Data

- A data set can often be viewed as a collection of data objects.
- Other names for a data object include record, point, vector, pattern, event, case, sample, observation or entity.

Data

- Data objects are described by a number of attributes that capture the basic characteristics of an object.
- Other names for an attribute are variable, characteristic, field, feature, or dimension.
- A data set is usually a file, in which
  - The objects are records (or rows) in the file and
  - Each field (or column) corresponds to an attribute.

Data

Student ID	Year	Grade Point Average (GPA)
51034262	1	3.24
51052663	2	3.51
51082246	3	3.62

Attributes

- An attribute is a property or characteristic of an object that may vary, either from one object to another or from one time to another.
- A measurement scale is a rule (function) that associates a numerical or symbolic value with an attribute of an object.
- The process of measurement is the application of a measurement scale to associate a value with a particular attribute of a specific object.

Different types of attributes

- We can define four types of attributes
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Nominal and ordinal attributes are collectively referred to as categorical or qualitative attributes.
- Interval and ratio attributes are collectively referred to as quantitative or numeric attributes.

Different types of attributes

- Nominal
  - The values of a nominal attribute are just different names.
  - They provide only enough information to distinguish one object from another.
  - Examples: eye color, gender.

Different types of attributes

- Ordinal
  - The values of an ordinal attribute provide enough information to order objects.
  - Example: grades.
- Interval
  - For interval attributes, the differences between values are meaningful.
  - Example: calendar dates.
- Ratio
  - For ratio variables, both differences and ratios are meaningful.
  - Example: monetary quantities, mass, length.

Different types of attributes

- Another way to distinguish between attributes is by the number of values they can take.
- Based on this criterion, attributes can be classified as either discrete or continuous.

Different types of attributes

- Discrete
  - A discrete attribute has a finite or countably infinite set of values.
  - Such attributes can be categorical, such as gender, or numeric, such as counts.
  - Binary attributes are a special case of discrete attributes and assume only two values, e.g. true/false, yes/no, male/female, or 0/1.

Different types of attributes

- Continuous
  - A continuous attribute is one whose values are real numbers
  - Examples include temperature, height or weight.
  - Continuous attributes are typically represented as floating point variables.

Types of data sets

- We consider the following different types of data sets
  - Record data
  - Transaction or market basket data
  - Data matrix
  - Sparse data matrix

Types of data sets

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	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

(d) Document-term matrix.

Record data

- A data set is usually represented as a collection of records.
- Each record consists of a fixed set of data fields (attributes).
- Record data is usually stored either in flat files or in relational databases.

Transaction or market basket data

- Transaction data is a special type of record data.
- Each transaction involves a set of items.
- Example: the set of products purchased by a customer during one shopping trip constitutes a transaction.

Data matrix

- If the data objects all have the same fixed set of numeric attributes, then they can be thought of as points (vectors) in a multi-dimensional space.
- This kind of data set can be interpreted as an m by n matrix where
  - There are m rows, one for each object.
  - There are n columns, one for each attribute.
- Standard matrix operations can be applied to transform and manipulate the data.

Sparse data matrix

- A sparse data matrix is a special case of a data matrix in which there are a large number of zeros in the matrix, and only the non-zero attribute values are important.
- Sparsity is an advantage because usually only the non-zero values need to be stored and manipulated.
- This results in significant savings with respect to computation time and storage.

Sparse data matrix

- An example is document data.
- A document can be represented as a term vector, where
  - Each term is a component (attribute) of the vector and
  - The value of each component is the number of times the corresponding term occurs in the document.
- This representation of a collection of documents is often called a document-term matrix.

Data quality

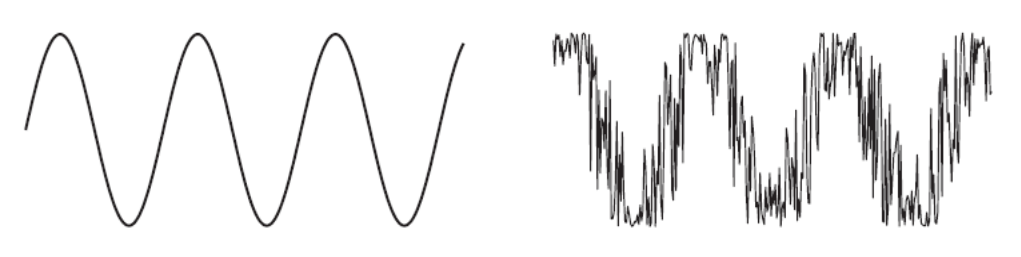
- Precision
  - The closeness of repeated measurements (of the same quantity) to one another.
  - This is often measured by the standard deviation of a set of values.
- Bias
  - A systematic variation of measurements from the quantity being measured.
  - This is measured by taking the difference between
    - the mean of the set of values and
    - the known value of the quantity being measured.

Data quality

- Suppose we have a standard laboratory weight with a mass of 1g.
- We want to assess the precision and bias of our new laboratory scale.
- We weigh the mass five times, and obtain the values: {1.015, 0.990, 1.013, 1.001, 0.986}.
- The mean of these values is 1.001.
- The bias is thus 0.001.
- The precision, as measured by the standard deviation, is 0.013.

Data quality: Noise and outliers

- Noise
  - Noise is the random component of a measurement error.



- Outliers
  - Data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set.
  - Values of an attribute that are unusual with respect to the typical values of that attribute.

Data quality: Missing values

- It is not unusual for an object to be missing one or more attribute values.
- There are several strategies for dealing with missing data
  - Eliminate data objects
  - Estimate missing values

Data quality: Missing values

- Eliminate data objects
  - If a data set has only a few objects that have missing attribute values, then it may be convenient to omit them.
  - However, even a partially specified data object contains some information.
  - If many objects have missing values, then a reliable analysis can be difficult or impossible.

Data quality: Missing values

- Estimate missing values
  - A missing attribute value of a point can be estimated by the corresponding attribute values of the other points.
  - If the attribute is discrete, then the most commonly occurring attribute value can be used.
  - If the attribute is continuous, then the average attribute value of similar points is used.

Data preprocessing

- There are a number of techniques for performing data preprocessing
  - Aggregation
  - Sampling
  - Dimensionality reduction
  - Discretization
  - Normalization

Aggregation

- Aggregation is the combining of two or more objects into a single object.
- There are several motivations for aggregation
  - The smaller data sets resulting from aggregation require less memory and processing time.
  - Aggregation can also provide a high-level view of the data.
  - Aggregate quantities, such as averages or totals, have less variability than the individual objects.
- A disadvantage of aggregation is the potential loss of interesting details.

Sampling

- Sampling is the selection of a subset of the data objects to be analyzed.
- Sometimes, it is too expensive or time consuming to process all the data.
- Using a sampling algorithm can reduce the data size to a point where a better, but more computationally expensive algorithm can be used.
- A sample is representative if it has approximately the same property (of interest) as the original set of data.

Sampling

- The simplest type of sampling is uniform random sampling.
- For this type of sampling, there is an equal probability of selecting any particular item.
- There are two variations on random sampling
  - Sampling without replacement
  - Sampling with replacement

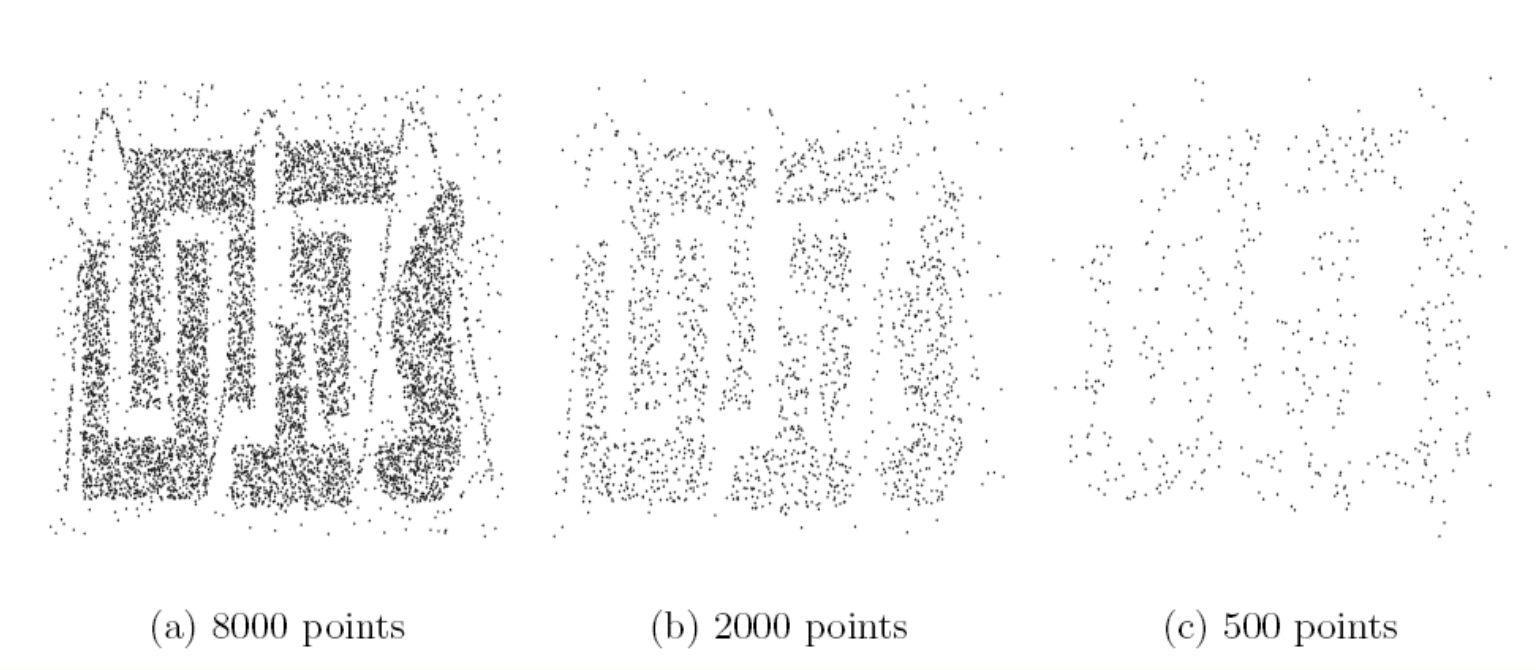
Sampling

- Sampling without replacement
  - As each item is selected, it is removed from the set of all objects.
- Sampling with replacement
  - Objects are not removed from the data set as they are selected.
  - The same object can be picked more than once.

Sampling

- Once a sampling technique has been selected, it is still necessary to choose the sample size.
- For larger sample sizes
  - The probability that a sample will be representative will be increased.
  - However, much of the advantage of sampling will also be eliminated.
- For smaller sample sizes
  - There may be a loss of important information.

Sampling



(a) 8000 points (b) 2000 points (c) 500 points

Dimensionality reduction

- The dimensionality of a data set is the number of attributes that each object possesses.
- It is usually more difficult to analyze high-dimensional data (curse of dimensionality).
- An important preprocessing step is dimensionality reduction.

Dimensionality reduction

- Dimensionality reduction has a number of advantages:
  - It can eliminate irrelevant features and reduce noise.
  - It can lead to a more understandable model which involves fewer attributes.
  - It may allow the data to be more easily visualized.
  - The amount of time and memory required for processing the data is reduced.

Dimensionality reduction

- The curse of dimensionality refers to the phenomenon that many types of data analysis become significantly harder as the number of dimensions increases.
- As the number of dimensions increases, the data becomes increasingly sparse in the space that it occupies.
- There may not be enough data objects to allow the reliable creation of a model that describes the set of objects.

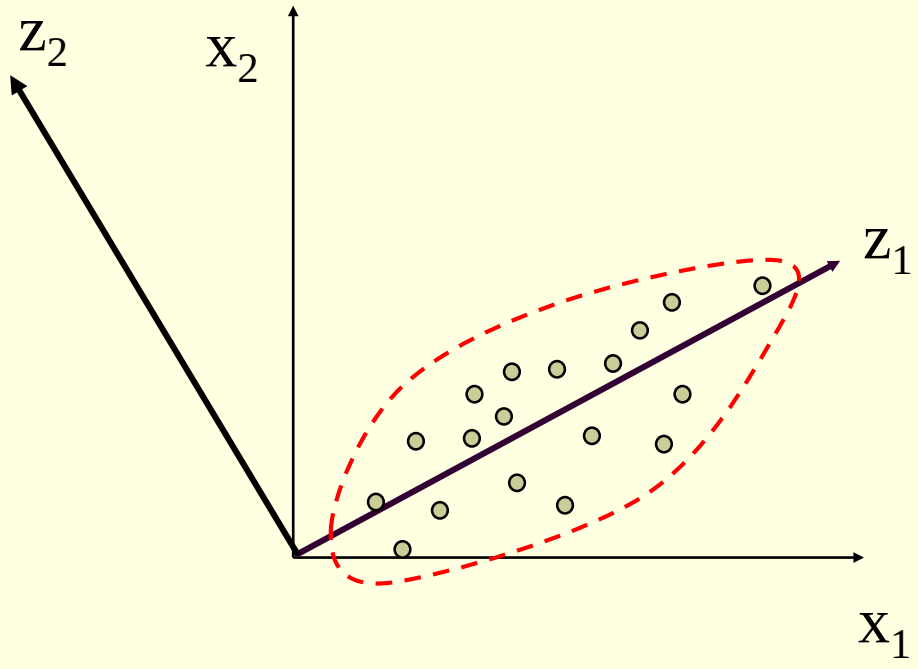
Dimensionality reduction

- There are a number of techniques for dimensionality reduction
  - Feature transformation
  - Feature subset selection

Feature transformation

- Feature transformation can be used to project data from a high-dimensional space to a low-dimensional space.
- Principal Component Analysis (PCA) is a feature transformation technique to find new attributes that are
  - linear combinations of the original attributes.
  - capture the maximum amount of variation in the data.

Feature transformation



Feature subset selection

- Another way to reduce the number of dimensions is to use only a subset of the features.
- This approach is effective if redundant and irrelevant features are present.
- Redundant features duplicate much or all of the information contained in one or more other attributes.
- Irrelevant features contain almost no useful information for the task at hand.

Feature subset selection

- The ideal approach to feature selection is to
  - Try all possible subsets of features.
  - Take the subset that produces the best result.
- Since the number of subsets involving n attributes is 2^n, such an approach is impractical in most situations.
- There are three standard approaches to feature selection
  - Embedded approaches
  - Filter approaches
  - Wrapper approaches

Embedded approaches

- Feature selection occurs naturally as part of the algorithm.
- The algorithm itself decides which attributes to use and which to ignore.

Filter approaches

- Features are selected before the algorithm is run.
- An evaluation measure is used to determine the goodness of a subset of attributes.
- This measure is independent of the current algorithm used.

Wrapper approaches

- These methods use the target algorithm as a black box to find the best subset of attributes.
- Typically, not all the possible subsets are considered.

Discretization

- In some cases, we prefer to use data with discrete attributes.
- It is thus necessary to transform a continuous attribute into a discrete attribute.

Discretization

- Transformation of a continuous attribute to a discrete attribute involves two subtasks
  - Deciding how many possible discrete values to have.
  - Determining how to map the values of the continuous attribute to these discrete values.

Discretization

- In the first step
  - The values of the continuous attribute are first sorted.
  - They are then divided into S intervals by specifying S-1 split points.
- In the second step
  - All the values in one interval are mapped to the same discrete value.

Normalization

- The goal of normalization or standardization is to make an entire set of values have a particular property.
- Normalization is necessary to avoid the case where a variable with large values dominates the result of the calculation.

Normalization

- Example 1: We suppose
  - $\bar{x}$ is the mean of the attribute values
  - σ_x is their standard deviation
  - We can use the following transformation to create a new variable that has a mean of 0 and a standard deviation of 1.

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

Normalization

- Example 2: We suppose
  - $x_{\min}$ is the minimum attribute value
  - $x_{\max}$ is the maximum attribute value
  - We can use the following transformation to create a new variable that has a minimum value of 0 and a maximum value of 1.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Similarity and dissimilarity

- The similarity between two objects is a numerical measure of the degree to which the two objects are alike.
- Similarities are higher for pairs of objects that are more alike.

Similarity and dissimilarity

- The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different.
- Dissimilarities are lower for more similar pairs of objects.
- Frequently, the term distance is used as a synonym for dissimilarity.
- The term proximity is used to refer to either similarity or dissimilarity.

Dissimilarity between attribute values

- We consider the definition of dissimilarity measures for the following attribute types
  - Nominal
  - Ordinal
  - Interval/Ratio

Nominal

- Nominal attributes only convey information about the distinctness of objects.
- All we can say is that two objects either have the same attribute value or not.
- As a result, dissimilarity is defined as
  - 0 if the attribute values match
  - 1 otherwise

Ordinal

- For ordinal attributes, information about order should be taken into account.
- The values of the ordinal attribute are often mapped to successive integers.
- The dissimilarity can be defined by taking the absolute difference between these integers.

Interval/Ratio

- For interval or ratio attributes, the natural measure of dissimilarity between two objects is the absolute difference of their values.

Distance

- The Euclidean distance d between two points **x** and **y** is given by

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{u=1}^{n}(x_u - y_u)^2}$$

where

- n is the number of dimensions
- x_u and y_u are, respectively, the u-th attributes of **x** and **y**.

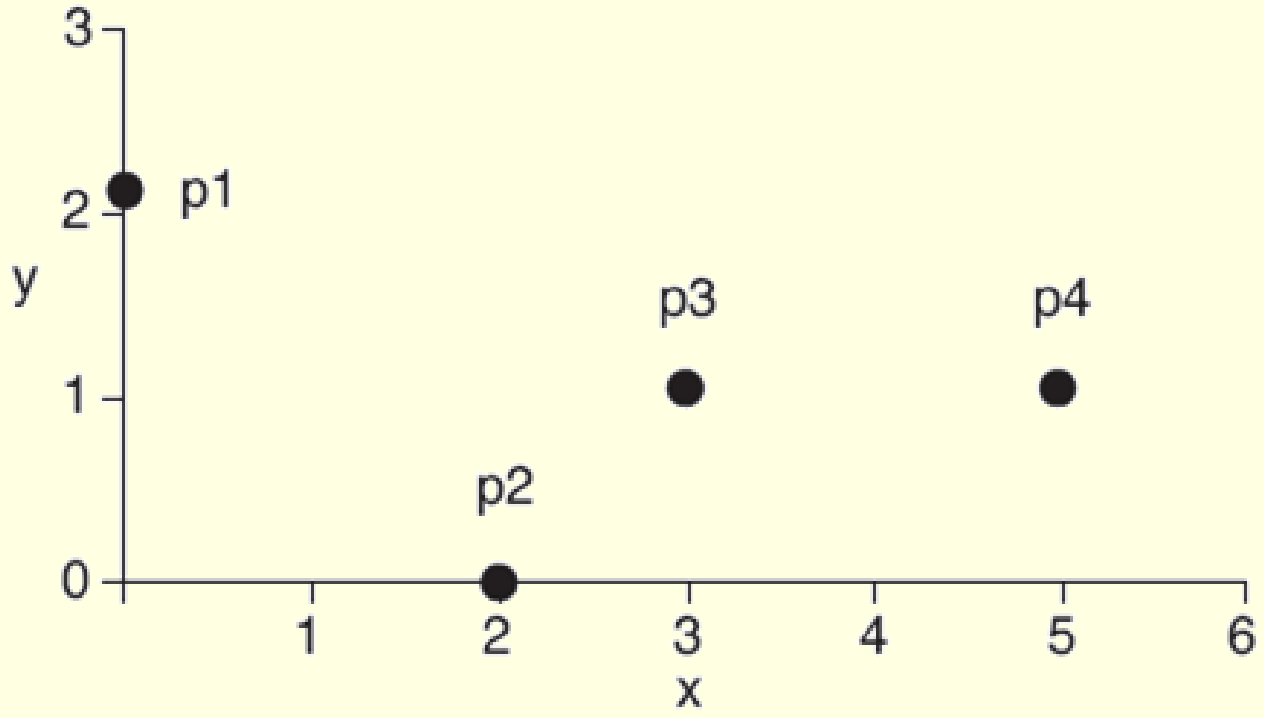
Distance

- The Euclidean distance measure is generalized by the Minkowski distance metric as follows:

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{u=1}^{n} \left| x_u - y_u \right|^h\right)^{\frac{1}{h}}$$

- Three most common examples of Minkowski distances are
  - h=1: City block distance (L_1 norm)
  - h=2: Euclidean distance (L_2 norm)
  - h=∞: Supremum distance (L_{max} or L_∞ norm), which is the maximum difference between any attribute of the objects.

Distance



Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

x and y coordinates of four points

	p1	p2	p3	P4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Euclidean distance matrix

Distance

	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

L_1 distance matrix

	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

L_∞ distance matrix

Distance

- A distance measure has some well-known properties
  - Positivity
    - $d(\mathbf{x},\mathbf{y}) \geq 0$ for all **x** and **y**
    - $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$
  - Symmetry
    - $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ for all **x** and **y**.
  - Triangle inequality
    - $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ for all points **x**, **y** and **z**.

Distance

- In the previous discussion, all attributes were treated equally when computing the distance.
- This is not desirable when some attributes are more important than others.
- To address these situations, the distance measure can be modified by weighting the contribution of each attribute:

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{u=1}^{n} w_u \left|x_u - y_u\right|^h\right)^{\frac{1}{h}}$$

Summary statistics

- Summary statistics are quantities that capture various characteristics of a large set of values using a small set of numbers.
- We consider the following summary statistics
  - Relative frequency and the mode
  - Measure of location: mean and median
  - Measure of spread: range and variance

Relative frequency and the mode

- Suppose we are given a discrete attribute x, which can take values $\{a_1,\ldots,a_s,\ldots,a_S\}$, and a set of m objects.
- The relative frequency of a value a_s is defined as

$$\text{relative frequency}(a_s) = \frac{\text{number of objects with attribute value } a_s}{m}$$

- The mode of a discrete attribute is the value that has the highest relative frequency.

Mean

- We consider a set of m objects and an attribute x.
- Let $\{x_1,\ldots\ldots,x_m\}$ be the attribute values of x for these m objects.
- The mean is defined as follows:

$$\text{mean}(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Median

- Let $\{x_{(1)}, \ldots\ldots x_{(m)}\}$ represent the values of x after they have been sorted in non-decreasing order.
- Thus, $x_{(1)}=x_{min}$ and $x_{(m)}=x_{max}$.
- The median is defined as follows:

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if m is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if m is even, i.e., } m = 2r \end{cases}$$

Mean and median

- The mean is sensitive to the presence of outliers.
- The median provides a more robust numerical summary of a set of values.
- To overcome problems with the mean, the notion of a trimmed mean is sometimes used.
  - A percentage p between 0 and 100 is specified.
  - The top and bottom (p/2)% of the data is thrown out.
  - The mean is then calculated in the normal way.

Range

- The simplest measure of spread is the range.
- Given an attribute x with a set of m values $\{x_1, \ldots\ldots x_m\}$, the range is defined as

$$\text{range}(x) = x_{\max} - x_{\min} = x_{(m)} - x_{(1)}$$

- However, using the range to measure the spread can be misleading if
  - most of the values are concentrated in a narrow band of values
  - there are also a relatively small number of more extreme values.

Variance

- The variance σ_x^2 of the values of an attribute x is defined as follows:

$$\text{variance}(x) = \sigma_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

- The standard deviation, which is the square root of the variance, is denoted as σ_x

Multivariate summary statistics

- The mean or median of a data set that consists of several attributes (multivariate data) can be obtained by computing the mean or median separately for each attribute.
- Given a data set, the mean of the data objects is given by:

$$\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots\ldots, \bar{x}_n)$$

Multivariate summary statistics

- For multivariate data, the spread of the data is most commonly captured by the covariance matrix **C**.
- The uv-th entry c_{uv} is the covariance of the u-th and v-th attributes of the data.

$$c_{uv} = \text{covariance}(x_u, x_v)$$

- This covariance is given by

$$\text{covariance}(x_u, x_v) = \frac{1}{m-1}\sum_{i=1}^{m}(x_{iu} - \bar{x}_u)(x_{iv} - \bar{x}_v)$$

Multivariate summary statistics

- The covariance of two attributes is a measure of the degree to which two attributes vary together.
- However, this measure depends on the magnitudes of the variables.
- In view of this, we perform the following operation on the covariance to obtain the correlation coefficient r_{uv}.

$$r_{uv} = \frac{\text{covariance}(x_u, x_v)}{\sigma_u \sigma_v}$$

where σ_u and σ_v are the standard deviations of x_u and x_v respectively.

- The range of r_{uv} is from -1 to 1.

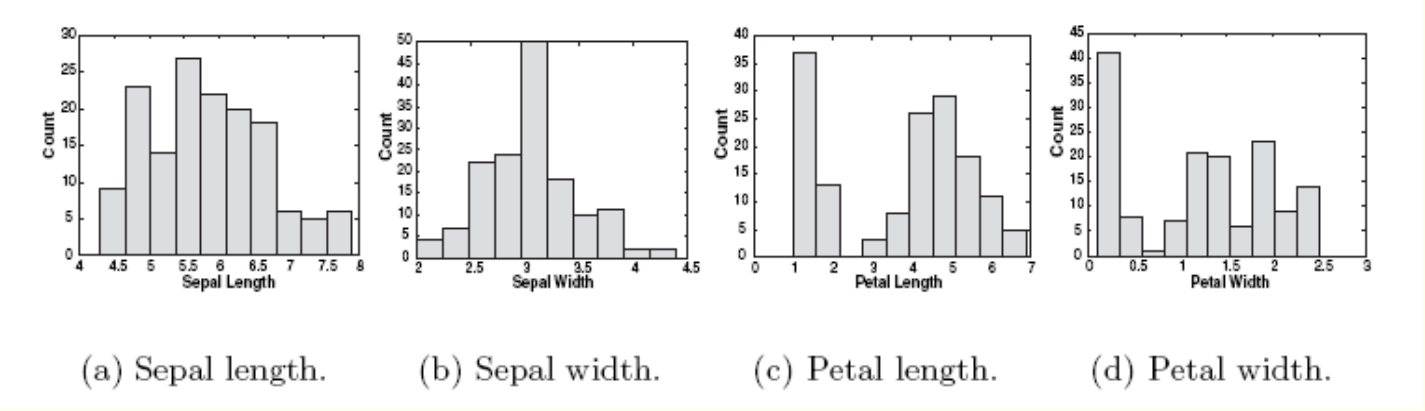
Data visualization

- The motivation of using data visualization is that people can quickly absorb large amounts of visual information and find patterns in it.
- We consider the following data visualization techniques
  - Histogram
  - Scatter plot

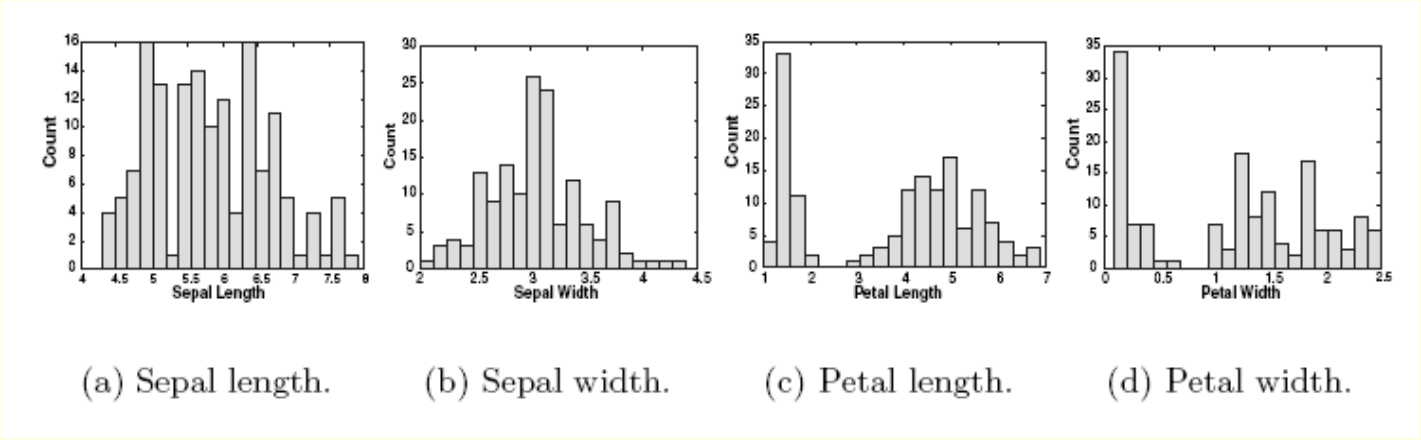
Histogram

- A histogram is a plot that displays the distribution of attribute values by
  - dividing the possible values into bins and
  - showing the number of objects that fall into each bin.
- Each bin is represented by one bar.
- The area of each bar is proportional to the number of values that fall into the corresponding range.

Histogram



Histograms of four Iris attributes (10 bins)



Histograms of four Iris attributes (20 bins)

Scatter plot

- A scatter plot can graphically show the relationship between two attributes.
- In particular, it can be used to judge the degree of linear correlation of the attributes.

Scatter plot