slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

DATA PRELIMINARIES

CSCI 452: Data Mining

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

id	Refund	Marital Status	Annual Income	Defaulted Barrower
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

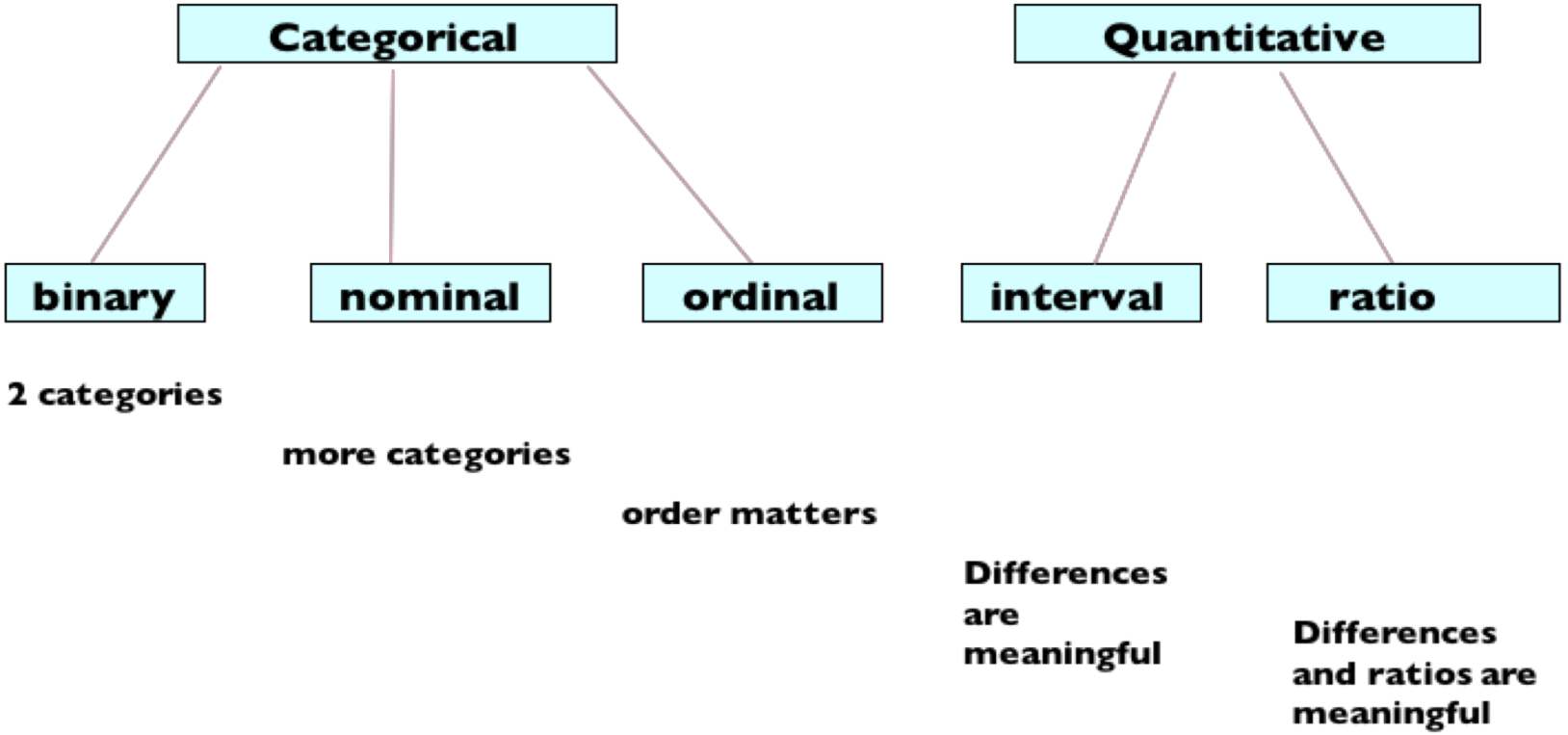
Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- 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
      - ID has no limit but age has a maximum and minimum value

Different Types of Attributes

- Binary
- Nominal
  - *Examples:* ID numbers, eye color, zip codes
- Ordinal
  - *Examples:* rankings, size in {small, medium, large}
- Interval
  - *Examples:* calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - *Examples:* time, monetary quantities, length

Attribute Types



Properties of Attribute Values

- The type of an attribute depends on which of the properties it possesses

Nominal	Eye color	= ≠
Ordinal	Size {small, medium, large}	= ≠ < > ≤ ≥
Interval	Calendar dates	= ≠ < > ≤ ≥ + −
Ratio	Time	= ≠ < > ≤ ≥ + − × ÷

Ordinal vs. Interval vs. Ratio

- Ordinal
  - Order matters, but not the difference between values
  - Difference between 7 and 5 may not be same as difference between 5 and 3
- Interval
  - Difference between two values is meaningful
  - 100 degrees – 90 degrees is same difference as 90 degrees – 80 degrees
  - Temperature of 100 degrees is not twice as hot as 50 degrees
- Ratio
  - Clear definition of 0.0; none of a variable at 0.0
  - Weight of 8 grams is twice the weight of 4 grams
    - (Temperature 100 Kelvin is twice as hot as 50 Kelvin; Kelvin is Ratio; 0.0 Kelvin means “no heat”)

Discrete and Continuous Attributes

Discrete

- Has finite attribute values
- Often represented as integer variables
- Examples: zip codes, counts, {1,2,3,…}
- (Note: binary 0/1 attributes are special case of discrete.)

Continuous

- Has real numbers as attribute values
- Often represented as doubles (floating-pt variables)
- Examples: height, temperature, 3.14159
- (Practically, real values can only be measured and represented using a finite number of digits)

Characteristics of Data Sets

- Dimensionality: number of attributes that objects possess
- Sparsity: most attributes of objects have values of 0
- Resolution: granularity of measurements

Dimensionality

- Univariate: Measurement made on one variable per subject
- Bivariate: Measurement made on two variables per subject
- Multivariate: Measurement made on more than two variables per subject

Types of Datasets

- Record Data
- Graph-Based Data
- Ordered Data

Graph-Based Data

- Data with Relationships among Objects: web pages, social media data
- Data with Objects That Are Graphs: structure of chemical compounds

Ordered Data

- Temporal Data
  - Each record has a time associated with it
    - Example: customer transactions
- Sequence Data
  - Dataset has sequence of individual entities (such as sequence of words or letters)
    - *Example:* DNA sequence (ATGC possible letters

Ordered Data (cont'd)

- Time Series
  - Series of measurements taken over time
    - *Example:* financial stock price data
  - Temporal autocorrelation: if two measurements are close in time, then the value of those measurements are often very similar
- Spatial Data
  - Each record has a position or area
    - *Example:* geographical locations
  - Spatial autocorrelation: objects that are physically close tend to be similar

Data Quality

- Unrealistic to expect that data will be perfect
- Some data mining algorithms are more susceptible to data quality issues
- Want to avoid "garbage in garbage out"
- **Data cleaning** phase for detection and correction of data issues often necessary during preprocessing

Measurement and Data Collection Errors

- Measurement error: any problem resulting from the measurement process; value recorded differs from true value to some extent
- Data collection error:
  - data objects are omitted
  - attribute values are missing for some objects
  - inappropriately including a data object

Noise

- Noise: the random component of a *measurement error*
  - Elimination of noise is very difficult or impossible for some measurements
  - Data mining techniques try to be robust enough to still produce acceptable results even when noise is present
  - We will see how noise is modeled in the underlying statistics and machine learning

Precision, Bias, Accuracy

- Assume we make repeated measurements
  - *Example:* weighing mass of object; .01 or .001 difference between measurements
- Precision: the closeness of repeated measurements to one another
  - Often measured by standard deviation
- Bias: a systematic variation of measurements
  - Often measured as difference of measurement average compared to true value
  - Similar to “accuracy”

Precision, Bias, Accuracy

- Measurements: {1.015, 0.990, 1.013, 1.001, 0.986}
- True mass: 1.000 g
- Mean of measurements: 1.001
- Bias: 0.001
  - 1.001 – 1.000 = 0.001
- Precision: 0.013
  - Standard deviation

Precision, Bias, Accuracy

- Often, data sets do not come with information on the precision of the data.
  - Remains to be discovered by the data analyst

Outliers

1. Data objects that have characteristics from most other data objects
   - In fraud detection, the goal is identifying these outliers
2. Value of an attribute is very unusual with respect to the typical value
   - Do we have a "data error?" or is some individual really eight foot tall?

- Various statistical definitions for what an outlier is.
- Outliers can be legitimate data objects or values (and may be of interest).
  - Outliers different from noise

Missing Values

- Often, values for some attributes are missing for some objects in data sets
  - *Example:* individuals who decline to provide their weight in a survey
- What to do?

Strategies for Dealing with Missing Data

- Eliminate data objects that have missing values
- Eliminate data attributes if any objects are missing that value
- Estimate missing values
  - Data set may contain similar data points
- Ignore missing values
  - If data mining method is robust

Inconsistent Values

- *Example:*
  - Data object with address, city, zip code in three separate fields
  - But address / city is in a different zip code

- Some inconsistencies are easy to detect (and fix) automatically; others are not.

Duplicate Data

- *Example:*
  - many people receive duplicate mailings because they are in a database multiple times under slightly different names

Other Issues

- Timeliness
  - Data starts to age as soon as it has been collected
  - *Example:* general population of users interact with Facebook differently than they did so 2 years ago
- Relevance
  - Sampling bias: occurs when a sample is not representative of the overall population
  - *Example:* survey data describes only those who responded to the survey

Other Issues

- The data sets needs to contain attributes which are relevant for the overall problem
  - *Example:* Constructing an accurate model that predicts the accident rate for drivers might be fruitless without features such as:
    - age, previous accident history, # of speeding tickets, etc.

Knowledge about the Data

- Ideally data sets are accompanied by *documentation* that describes different aspects of the data
  - Read it!
  - *Example:* contains information that missing values for a particular field are coded as -9999
  - Should also document the type of feature (nominal, etc.) and its measurement scale (meters or feet, etc.)

Exploring Data

- Data Exploration: a preliminary investigation of the data in order to better understand its specific characteristics
  - Aid in selection appropriate preprocessing and data analysis techniques
  - Patterns can sometime be found simply by *visualizing* the data (and then can be used to explain the data mining results)
  - Summary statistics also used

Summary Statistics

- Capture various characteristics of a large set of values
- Common summary statistics:
  - Mean
  - Standard deviation
  - Range
  - Mode
- Most summary statistics can be calculated in a single pass through the data.

Frequency

$$frequency(v_i) = \frac{\text{number of objects with attribute } v_i}{n}$$

Class	Size	Frequency
Freshman	200	0.33
Sophomore	160	0.27
Junior	130	0.22
Senior	110	0.18

Often used with *categorical* values.

Mode

The mode (especially with discrete / continuous data) may reveal value that symbolizes a missing value.

The value that has the highest frequency.

Class	Size	Frequency
Freshman	200	0.33
Sophomore	160	0.27
Junior	130	0.22
Senior	110	0.18

Often used with *categorical* values.

Percentiles

- For ordered data, percentile is useful.
- Given an ordinal or continuous attribute x and a number p between 0 and 100, the p^{th} percentile x_p is a value of x such that $p\%$ of the observed values are less than x_p.
- *Example:* the 75^{th} percentile is the value such that 75% of all values are less than it.

Mean

- Measure the "location" of a set of values
- Mean is a very, very common measurement
  - But is sensitive to outliers

$$mean(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{given } \{x_1, \ldots, x_n\}$$

Median

- Commonly used instead of mean if outliers are present
- Median is the middle value if odd number of values are present; average of the two middle values if even number of values
- Ordered set of *n* values: $\{x_1, \ldots, x_n\}$

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

Mean vs. Median

- If the distribution of values is skewed, then the median is a better indicator of the middle, compare to the mean.

Trimmed Mean

- Specify a percentage p between 0 and 100
- Top and bottom $(p/2)\%$ of data is not used in mean calculation
  - $p=0$, corresponds to standard mean
  - $p=100$, corresponds to median calculation

Range

- "Measure of spread"
- Ordered set of *n* values: $\{x_1, \ldots, x_n\}$

$$range(x) = \max(x) - \min(x) = x_n - x_1$$

- Can be misleading if most values are concentrated, but a few values are extreme

Variance / Standard Deviation

- "Measure of spread"
- Ordered set of *n* values: $\{x_1, \ldots, x_n\}$

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

$$sd(x) = s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Because variance and standard deviation measures use the mean, they can also be sensitive to outliers.

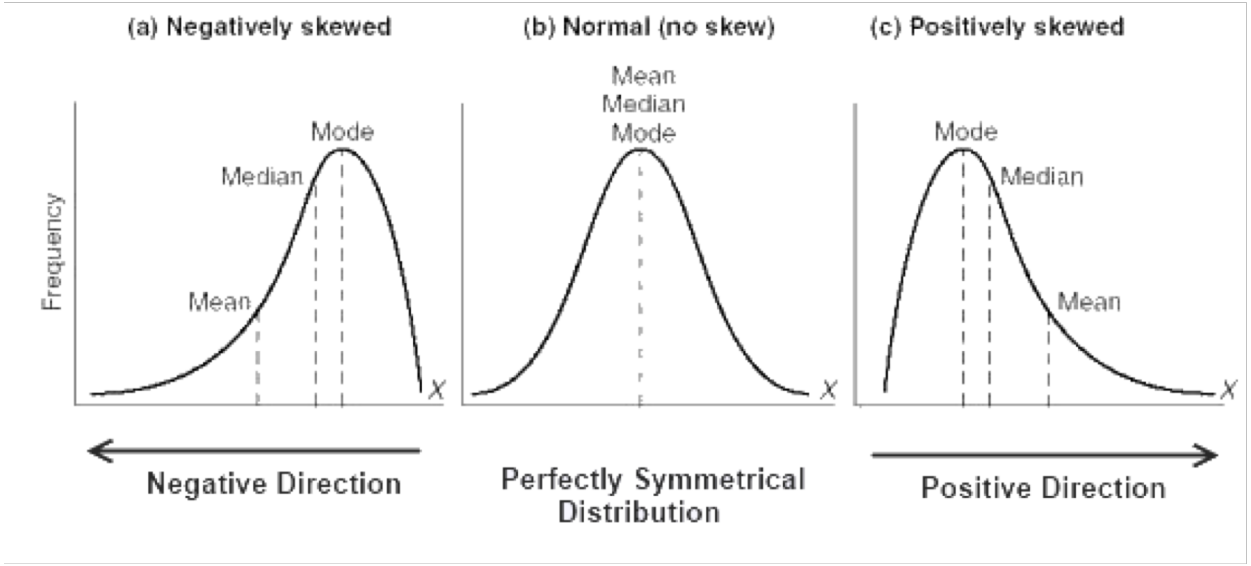
Other Measures of Spread

- Absolute Average Deviation
- Median Absolute Deviation
- Interquartile Range

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Skewness

- Measures the degree to which the values are symmetrically distributed about the center



Visualization

- The display of information in a graphic or tabular format.
- Many visualization formats exist (contour plots, graphs, heat maps) to display high-dimensional information.
- Since this isn't a visualization course, we'll mainly use "traditional" two-dimensional graphic types.
- How can we transform a dataset with many attributes into two dimensions?
  - Selection: Typically by selecting two dimensions at a time
  - Can also only "select" a subset of records to display
  - Other techniques also exist

Iris Data Set

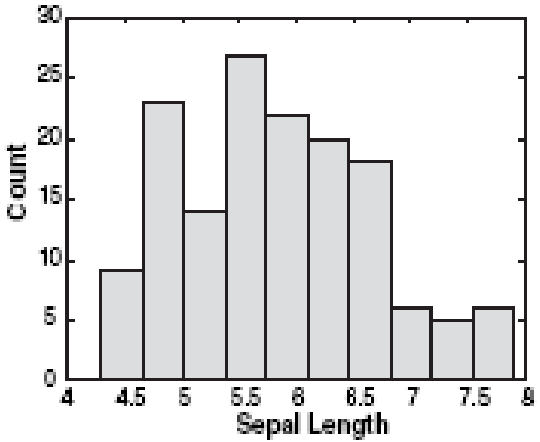
- Next few slides will demonstrate visualization using the classic Iris dataset
  - Freely available from UCI (University of California at Irvine) Machine Learning Lab
  - Relatively very small
  - 150 records of Iris flowers (50 for each species)
  - Attributes:
    - Sepal length (centimeters)
    - Sepal width (centimeters)
    - Petal length (centimeters)
    - Petal width (centimeters)
    - Class (species of Iris) {Setosa, Versicolour, Virginica}

Histogram

- For showing the distribution of values
- Divide values into bins; show number of objects that fall into each bin
- Shape of histogram depends on number of bins

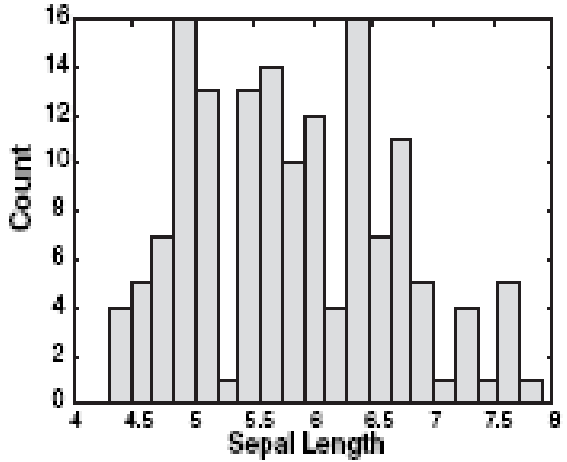
43 44 44 44 45 46 46 46 46 47 47 48 48 48 48 48 49 49 49 49 49 49 50
50 50 50 50 50 50 50 50 50 51 51 51 51 51 51 51 51 51 52 52 52 52 53
54 54 54 54 54 54 55 55 55 55 55 55 55 56 56 56 56 56 56 57 57 57 57
57 57 57 57 58 58 58 58 58 58 58 59 59 59 60 60 60 60 60 60 61 61 61
61 61 61 62 62 62 62 63 63 63 63 63 63 63 63 63 64 64 64 64 64 64 64
65 65 65 65 65 66 66 67 67 67 67 67 67 67 67 68 68 68 69 69 69 69 70
71 72 72 72 73 74 76 77 77 77 77 79

Sepal length data



Histogram (10 bins)

Bins of **equal width**



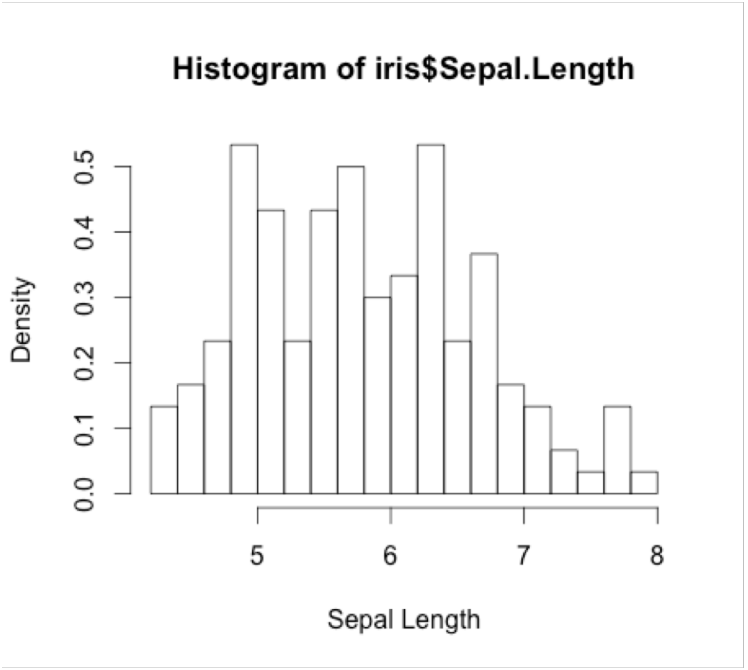
Histogram (20 bins)

Histogram

- Previous slide showed histogram of continuous attribute
- For categorical attribute, each category is a bin.
  - If there are too many bins, then values need to be combined in some way.

Relative Frequency Histogram

- Instead of counts on the y-axis, the relative frequency (density) is used.

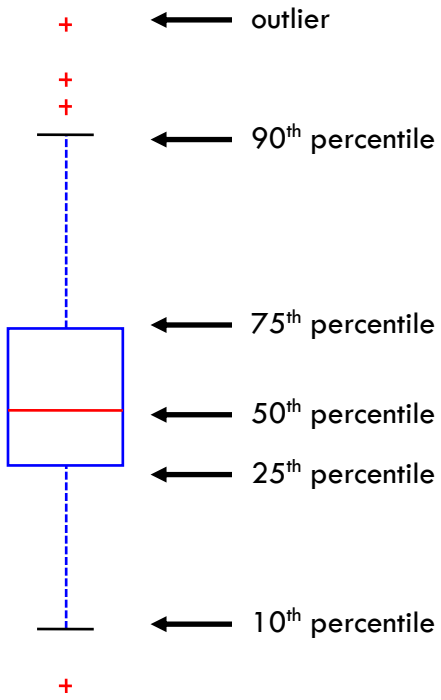



```
> hist(iris$Sepal.Length,freq = F,xlab="Sepal Length",breaks=20)
```

Box Plot

Note: only 50% of the data is in the box!

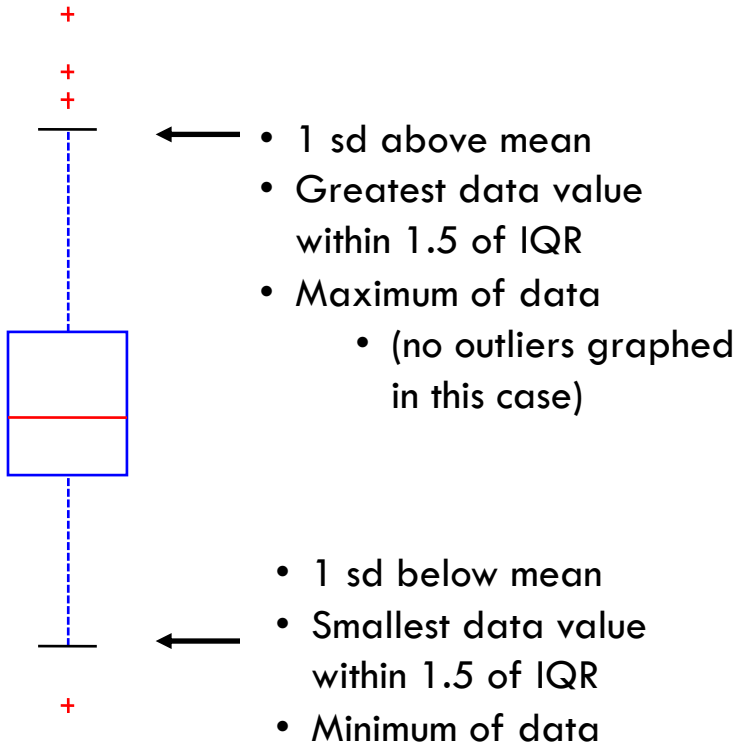
- Box plots show the distribution of the values for a single *numerical* attribute.
- Whiskers: top and bottom lines of the box plot



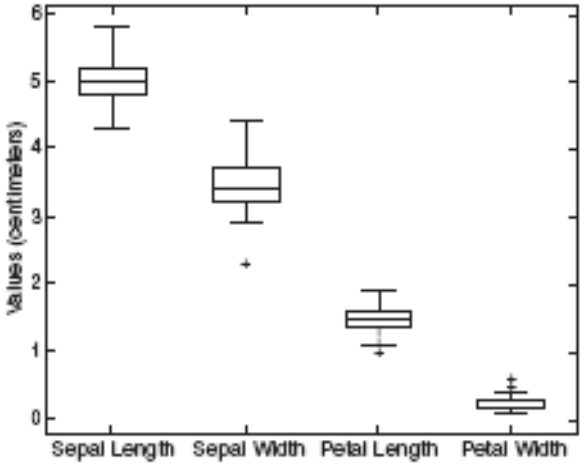
Box Plot

Note: only 50% of the data is in the box!

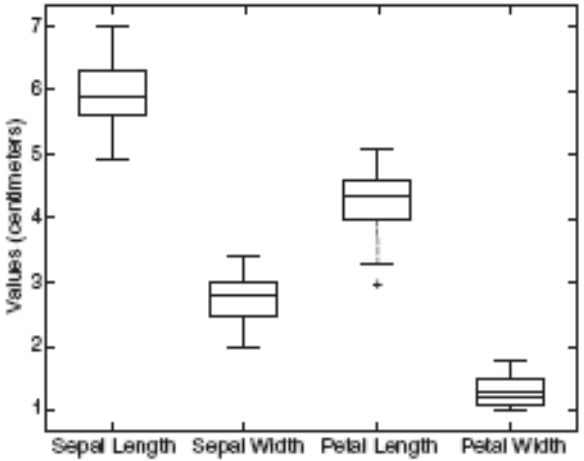
- Whiskers can represent several possible alternate values
- Best to describe the convention used in a legend along the chart



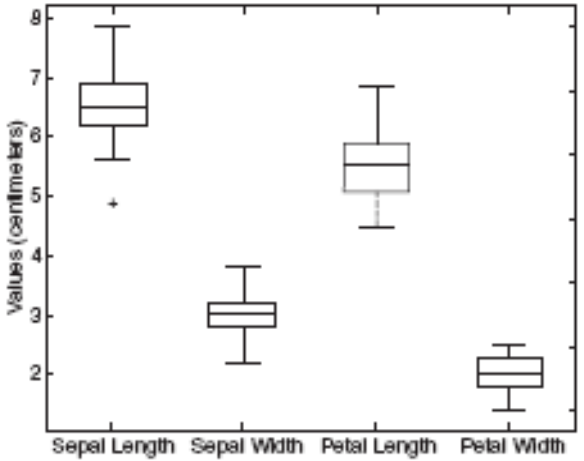
Box plots are relatively compact; many of them can be shown on the same plot.



(a) Setosa.



(b) Versicolour.

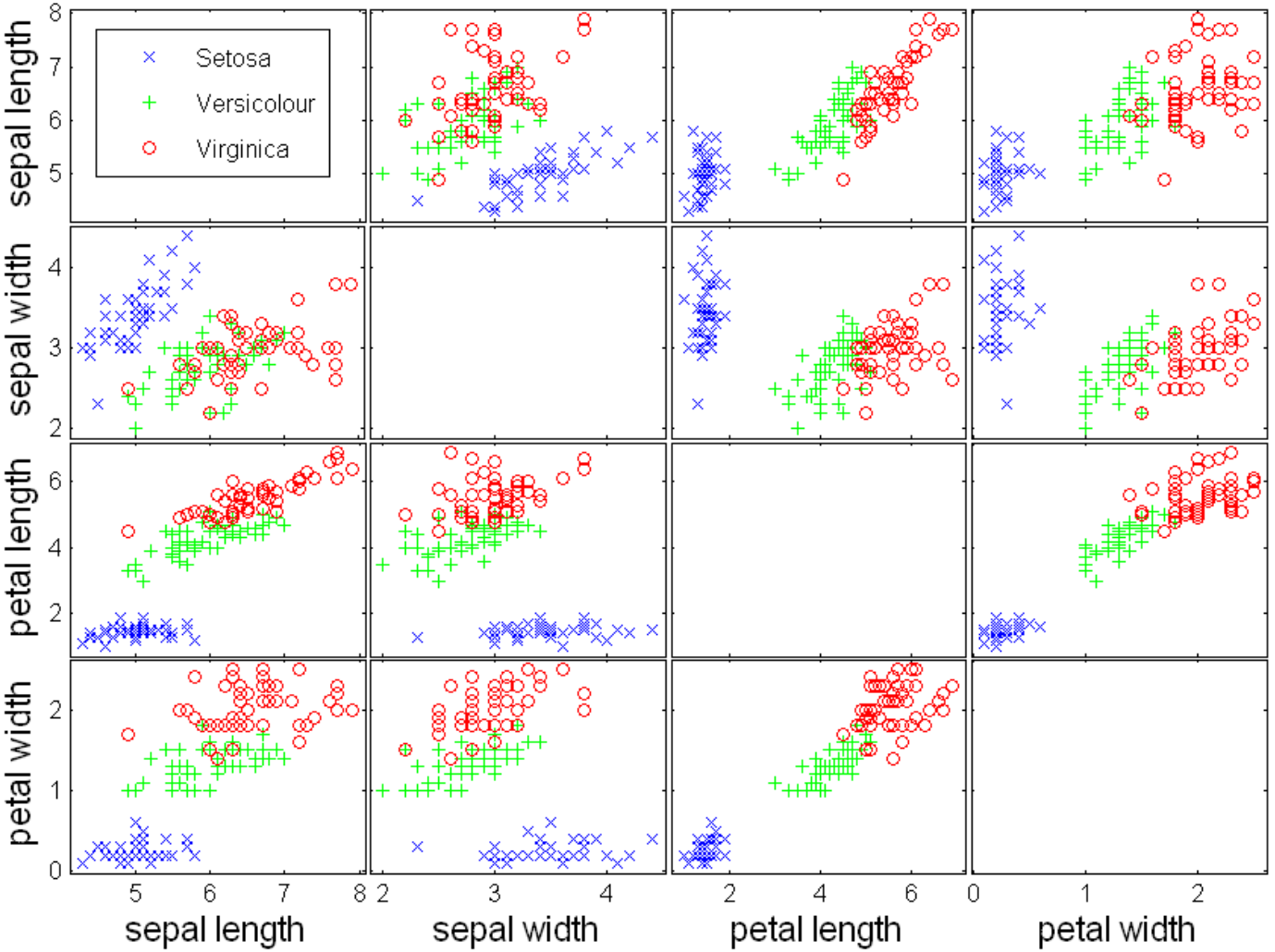


(c) Virginica.

Scatter Plot

- Data objects are plotted as a point in a 2d-plane: one attribute on x-axis, the other on y-axis
  - Assumed that both attributes are discrete or continuous
- Scatter plot matrix: organized way to examine a number of scatter plots simultaneously
  - Scatter plots for multiple pairs of attributes

When class labels are available, a scatter plot matrix can visualize how much two attributes separate the classes.



References

- *Introduction to Data Mining*, 1st edition, Tan et al.
- http://en.wikipedia.org/wiki/Box_plot