Data Mining Issues

Major issues in data mining

1. **Mining methodology and user interaction issues**

- Mining different kinds of knowledge in databases:

Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery task. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

- Interactive mining of knowledge at multiple levels of abstraction:

Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

- Incorporation of background knowledge:

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction

- Pattern evaluation—the interestingness problem:

A data mining system can uncover thousands of patterns. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

2. **Performance issues**

- Efficiency and scalability of data mining algorithms:

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. The running time of a data mining algorithm must be predictable and acceptable in large databases.

- Parallel, distributed, and incremental mining algorithms:

The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such

algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

3. **Issues relating to the diversity of database types:**

- Handling of relational and complex types of data:

Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. Specific data mining systems should be constructed for mining specific kinds of data.

- Mining information from heterogeneous databases and global information systems:

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining.

****

data set, you may construct a decision tree to predict the missing values for income.

❖ **Noisy Data**

"What is noise?" Noise is a random error or variance in a measured variable. Let's look at the following data smoothing techniques.

1. **Binning:** Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. Figure 3.2 illustrates some binning techniques. In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values).

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

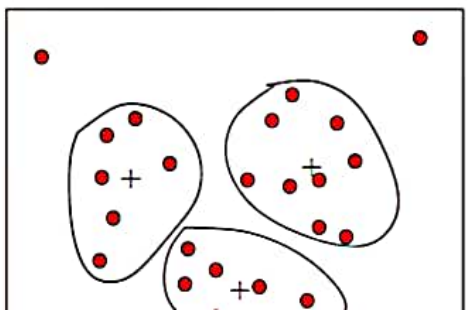
Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Binning methods for data smoothing.

   a. In **smoothing by bin means**, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9.
   b. Similarly, **smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.

   c. **In smoothing by bin boundaries**, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal width, where the interval range of values in each bin is constant.

2. **Clustering :** Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.



given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal width, where the interval range of values in each bin is constant.

2. **Clustering :** Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

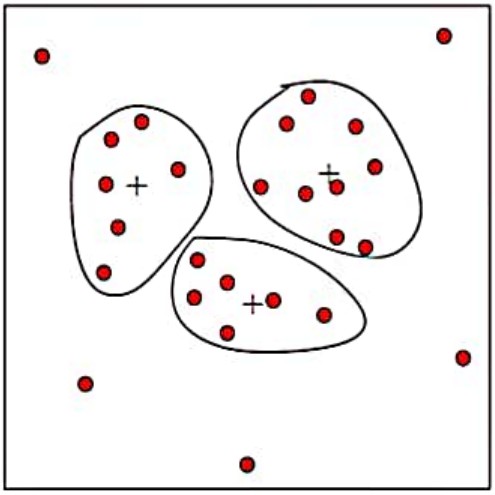


Figure 3.3: Outliers may be detected by clustering analysis.

3. **Combined computer and human inspection:** Outliers may be identified through a combination of **computer and human inspection**. In one application, for example, an information-theoretic measure was used to help identify outlier patterns in a handwritten character database for classification. The measure's value reflected the **"surprise"** content of the predicted character label with respect to the known label. Outlier patterns may be informative (e.g., identifying useful data exceptions, such as different versions of the characters "0"or "7"), or **"garbage"**

(e.g., mislabeled characters). Patterns whose surprise content is above a threshold are output to a list. A human can then sort through the patterns in the list to identify the actual garbage ones.

This is much faster than having **to manually search through the entire database.** The garbage patterns can then be removed from the (training) database. The garbage patterns can be excluded from use in subsequent data mining.

4. **Regression:** Data smoothing can also be done by regression, a technique that conforms data values to a function. **Linear regression** involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. **Multiple linear regression** is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Measuring Performance Of Classifier Dog

→ Confusion Matrix

		Actual class	
		Yes	No
Predicted class	Yes	TP	FP
	No	FN	TN

- It is a model that is used to describe the performance of a classifier mod
- It is a table that categorice predictions according to whether they match the actual values
- Tp :
- Fp : Falsily positive
-
-

→ pricision
→ recall
→ accurency

? $\text{Pricision} = \frac{Tp}{Tp+Fp}$

$\text{Recall} = \frac{Tp}{Tp+FN}$

$\text{Accurency} = \frac{Tp+TN}{TP+Fp+EN+TN}$

? Suppose a Computer prgm to recognizing dogs in a photogram identifies 8 dogs in a picture contain 12 dogs and some cats. of the 8 dogs identified, 5 actually are dogs. while the rest are cats. Compute the pricision and recall of the Computer prgm

	Actual: Dog	Actual: Cat	
Dog	5	3	8
Cat	7		
	12		

$\text{Pricision} = \frac{Tp}{Tp+Fp} = \frac{5}{5+3} = 5/8$

$\text{Recall} = \frac{Tp}{Tp+FN} = \frac{5}{5+7} = 5/12$

? Let there be 10 balls (6 White & 4 red) in a box and it is required to pick up red balls from them. Suppose if he pick up 7 balls as red of which 2 are actualy red ball What are the value of precision and recall in picking red ball

	red	White	
red	2	5	7 predicted red
White	2	1	total = 10
	4 actual red	6	

$$\text{Precision} = \frac{Tp}{Tp + Fp} = \frac{2}{7}$$

$$\text{Recall} = \frac{Tp}{Tp + FN} = 2/4$$

? a database contains 80 record on a particular topic whose 55 are relevent of certain investigations. a search was conducted on that topic and 50 records are retrieved. 40 were relevent Comp

	relevent	not relevent	
relevent	40	10	50
not relevent	15	15	total = 80
	55		

$$\text{Precision} = \frac{Tp}{Tp + Fp} \quad \text{Re}$$

$$= \frac{40}{50}$$

$$\text{Recall} = \frac{Tp}{Tp + FN}$$

$$= \frac{40}{55}$$

21 Good if

? Consider a 2 class classification problem of predicting whether a photograph contain Man and Women. Suppose we have a test dataset of 10 records with expected outcome and a set of prediction from our classification algm.

Expected	Predicted
Man	Woman
Man	Man
Woman	Woman
Man	Man
Woman	Man
Woman	Woman
Woman	Woman
Man	Man
Man	Woman
Woman	Woman

Complete the Confusion Matrix

Compute accuracy precision recall

Predicted \ Actual	Man	Woman	
Man	3	1	4
Woman	2	4	6
	5	5	

Precision : $\frac{TP}{TP+FP} = \frac{3}{3+1} = 3/4$

Recall : $\frac{TP}{TP+FN} = \frac{3}{3+2} = 3/5$

Accuracy : $\frac{TP+TN}{TP+FP+FN+TN} = \frac{3+4}{3+1+2+4} = 7/10$

1) Suppose 10000 patterns gets tested for flue, out of 9000 are actually health and 1000 sick. for the sick was +ve for 620 and -ve for 380 for health people the same test was for 180 and -ve for 8820. Construct a confusion matrix for the data and compute the accuracy, precision, recall

Predicted \ Actual	Sick	Health	
Sick	620	180	800
Health	380	8820	
	1000	9000	

Precision : $\frac{Tp}{Tp+Fp} = \frac{620}{620+180} = \frac{620}{800}$

Recall : $\frac{Tp}{Tp+FN} = \frac{620}{620+380} = \frac{620}{1000}$

Accuracy : $\frac{Tp+TN}{Tp+FN+Ep+TN} = \frac{620+8820}{620+180+380+8820} = \frac{9440}{10000}$

Suppose that the data for analysis includes the attribute age. The

attributes that can be useful for knowledge discovery.

Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, **mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.**

Data reduction strategies include the following

1. **Data Cube Aggregation**, where **aggregation operations are applied to the construction of a data cube.**
2. **Dimension Reduction**, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.
3. **Data compression,** where encoding mechanisms are used to reduce the data set size.
4. **Numerosity reduction**, where the the original data volume is replaced by alternative, smaller forms of data representation. These techniques may be parametric or nonparametric. For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.) Regression and log-linear models are examples. Nonparametric methods for storing reduced representations of the data include histograms, clustering, sampling.
5. **Discretization and concept hierarchy generation**, where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction and are a powerful tool for data mining.

➲ Data Cube Aggregation

Imagine that you have collected the data for your analysis. These data consist of the AllElectronics sales per quarter, for the years 2008 to 2010. You are, however,

Name - Anjana Sreekumar
Roll No - 17

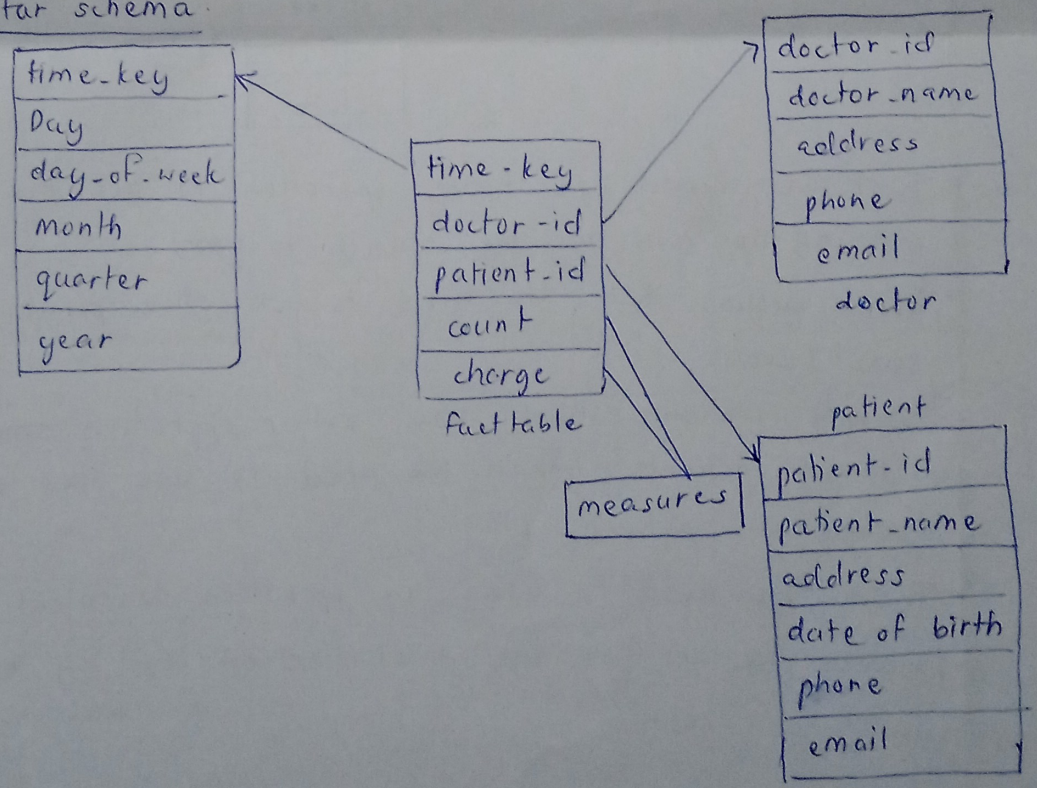
DM Tutorial - 1

(1) Suppose that a datawarehouse consist of the three dimensions time, doctor and patient and the two measures count and charge, where charge is the fee that a doctor charges a patient for visit.

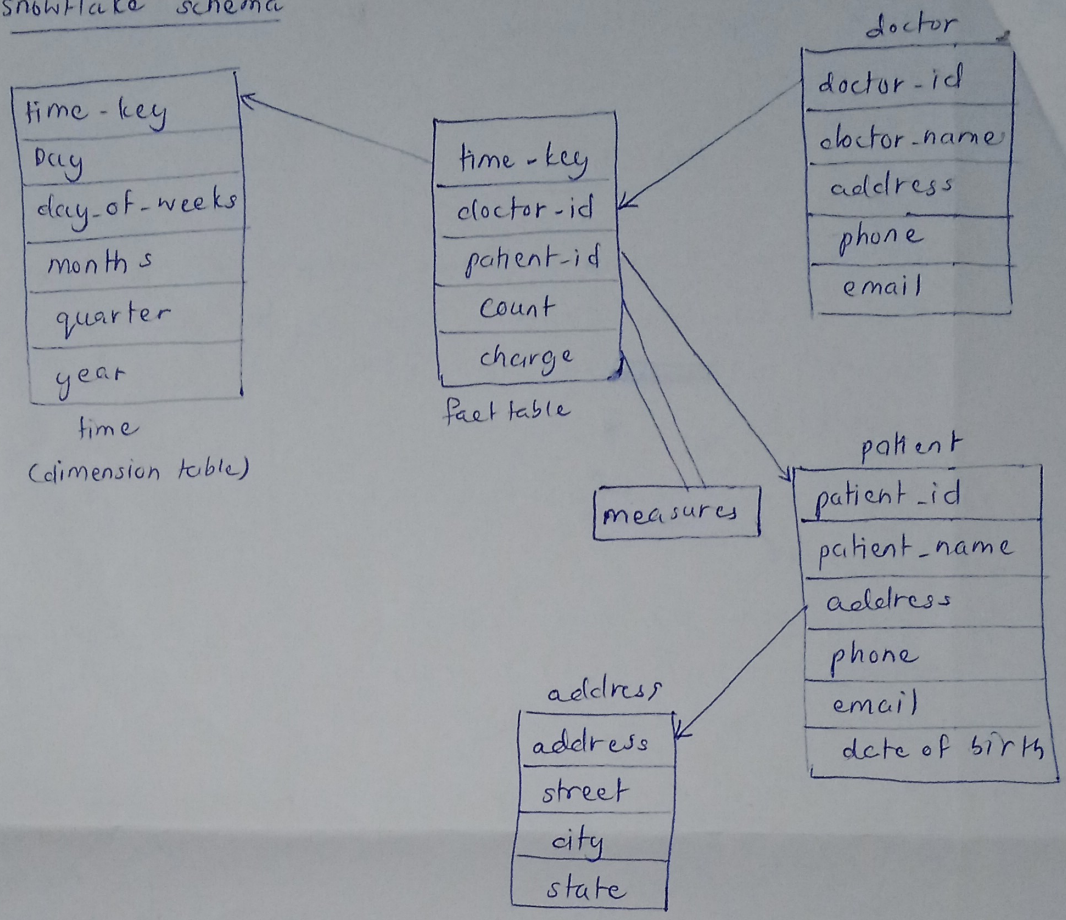
a) Draw a star and snowflake schema diagram of for the data warehouse.

b) Starting with the base cuboid, what specific OLAP operation should be performed in order to list the total fee collected by each doctor in 2009 ?

a → Star schema.



Snowflake schema



b → First we should use Rollup operation to get the year 2004 (rolling-up from day then month to year).

- After getting that, we need to use slice operation to select (2004).
- Second, we ~~use~~ should use rollup operation again to get all patients. Then we need to use slice operation to get all patients.
- Then, we need to use slice operation. to select (all).
- Finally, we get list the total fee collected by each doctor in 2004.

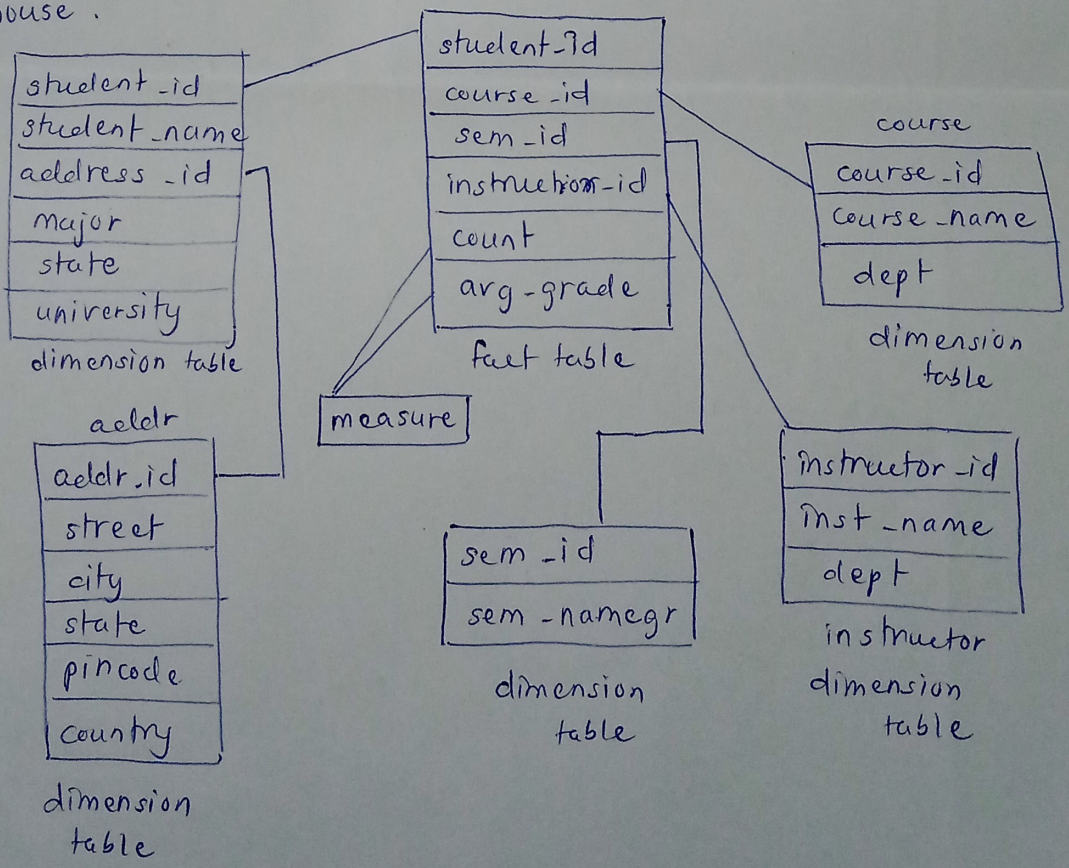
So,

1. rollup from day to month to year.

2. slice for year = "2004"
3. rollup on patient from individual patient to all
4. slice for patient = "all"
5. get the list of total fee collected by each doctor in 2004.

(2) Suppose the datawarehouse for Big-university consist of the following 4 dimension : Student, course, semester and instructor and 2 measures - count and average when at the lowest conceptual level, the avg grade measures store the actual course grade of the student. At higher conceptual level, avg grade store the avg grade for the given combination.

a) Draw a snow flake schema diagram for the data wore house.



b) Standing with the base cuboid [student, course, sem, instructor], what specific OLAP operations should one perform in order to test the average grade of cs course for each Big university student.

i. Roll up on course from course_id to dept.

ii. Roll up on student from student_id to university.

iii. Dice on course, student with dept = "cs" and university = "Big university"

iv. Drill-down on student from university to student_name

Accuracy : $\frac{Tp+TN}{Tp+FN+Fp+TN} = \frac{620+8820}{620+180+380+8820} = \frac{9440}{10000}$

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

a) What is the mean of the data? What is the median?

b) Use smoothing by bin means to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data

c) How might you determine outlier in the data?

d) What other method are there for data smoothing

e) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].

f) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 yr

g)

a)

Mean $= \frac{809}{27} = 29.96$

Median $= 25$

b) Bin 1 : 13, 15, 16

Bin 2 : 16, 19, 20

Bin 3 : 20, 21, 22

Bin 4 : 22, 25, 25

Bin 5 : 25, 25, 30

Bin 6 : 33, 33, 35

Bin 7 : 35, 35, 35

Bin 8 : 36, 40, 45

Bin 9 : 46, 52, 70

After Smoothing :-

Bin 1 : 14.66, 14.66, 14.66

Bin 2 : 18.33, 18.33, 18.33

Bin 3 : 21, 21, 21

Bin 4 : 24, 24, 24

Bin 5 : 26.6, 26.6, 26.6

Bin 6 : 33.6, 33.6, 33.6

Bin 7 : 35, 35, 35

Bin 8 : 40.3, 40.3, 40.3

Bin 9 : 56, 56, 56

e) $v = 35$

$min_D = 13$ new_min = 0

$max_D = 70$ new_max = 1

$$v' = \frac{v - min_A}{max_D - min_D} (new\ max_D - new\ min_D) + new\_min_D$$

$$= \frac{35 - 13}{70 - 13} (1 - 0) + 0$$

$$= \frac{22}{57}$$

$$= 0.385$$

c) The data that is deviated more from mean is called outliers outliers get eliminated by smoothing values that are outside the given range of data

d)

f) $v = 35$

$\sigma = 12.94$

$$v' = \frac{v - \bar{D}}{\sigma_D}$$

$$= \frac{35 - 29.96}{12.94}$$

$$= 0.389$$

g) Decimal Scale

$$v' = \frac{v}{10^2} = \frac{35}{100} = 0.35$$

b) Z-Score normalization

2. Suppose a hospital tested the age and body fat data for 18 men.