac{\frac{n}{2} - c}{f} \right] \times h$$

- l = lower limit of median class
- n = total number of observations
- c = cumulative frequency of the preceding class
- f = frequency of each class
- h = class size

Marks	**Number of students**	**Cumulative frequency**	
0 - 20	6	0 + 6	6
20 - 40	20	6 + 20	26
40 - 60	37	26 + 37	63
60 - 80	10	63 + 10	73
80 - 100	7	73 + 7	80

Solution:

We need to calculate the cumulative frequencies to find the median.

N = sum of f = 80, N/2 = 80/2 = 40

Since n is even, we will find the average of the $n/2^{th}$ and the $(n/2 +1)^{th}$ observation i.e. the cumulative frequency greater than 40 is 63 and the class is 40 - 60. Hence, the median class is 40 - 60.

l = 40, f = 37, c = 26, h = 20

Median = l + [(n/2−c)/f] × h == 40 + [(37 - 26)/40] × 20 == 40 + (11/40) × 20 = = 40 + (220/40) = 40 + 5.5 == 45.5

Measuring the Central Tendency:Mode and midrange

- *Mode*
  - ***Value that occurs most frequently in the data***
  - ***Unimodal, bimodal, trimodal,multimodal***
  - At the other extreme, if each data value occurs only once, then there is no mode.

- Midrange
  - The **midrange** can also be used to assess the central tendency of a numeric data set.
  - It is the average of the largest and smallest values in the set.

Measuring the Central Tendency: Example

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

<u>*Claculate mean, median,mode and midrange*</u>

Mean=58 ~~avg~~

Median=54 $\frac{52+56}{2} = 54$

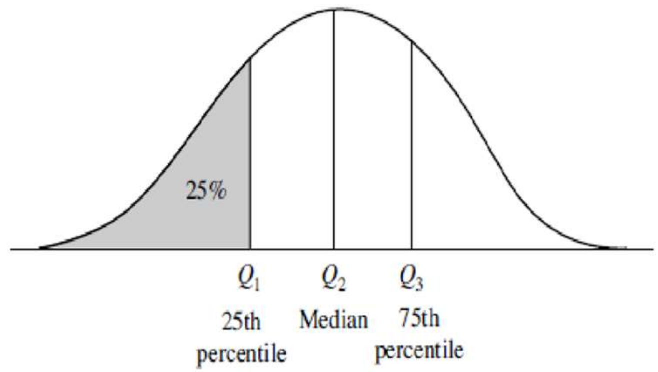
Modes = bimodal. The two modes are 52 and 70

Midrange=70 $\rightarrow \frac{min + max}{2} = \frac{30+110}{2} = 70$

Measuring the Dispersion of Data
(Range, Quartiles, Variance ,Standard Deviation, and Interquartile Range)

- **Range, Quartiles, and Interquartile Range**
  - The **range** of the data set is the difference between the largest (max()) and smallest (min()) values.
  - **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
  - The ***kth q-quantile*** for a given data distribution is the value x such that at most k/q of the data values are less than x and at the most $(q-k)/q$ of the data values are more than x, where k is an integer such that $0 < k < q$. There are q-1 q-quantiles.
  - The **2-quantile** is the data point dividing the lower and upper halves of the data distribution.It corresponds to the median
  - The **4-quantiles** are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**
  - The **100-quantiles** are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.
  - The median, quartiles, and percentiles are the most widely used forms of quantiles.

Quartiles:The quartiles are the three values that split the sorted data set into four equal parts



A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

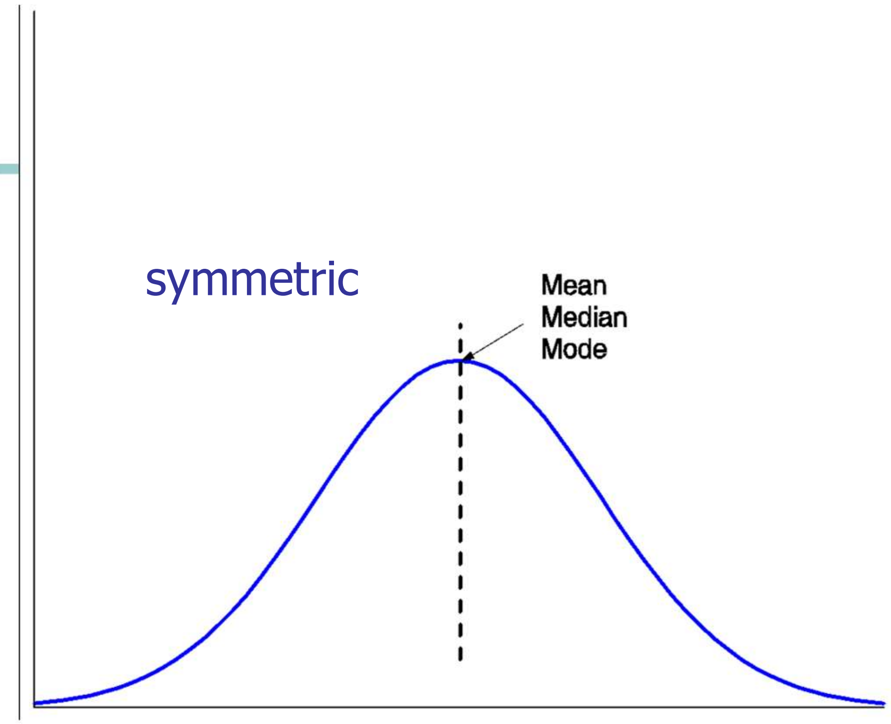
- The quartiles give an indication of a distribution's center, spread, and shape
- *Q*1, is the 25th percentile
- *Q*3, is the 75th percentile
- The second quartile Q2 is, the 50th percentile. As the median, it gives the center of the data distribution.

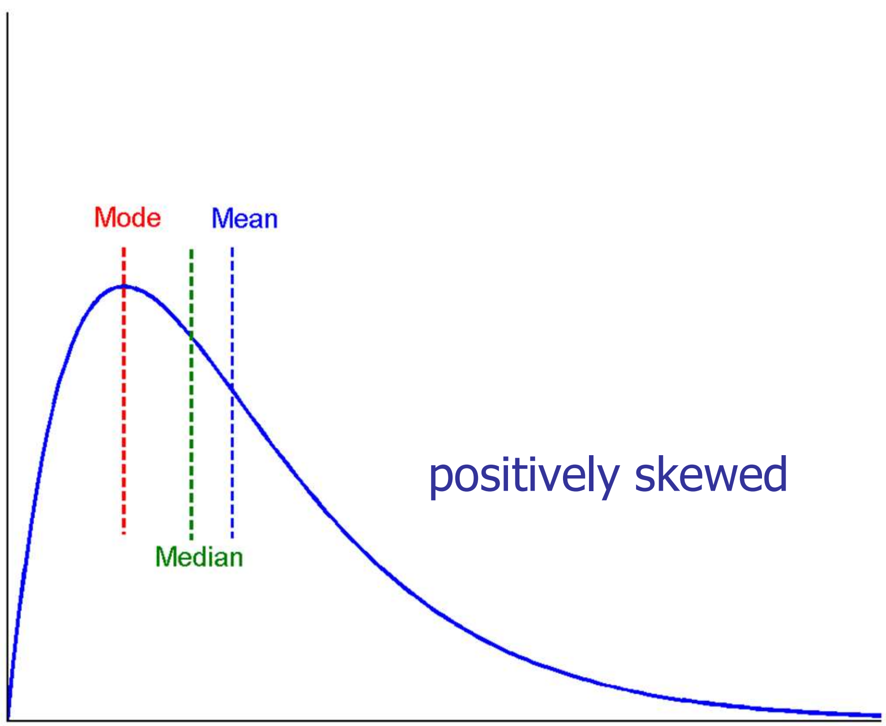
- **InterQuartile Range** (**IQR**) :The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data.
  IQR = *Q*3 -*Q*1.

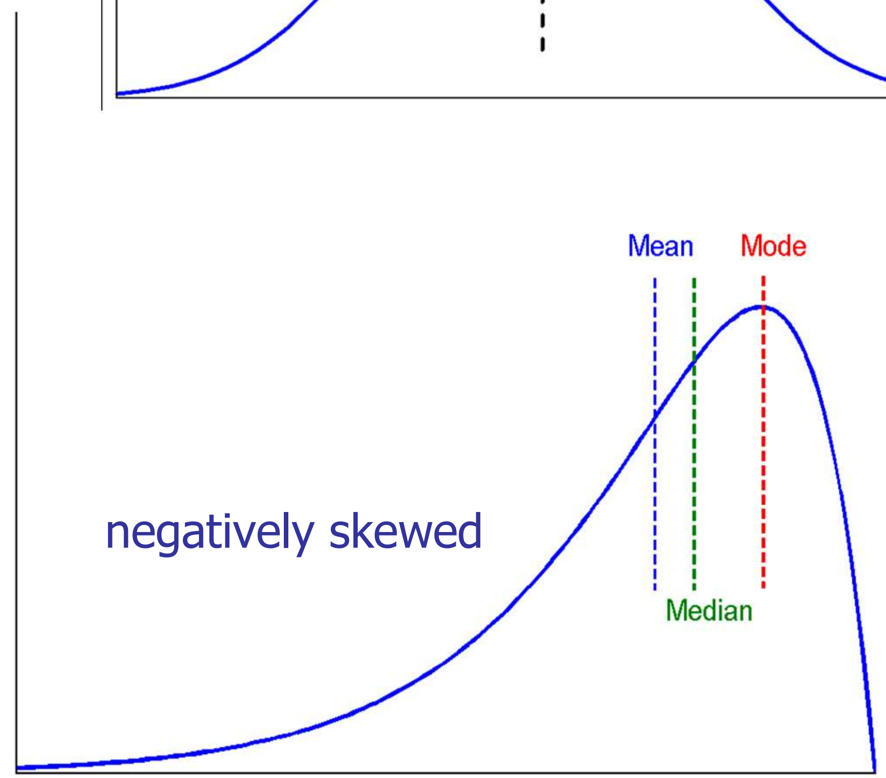
Ex: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110
the quartiles for this data are the third, sixth, and ninth values respectively, in the sorted list
Therefore, *Q*1 =47 and *Q*3 =63. Thus, the IQR=63-47 =16

Median, mean and mode of symmetric, positively and negatively skewed data

- In the symmetric distribution, the median (and other measures of central tendency) splits the data into equal-size halves.



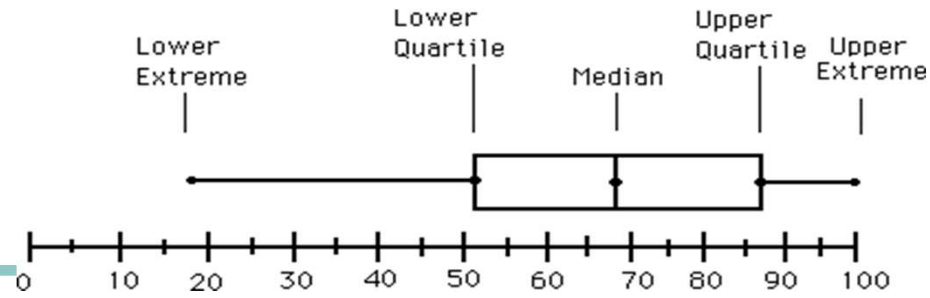




Five Number Summary

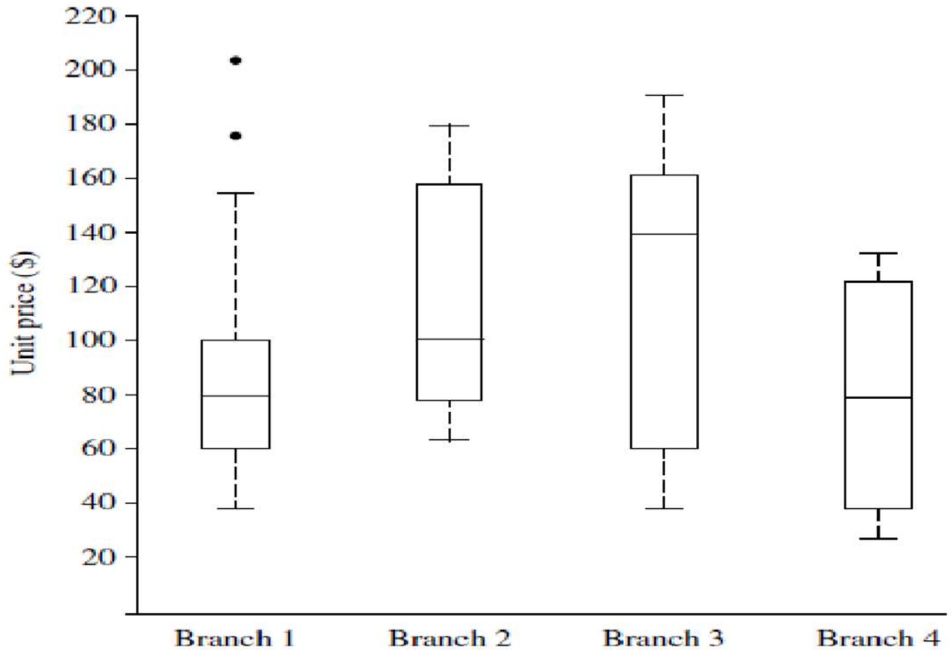
- ***Five-number summary** of a distribution* (*Minimum, Q1, Median, Q3, Maximum)*
  - It is more informative to also provide the two quartiles Q1 and Q3, along with the median
  - A common rule of thumb for identifying suspected **outliers** is to single out values falling at least 1.5*IQR* above the third quartile or below the first quartile.
  - Because *Q*1, the median, and *Q*3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the *five-number summary.*
  - The **five-number summary** of a distribution consists of the median (*Q*2), the quartiles *Q*1 and *Q*3, and the smallest and largest individual observations, written in the order of *Minimum*, *Q*1, *Median*, *Q*3, *Maximum*.

Boxplot Analysis



- **Boxplots** are a popular way of visualizing a distribution.
- A boxplot incorporates the five-number summary as follows:
  - Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
  - The median is marked by a line within the box.
  - Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.
  - When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually. To do this in a box plot, the whiskers are extended to the extreme low and high observations *only if* these values are less than 1.5*IQR* beyond the quartiles. Otherwise, the whiskers terminate at the most extreme observations occurring within 1.5*IQR* of the quartiles. The remaining cases are plotted individually.

Boxplot Analysis: Example



3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

In this boxplot :

For branch 1, we see that the median price of items sold is $80, *Q*1 is $60, and *Q*3 is $100.

Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

Variance and Standard Deviation

- They indicate how spread out a data distribution is.
- The **standard deviation**, σ, of the observations is the square root of the variance,
- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
- The **variance** of *N* observations, $x1,x2, \ldots ,xN$, for a numeric attribute *X* is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

The basic properties of the standard deviation, σ, as a measure of spread are as follows:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

Measuring the Central Tendency

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Claculate variance and standard deviation.

Mean=58

Median=54

Modes = bimodal. The two modes are 52 and 70

Midrange=70

Variance= 379.17

Std. Deviation= 19.47

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$
$$= \frac{(30 - 58)^2 + \ldots + (110 - 58)^2}{12}$$
$$= \frac{4550}{12}$$
$$= 379.1666666667$$
$$\sigma = \sqrt{379.1666666667}$$
$$= 19.472202409247$$

Problems

2.2 Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

13

13

(a) What is the *mean* of the data? What is the *median*? 30,25

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). 25,35, Bimodal

(c) What is the *midrange* of the data? 41.5

(d) Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data? 20,35

(e) Give the *five-number summary* of the data. 13,20,25,35,70

(f) Show a *boxplot* of the data.

median 1st half = 7th item = 20

median 2nd half = 35

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

Measuring Data Similarity and Dissimilarity

DOUBT

Box Plot Components:
Draw a scale on the vertical axis.
Mark the five key values (min, Q1, Q2, Q3, max) on the scale.
Join Q1 to Q3 with a vertical line to form the box.
Draw horizontal lines (whiskers) from Q1 to the min and from Q3 to the max.

slide (29- 48)

Proximity(Similarity and Dissimilarity)

Similarity

- Numerical measure of how alike two data objects are.
- Value is higher when objects are more alike
- Often falls in the range [0,1]

Dissimilarity (e.g., distance)

- Numerical measure of how different two data objects are.
- Lower when objects are more alike.
- Minimum dissimilarity is often 0.
- Upper limit varies

Data Matrix and Dissimilarity Matrix

Data matrix n-by-p matrix (n objects p attributes) : 2 mode matrix.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Data Matrix and Dissimilarity Matrix

Dissimilarity matrix : n X n Matrix

- In general, $d(i, j)$ is a non-negative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ.

- Note that $d(i, i) = 0$; i.e. the difference between an object and itself is 0. Furthermore, $d(i, j) = d(j, i)$

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

- Measures of similarity can often be expressed as a function of measures of dissimilarity. For example, for nominal data, $sim(i, j) = 1 - d(i, j)$ where $sim(i, j)$ is the similarity between objects i and j.
- Also called as **one-mode** matrix

Proximity Measure for Nominal Attributes

- A nominal attribute Can take 2 or more states,

- "How is dissimilarity computed between objects described by nominal attributes?"

Method 1: Simple matching :The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i,j)=\frac{p-m}{p}$$

Where

- p is the total number of attributes describing the objects.
- m is the number of matches (i.e., the number of attributes for which i and j are in the same state)

Proximity Measure for Nominal Attributes

Example : **Dissimilarity between nominal attributes.**

Object Identifier	test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

Here p= 1
d(*i*, *j*) evaluates to 0 if objects *i* and *j* match, and 1 if the objects differ.

$$d(i,j)=\frac{p-m}{p}$$

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}.$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

From this, we see that all objects are dissimilar except objects 1 and 4 (i.e.,d(4,1)= 0.

Proximity Measure for Nominal Attributes

Example : **Dissimilarity between nominal attributes.**

Alternatively, similarity can be computed as

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}.$$

Method 2: Use a large number of binary attributes

creating a new binary attribute for each of the M nominal states

how can we compute the dissimilarity between two binary attributes?

One approach involves computing a dissimilarity matrix from the given binary data. If all binary attributes are thought of as having the same weight, we have the 2 x 2 contingency table as shown below

Object j

Object i

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Where q = the number of attributes that equal 1 for both objects i and j,

r = the number of attributes that equal 1 for object i but equal 0 for object j,

s = the number of attributes that equal 0 for object i but equal 1 for object j,

t = the number of attributes that equal 0 for both objects i and j.

p = The total number of attributes = $q + r + s + t$.

Proximity Measure for Binary Attributes

- *A contingency table for binary data*

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- *Distance measure for symmetric binary variables:*

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- *Distance measure for asymmetric binary variables:*

$$d(i,j) = \frac{r+s}{q+r+s}$$

- *Jaccard coefficient (similarity measure for asymmetric binary variables):*

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

$= 1 - d(i,j)$

Dissimilarity between Binary Attributes

<u>*Example*</u>

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- *Gender is a symmetric attribute*
- *The remaining attributes are asymmetric binary*
- *Let the values Y and P be 1, and the value N be 0*

Suppose distance is computed based on only asymmetric attributes

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

$$d(i, j) = \frac{r+s}{q+r+s}$$

- These measurements suggest that Jim and Mary are unlikely to have a similar disease because they have the highest dissimilarity value among the three pairs.

Of the three patients, Jack and Mary are most likely to have a similar disease.

Dissimilarity of Numeric Data:

- Euclidean distance: The most popular distance measure

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

where i = (x_{i1}, x_{i2}, …, x_{ip}) and j = (x_{j1}, x_{j2}, …, x_{jp}) are two objects described by numeric attributes.

- Manhattan (or city block) distance: named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks).
- The distance between two points measured along axes at right angles

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

Dissimilarity of Numeric Data: Minkowski Distance

The Euclidean and the Manhattan distance satisfy the following mathematical properties:

- **Non-negativity:** $d(i, j) >= 0$: Distance is a non-negative number.

- **Identity of indiscernibles:** $d(i, j) = 0$:The distance of an object to itself is 0.

- **Symmetry:** $d(i, j) = d(j, i)$: Distance is a symmetric function.

- **Triangle inequality:** $d(i, j) <= d(i, k)+d(k,j)$: Going directly from object i to object j in space is no more than making a detour over any other object k.

A measure that satisfies these conditions is known as **metric**.

Example: Euclidean and Manhattan distance

Euclidean distance and Manhattan distance. Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects The Euclidean distance between the two is $\sqrt{2^2 + 3^2} = 3.61$. The Manhattan distance between the two is $2 + 3 = 5$.

Dissimilarity of Numeric Data: Minkowski Distance

Minkowski distance: A generalization of Euclidean and Manhattan distances

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where i = (x_{i1}, x_{i2}, …, x_{ip}) and j = (x_{j1}, x_{j2}, …, x_{jp}) are two objects described by p numeric attributes and h is a real number such that h >= 1.

- Also called as L_p norm where p refers to h.

Special Cases of Minkowski Distance

- *h = 1: Manhattan (city block, L_1 norm) distance*

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- *h = 2: (L_2 norm) Euclidean distance*

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- *$h \to \infty$. "supremum" (L_{max} norm, L_∞ norm, Chebyshev distance) distance.*

To compute it, we find the attribute *f* that gives the maximum difference in values between the two objects.

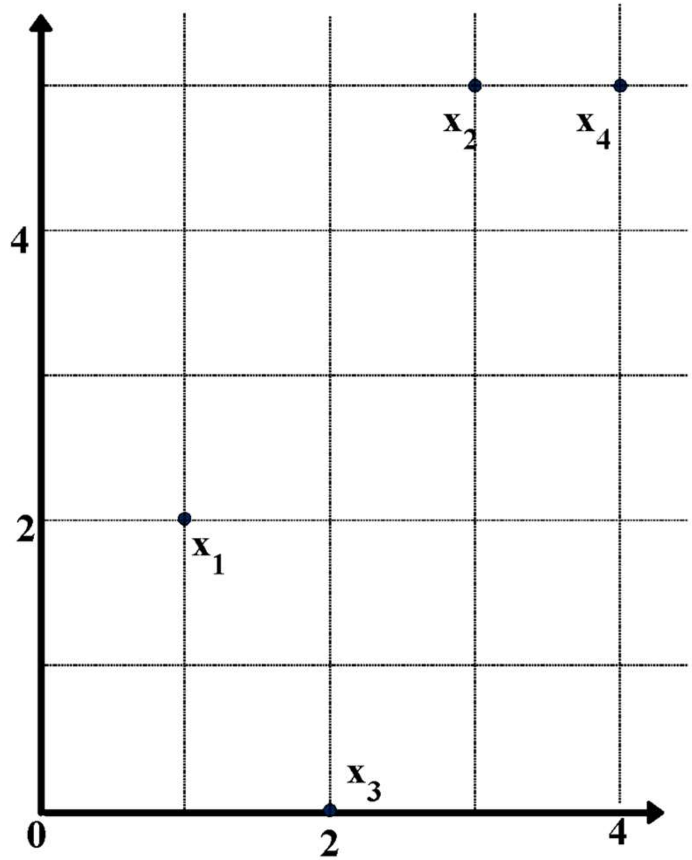
This difference is the supremum distance, defined more formally as:

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: **Supremum distance.** Let's use the two objects, $\boldsymbol{x}_1 = (1, 2)$ and $\boldsymbol{x}_2 = (3, 5)$ The second attribute gives the greatest difference between values for the objects, which is $5 - 2 = 3$. This is the supremum distance between both objects.

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

- Suppose that f is an attribute from a set of ordinal attributes describing n objects. The dissimilarity computation with respect to f involves the following 3 steps:

1. Let the value of f for the i-th object is xif , and f has Mf ordered states, representing the ranking: 1, 2, …. , Mf . Replace each xif by its corresponding rank, rif = {1,2 f1,.... , Mf }.
2. Map the range of each attribute onto [0.0, 1.0] so that each attribute has equal weight. Perform such data normalization by replacing the rank rif of the i-th object in the f th attribute by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

3. Dissimilarity can then be computed using any of the distance measures for numeric attributes, using zif to represent the f value for the i-th object.

Example: Dissimilarity for Ordinal Variables

Object Identifier	test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

Step 1: There are three states for test-2: fair, good, excellent, that is, Mf = 3.
Replace each value for test-2 by its rank, So objects 1 to 4 will be assigned ranks 3, 1, 2, and 3

Step 2: Normalize the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0.

Step 3: we can use, say, the Euclidean distance , which results in the following dissimilarity matrix:

$$
\begin{array}{c c}
 & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}
\end{array}
$$

- Objects 1 and 2 are the most dissimilar, as are objects 2 and 4 (i.e., d(2,1)=1.0 and d(4,2)= 1.0).

- sim(i, j)= 1 - d(i,,j).

Attributes of Mixed Type

Suppose that the data set contains p attributes of mixed type. The dissimilarity d(i, j) between objects i and j is defined as

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

where the indicator $\delta_{ij}^{(f)}=0$ if either (1) x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or object j), or (2) $x_{if}=x_{jf}=0$ and attribute f is asymmetric binary; otherwise, $\delta_{ij}^{(f)}=1$. The contribution of attribute f to the dissimilarity between i and j (i.e., $d_{ij}^{(f)}$) is computed dependent on its type:

- If f is numeric: $d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{max_h x_{hf}-min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.
- If f is nominal or binary: $d_{ij}^{(f)}=0$ if $x_{if}=x_{jf}$; otherwise, $d_{ij}^{(f)}=1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if}=\frac{r_{if}-1}{M_f-1}$, and treat z_{if} as numeric.

Example: Dissimilarity between attributes of mixed type

$d_{ij}^{(3)}$ $d_{ij}^{(3)}$

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	*test-1* (nominal)	*test-2* (ordinal)	*test-3* (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

We have computed Dissimilarity matrix for test-1 (which is nominal) and test-2 (which is ordinal),
Let us compute the dissimilarity matrix for the third attribute, test-3 (which is numeric).That is, we must compute $d_{ij}^{(3)}$

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.

Using the above case of numeric attributes, let maxhxh = 64 and minhxh = 22.

The difference between the two is used below to normalize the values of the dissimilarity matrix. The resulting dissimilarity matrix for test-3 is

$$\begin{array}{c|c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix} \end{array}$$

Dissimilarity matrix for test-3 attribute

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

We can now use the dissimilarity matrices for the three attributes in our computation of Eq. (2.22). The indicator $\delta_{ij}^{(f)} = 1$ for each of the three attributes, f. We get, for example, $d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65$. The resulting dissimilarity matrix obtained for the data described by the three attributes of mixed types is:

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & & & & \\ 2 & & & & \\ 3 & & & & \\ 4 & & & & \end{array}$$

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}.$$

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Test1 (Nominal)

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Test2 (Ordinal)

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Test3 (Numeric)

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If x and y are two vectors (e.g., term-frequency vectors), then

  cos(x, y) = (x · y) /||x|| ||y|| ,

  where · indicates vector dot product, ||x||: the Euclidean norm of vector x, defined as

$\boxed{\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}}$ **= length of vector x**

- The Cosine measure computes the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

- A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match.

- The closer the cosine value to 1, the smaller the angle and the greater the match between vectors.

Example: Cosine Similarity

$\cos(x, y) == (x \cdot y) / ||x|| \, ||y||$,

where · indicates vector dot product, ||d|: the length of vector d

Ex: Find the **similarity** between documents 1 and 2.

- x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
- y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

- $x \cdot y = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

- $||x|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5} = 6.481$

- $||y|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5} = 4.12$

- $\cos(x, y) = 0.94$ -----Quite similar closer to 1 → similar

Problem 1

i j

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance* between the two objects.

(b) Compute the *Manhattan distance* between the two objects.

(c) Compute the *Minkowski distance* between the two objects, using $h = 3$.

(d) Compute the *supremum distance* between the two objects.

Euclidean distance: The most popular distance measure

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

Minkowski distance: A generalization of Euclidean and Manhattan distances

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

Manhattan (or city block) distance: named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks).

The distance between two points measured along axes at right angles

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

- $h \to \infty$. *"supremum" (L_{max} norm, L_∞ norm, Chebyshev distance) distance.*

To compute it, we find the attribute f that gives the maximum difference in values between the two objects.

This difference is the supremum distance, defined more formally as:

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: **Supremum distance.** Let's use the two objects, $x_1 = (1, 2)$ and $x_2 = (3, 5)$. The second attribute gives the greatest difference between values for the objects, which is $5 - 2 = 3$. This is the supremum distance between both objects.

Solution-Problem 1

(a) Compute the *Euclidean distance* between the two objects.

The Euclidean distance is computed using Equation (2.6).

Therefore, we have $\sqrt{(22-20)^2+(1-0)^2+(42-36)^2+(10-8)^2} = \sqrt{45} = 6.7082$.

(b) Compute the *Manhattan distance* between the two objects.

The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22-20|+|1-0|+|42-36|+|10-8| = 11$.

(c) Compute the *Minkowski distance* between the two objects, using $h = 3$.

The Minkowski disance is

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdot+|x_{ip}-x_{jp}|^h} \tag{2.10}$$

where h is a real number such that $h \geq 1$.

Therefore, with $h = 3$, we have $\sqrt[3]{|22-20|^3+|1-0|^3+|42-36|^3+|10-8|^3} = \sqrt[3]{233} = 6.1534$.

(d) Compute the *supremum distance* between the two objects.

The supremum distance is computed using Equation (2.8). Therefore, we have a supremum distance of 6.

Problem 2

It is important to define or select similarity measures in data analysis. However, there is no commonly-accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

Suppose we have the following two-dimensional data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Consider the data as two-dimensional data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

Solution –Problem2(a)

	Euclidean dist.	Manhattan dist.	supremum dist.	cosine sim.
$\boldsymbol{x_1}$	0.1414	0.2	0.1	0.99999
$\boldsymbol{x_2}$	0.6708	0.9	0.6	0.99575
$\boldsymbol{x_3}$	0.2828	0.4	0.2	0.99997
$\boldsymbol{x_4}$	0.2236	0.3	0.2	0.99903
$\boldsymbol{x_5}$	0.6083	0.7	0.6	0.96536

These values produce the following rankings of the data points based on similarity:

Euclidean distance: x_1, x_4, x_3, x_5, x_2

Manhattan distance: x_1, x_4, x_3, x_5, x_2

Supremum distance: x_1, x_4, x_3, x_5, x_2

Cosine similarity: x_1, x_3, x_4, x_2, x_5

Solution-Problem 2(b)

(b) The normalized query is (0.65850, 0.75258). The normalized data set is given by the following table

	A_1	A_2
$\boldsymbol{x_1}$	0.66162	0.74984
$\boldsymbol{x_2}$	0.72500	0.68875
$\boldsymbol{x_3}$	0.66436	0.74741
$\boldsymbol{x_4}$	0.62470	0.78087
$\boldsymbol{x_5}$	0.83205	0.55470

Recomputing the Euclidean distances as before yields

	Euclidean dist.
$\boldsymbol{x_1}$	0.00415
$\boldsymbol{x_2}$	0.09217
$\boldsymbol{x_3}$	0.00781
$\boldsymbol{x_4}$	0.04409
$\boldsymbol{x_5}$	0.26320

which results in the final ranking of the transformed data points: x_1, x_3, x_4, x_2, x_5