Lecture 2 – Preprocessing and Visualization

Data Mining, Spring 2016

Anders Hartzen, andershh@itu.dk

IT UNIVERSITY OF COPENHAGEN

Some slides adapted from Hector Martinez of University of Malta (former ITU).

Hello!

- Anders Hartzen
  - 2010: B.Sc. Software Development, ITU
  - 2012: M.Sc. Games Technology, ITU
  - Working at ITU since as
    - Research and Teaching assistant
    - This semester also External Lecturer for Data Mining
  - Trying to get PhD project funded
  - Email: andershh@itu.dk

Overview of today's lecture 1/2

- Knowing your data
  - Attributes and attribute types
- Measuring your data
  - Central tendency measures
  - Data dispersion measures
  - Data similarity/dissimilarity/distance
- Visualizing your data

Overview of today's lecture 2/2

- Cleaning your data
- Reducing data
- Transforming data
- Conclusion: Why preprocessing and data visualization?

Some slides adapted from Hector Martinez.

Knowing your data

Data Object

- A data set is made up of **data objects**, also known as
  - samples, examples, instances, data points, data tuple and tuple
- Data objects describe the entities the data set has data on
  - e.g. a customer in a customer-database
- Each row in a database is a data object

Name	Address	Position
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Attribute

- An **attribute** is a **data field**, which describes a characteristic of a data object
  - E.g. name, address or position
- Also known as
  - Dimension, feature and variable
- Many types of attributes

Name	**Address**	**Position**
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Attribute Types

- *Qualitative*
  - Nominal
  - Binary
  - Ordinal
- *Quantitative*
  - Numeric
    - Interval-Scaled
    - Ratio-Scaled

- Not necessarily mutually exclusive
- Other types
  - Discrete
  - Continuous
  - String
  - etc.
  - May vary from tool to tool

Nominal Attribute

- Nominal
  - *"of, relating to, or constituting a name"* Merriam-Webster Dictionary
- Symbol or name of *things*
  - e.g. code or category
- No meaningful order between possible values for the nominal attribute
- Also known as **categorical** or **enumeration**

Name	**Address**	**Position**
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Nominal Attribute

- Can be encoded used integers
  - e.g. Student = 1; Professor = 2 etc.
- When encoded as integers, can we then use nominal attributes quantitatively?
  - e.g. subtract one from another?
  - Or calculate the average?

Name	Address	Position
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Nominal Attribute

- Can be encoded used integers
  - e.g. Student = 1; Professor = 2 etc.
- When encoded as integers, can we then use nominal attributes quantitatively?
  - e.g. subtract one from another?
    - 2 -1 aka Professor - Student
  - Or calculate the average?
- Answer: NO!

Name	**Address**	**Position**
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Nominal Attribute

- Nominal attributes should never be used quantitatively

Name	Address	Position
John Doe	Happy Road 2	Student
Jane Doe	Spring Way 1	Student
Joan Petersen	Sunset Blvd.	Professor

Binary Attribute

- Is a nominal attribute, but with the restriction that it can only have two possible values:
  - 0 = Usually means that the attribute is absent or “turned off”
  - 1 = Usually means that the attribute is present or “turned on”
- Also known as boolean when 1 and 0 correspond to *true* and *false*

Name	Married	Flu Positive
John Doe	0	1
Jane Doe	1	0
Joan Petersen	0	1

Binary Attribute

- Symmetric binary attribute
  - Both possible values are equally valuable and has same weight
  - e.g. Married
- Asymmetric binary attribute
  - Both possible values not equally important, i.e. one outcome is better than the other
  - Convention: Most important outcome = 1
  - e.g. Flu Positive

Name	**Married**	**Flu Positive**
John Doe	0	1
Jane Doe	1	0
Joan Petersen	0	1

Ordinal Attribute

- Similar to nominal attribute, but where possible values have an order or ranking between them
  - Example: Drink Size
- Magnitude i.e. distance between possible values not known
  - We do not know how much bigger a large drink is compared to a medium one

Drink Name	Drink Size	Price
Juice	medium	1.99
Juice	large	2.99
Slush	small	0.99

Interval-Scaled Attribute

- Numerical attribute whose values are measured on an equal-size scale
- Possible values have order (e.g. -1 is before 2) and can be negative, zero and positive
- Differences in values can be compared and quantified

Date	Forecast	Temperature (celsius)
10/12/2015	Sunny	2
11/12/2015	Cloudy	5
12/12/2015	Snow	-3

Interval-Scaled Attribute

- However, when comparing different values we can not say a value is a *multiple* or *ratio* of another value
  - e.g. A is two times larger than B
- Example: Temparture (celsius)
  - Celsius scale has no zero-point (i.e. 0 celsius not equal to “no temparture”)
- Other example: Dates
  - Year 0 (Gregorian calendar) not the beginning of time

Date	**Forecast**	**Temperature (celsius)**
10/12/2015	Sunny	2
11/12/2015	Cloudy	5
12/12/2015	Snow	-3

Ratio-Scaled Attribute

- Numeric attribute that has a zero-point
- Therefore we can say when comparing different values that a value is a *multiple* or *ratio* of another value
- e.g. A is two times larger than B
- Example: Kelvin temperature scale
  - Kelvin = 0 means zero kinectic energy for atomic particles

Date	**Forecast**	**Temperature (Kelvin)**
10/12/2015	Sunny	276
11/12/2015	Cloudy	281
12/12/2015	Snow	270

Discrete vs Continuous Attributes

Discrete Attribute

- Has a finite set of possible values
  - Values may or may not be represented as integers
- Examples: Position, Drink Size, Flu Positive, Married, Forecast, Age
- *Countable* infinite: Attribute that theoretically can have infinite values, but doesn't have in practice e.g. zip-codes

Continuous Attribute

- Is the opposite of discrete, i.e. has infinite set of possible values
- Usually represented by floating-point value
- Example: Measurements like height, width, weight, distance etc.

Measuring your data

Measuring your data

- Central tendency
  - Where do most values for an attribute fall?
- Central tendency measures
  - Mean
  - Median
  - Mode
- Data dispersion
  - How are the data spread out?
- Data dispersion measures
  - Variance and Standard Deviation
  - Range
  - Quantiles
  - Five-Number Summary

Central Tendency - Mean

- Most common and effective measure of the “center” of data set
- Formula - Let X_1, X_2, …., X_N be a set of N values for a Numeric attribute, then the mean is:

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

Central Tendency - Mean

- Example – Let us compute the mean Age
- Mean Age = (24+24+28+75+80) / 5 = 46.2
- Any problems here?

Name	Position	Age
Susie	Student	24
Bob	Student	24
Joan	TA	28
Robert	Professor	75
Arthur	Chancellor	80

Central Tendency - Mean

- The problem with mean is its sensitivity to outlier values
- Trimmed mean
  - Mean calculated after removing extreme outlier values
- Weighted mean
  - Mean calculated using weights for each value

Name	Position	Age
Susie	Student	24
Bob	Student	24
Joan	TA	28
Robert	Professor	75
Arthur	Chancellor	80

Central Tendency – Median and Mode

- Median
  - Middle value in an ordered set of data values that separates lower half from upper half
  - Example: 28 for Age attribute
- Mode
  - The value that occurs most frequently in the set of data values
  - Example: 24 for Age attribute

Name	Position	Age
Susie	Student	24
Bob	Student	24
Joan	TA	28
Robert	Professor	75
Arthur	Chancellor	80

Central Tendency – Midrange

- Midrange
  - The average of the lowest and highest value in the set of data values
- Example – Age attribute:
  - (24 + 80) / 2 = 52

Name	Position	Age
Susie	Student	24
Bob	Student	24
Joan	TA	28
Robert	Professor	75
Arthur	Chancellor	80

Symmetric/Asymmetric data

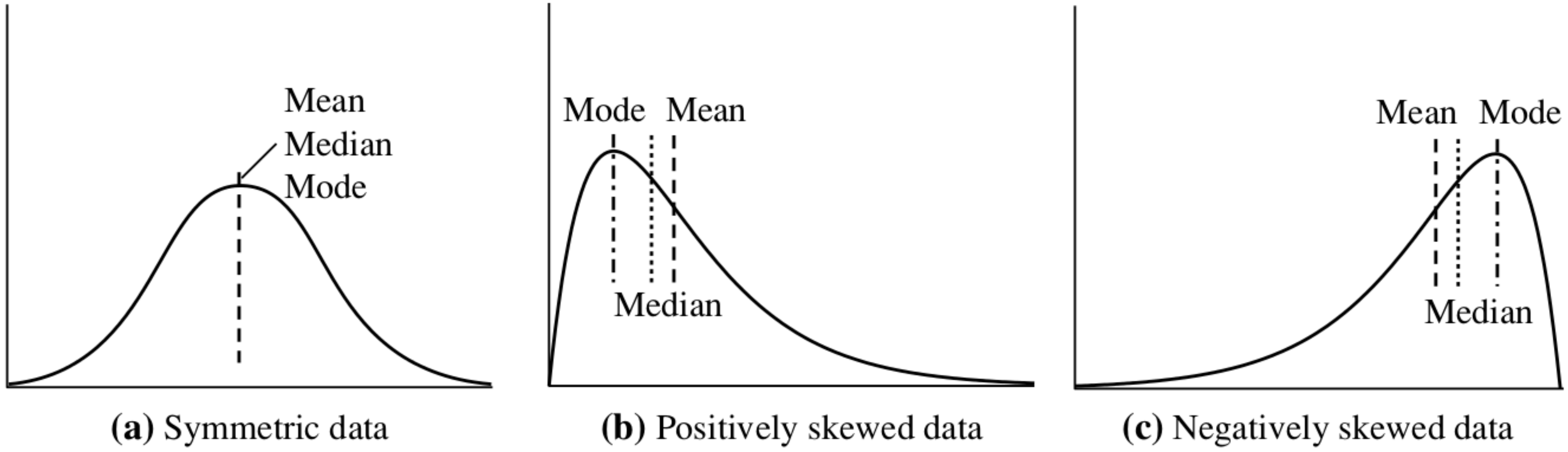
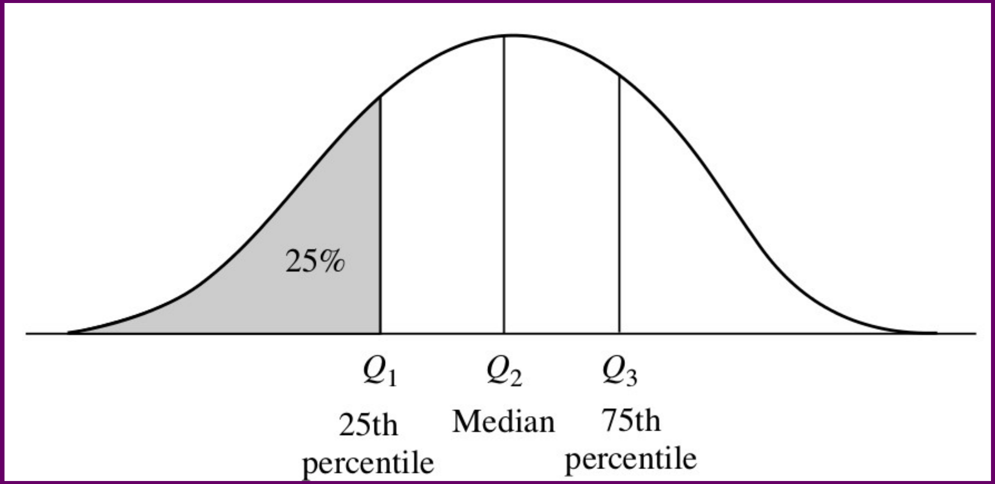


Figure 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

Data Dispersion – Range and Quantiles

- Range of a set of data values
  - Difference between largest and smallest value
- Quantiles
  - Selecting specific data points that divide sorted data distribution into equal-size sets
  - 4-Quantiles = Quartiles (image)
    - Interquartile Range (IQR) = Difference between Q_3 and Q_1
  - 100-Quantiles = Percentiles



Data Dispersion – Variance/Standard Deviation

- Measurement of how close the data values tend to be to the mean
  - Low standard deviation = values are close to mean
  - High standard deviation = values are spread out large range

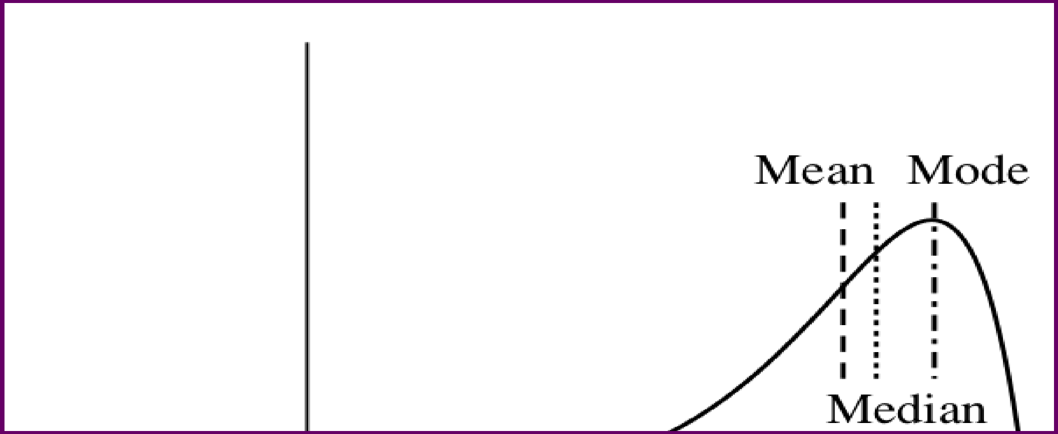
The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

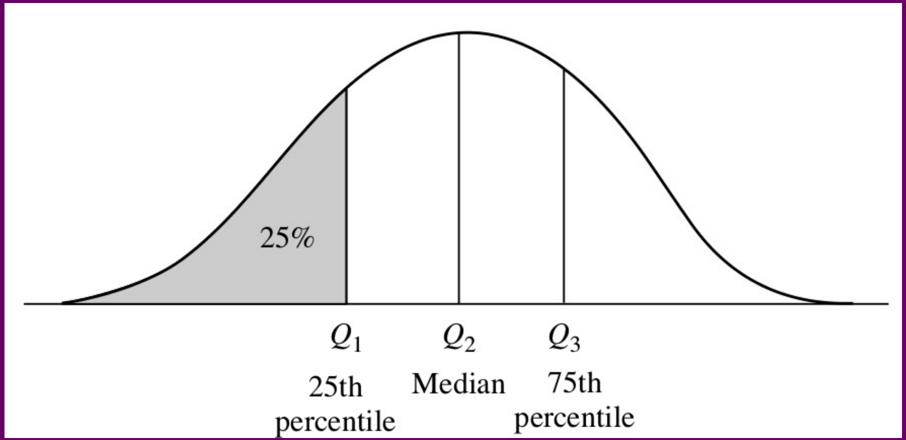
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2, \tag{2.6}$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

Data Dispersion – Five Number Summary

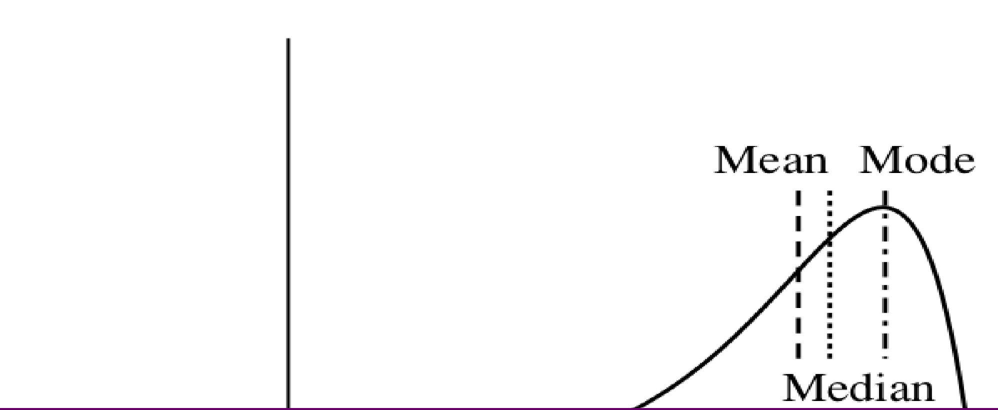
- No single measure is enough to describe skewed data
- Hence the Five-Number Summary
  - Minimum Value
  - Q_1
  - Median (Q_2)
  - Q_3
  - Maximum Value

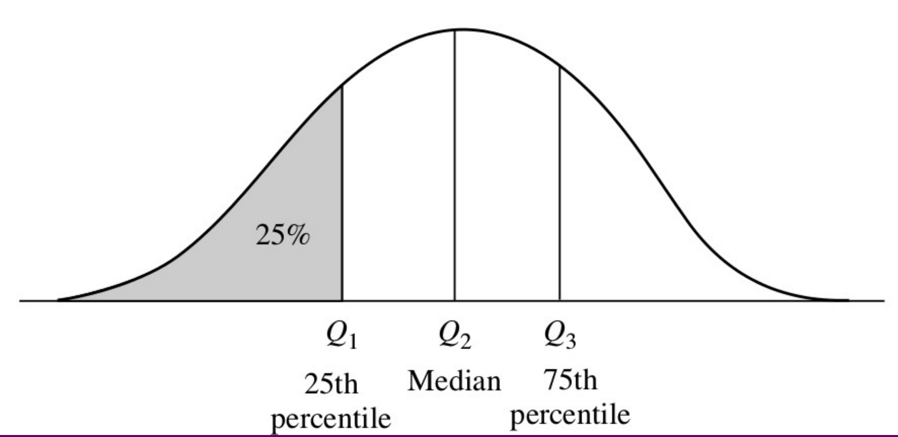




Data Dispersion – Five Number Summary

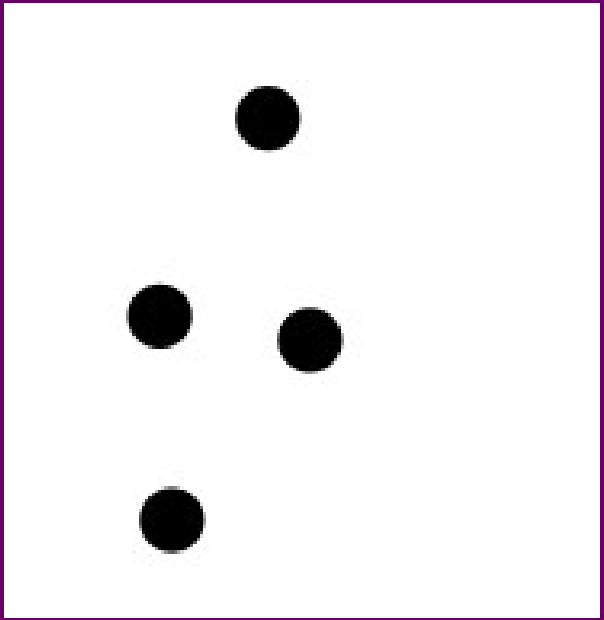
- The Five-Number Summary and its visualization (Boxplot) is used to detect outlier values in the data
- Outlier value
  - attribute value that is distant from the rest
  - can be the result of errors during data collection or represent odd behaviours
  - Rule of thumb: value is outlier if 1.5 IQR below Q_1 or above Q_3
    - Only rule of thumb, may not always be true!





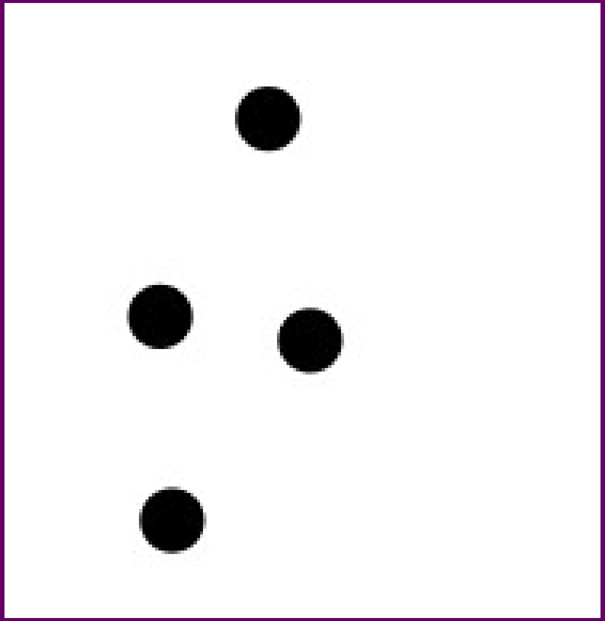
Data Similarity/Dissimilarity

- Used to measure “difference” between two data objects
  - Used in clustering, outlier anylysis and nearest-neighbor classification
- Similarity measure will typically return 0 if two data objects are completely unalike and 1 if they are the same
- Dissimilarity measure works the opposite way



Data Similarity/Dissimilarity

- Different dissimilarity measures for each attribute type
  - See section 2.4.2 – 2.4.5 in book
- Used when the data the object is only made up of one kind of attribute type
- But what do we do if the data objects consists of mixed attribute types?



Data Similarity/Dissimilarity

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

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}, \tag{2.22}$$

where the indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing (i.e., there is no mea-

surement of attribute f for object i or object j), or (2) $x_{if} = x_{jf} = 0$ and attribute

Data Similarity/Dissimilarity

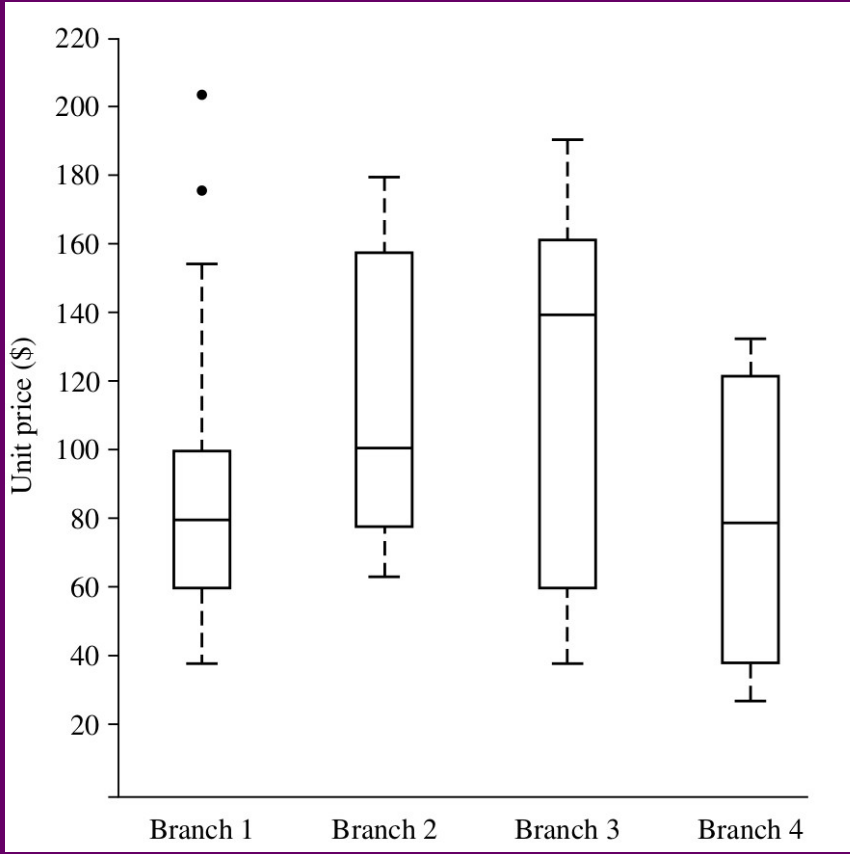
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

- To compute $|x_{if} – x_{jf}|$ (distance) we can use different distance measures
  - Euclidian, Manhattan, Minkowski

Visualizing your data

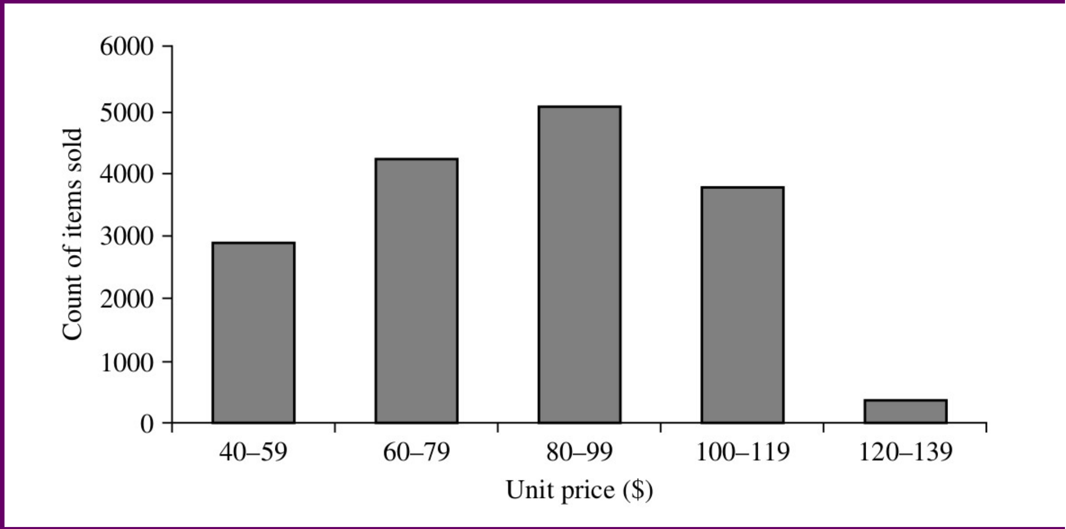
Boxplot

- Visualization of Five-Number Summary
  - Ends of box are Quartile 1 and 3, box length = IQR
  - Median marked by line inside box
  - Two lines extend from top and bottom of box to maximum and minimum values (within 1.5 IQR)
  - Values outside IQR x 1.5 are marked with dots
- Good for comparing attribute values across different data sets



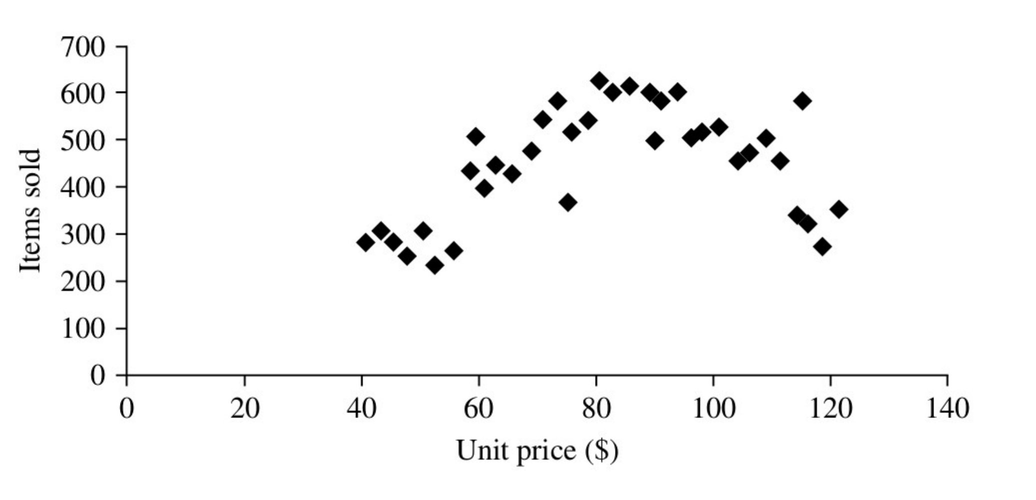
Histograms

- Visualization of distribution of attribute values
- Values divided into buckets/bins (Numeric)
  - Bucket range = width
  - Typically buckets are of the same width
- Can be used in data reduction



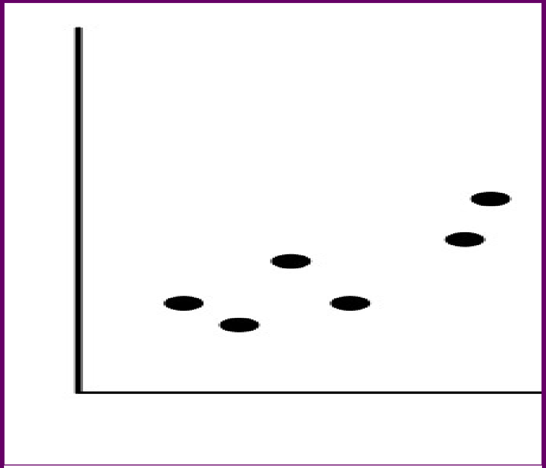
Scatter plots

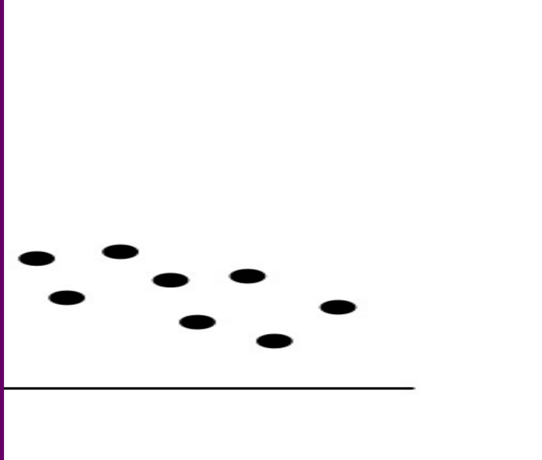
- Used for determining pattern, trend or relationship between two attributes
- Each pair of attribute values are treated as (x,y)-coordinates and are then plotted in to create the scatter plot
- Two attributes are correlated if one attribute imply the other



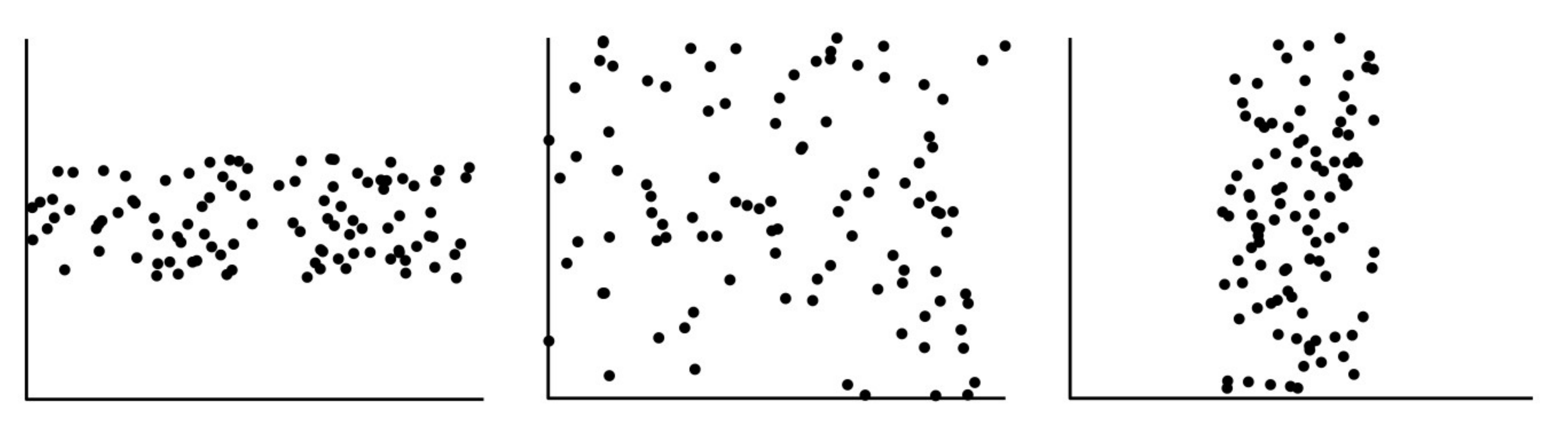
Scatter plots and data correlation

- Positive correlation = when attribute x increases, then attribute y **increases** (image a)
- Negative correlation = when attribute x increases, then attribute y **decreases** (image b)



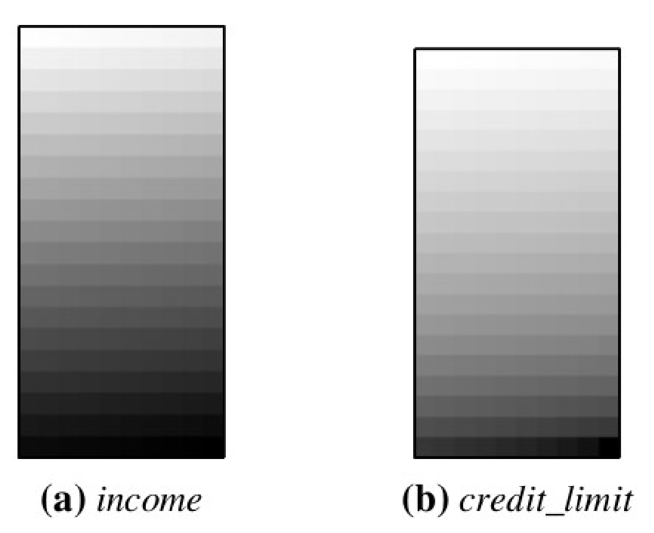


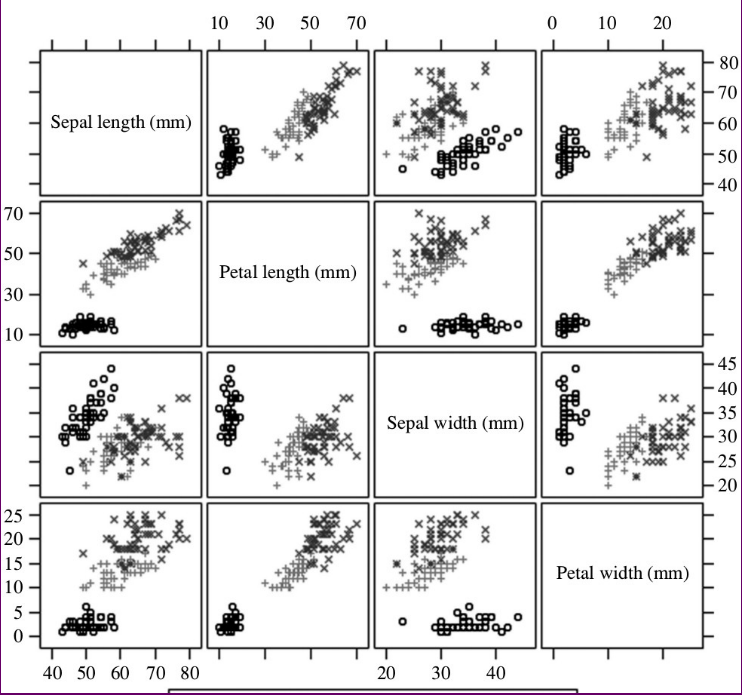
Scatter plots and data correlation

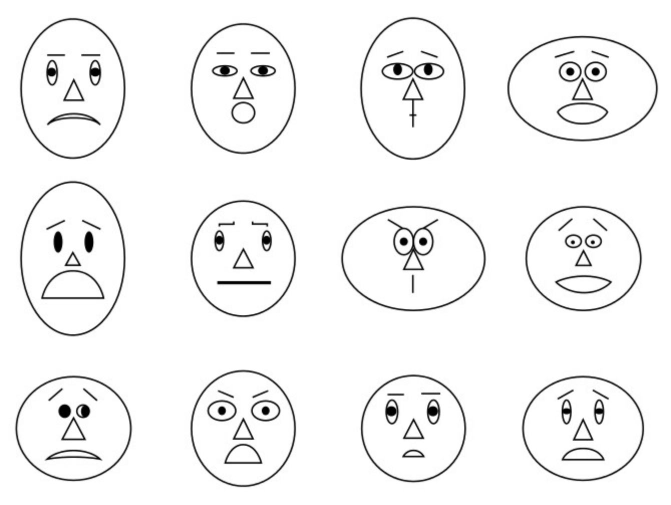


- Examples of no correlation present in three scatterplots

Other visualization types





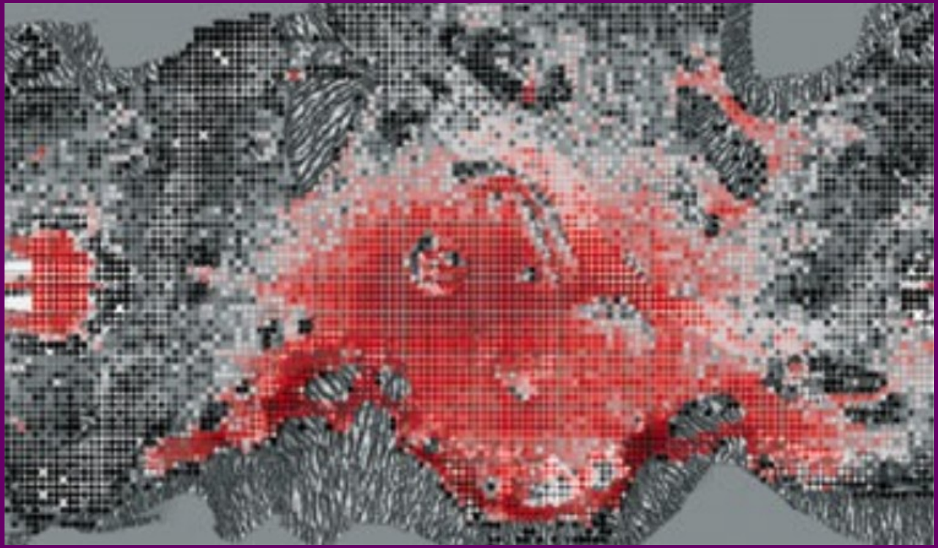


Other visualization types – heat maps

- Plot some attributes over a (virtual/physical) map
- The higher the value of the attribute, the higher the temperature
- The attributes are often counts
  - e.g. number of deaths

Heat maps and Halo 3

Number of deaths



Player Navigation

Source: How Microsoft Labs Invented a New Science of Play,Thompson, Wired, 2007 – Slide adapted from Hector Martinez

Cleaning your data

Data Cleaning

- Deal with missing data
- Smooth data i.e. identify and reduce noise and/or outliers
  - Binning, regression, outlier analysis
- Identify and remove redundant and inconsistent attributes
  - Pearson’s correlation, scatter plots

Missing Data

- Ignore tuple
  - Loss of data in other attributes. Use with caution!
- Fill in missing value manually
  - May be time consuming and not feasible because of data set size
- Use a global constant like “Unknown” to fill in missing value
  - Data mining program may think there is an exciting pattern or concept involving “Unknown”
- Use central tendency measure to fill in missing value
  - Use mean for symmetric data, otherwise use median

Missing Data

- Use attribute mean/median for all attribute values belonging to same class as tuple missing value
- Use the most probable value
  - Found for instance via regression

Missing Data

Every method involving inserting a replacement value may bias the data, i.e. the fill-in value may not be correct

Data Smoothing

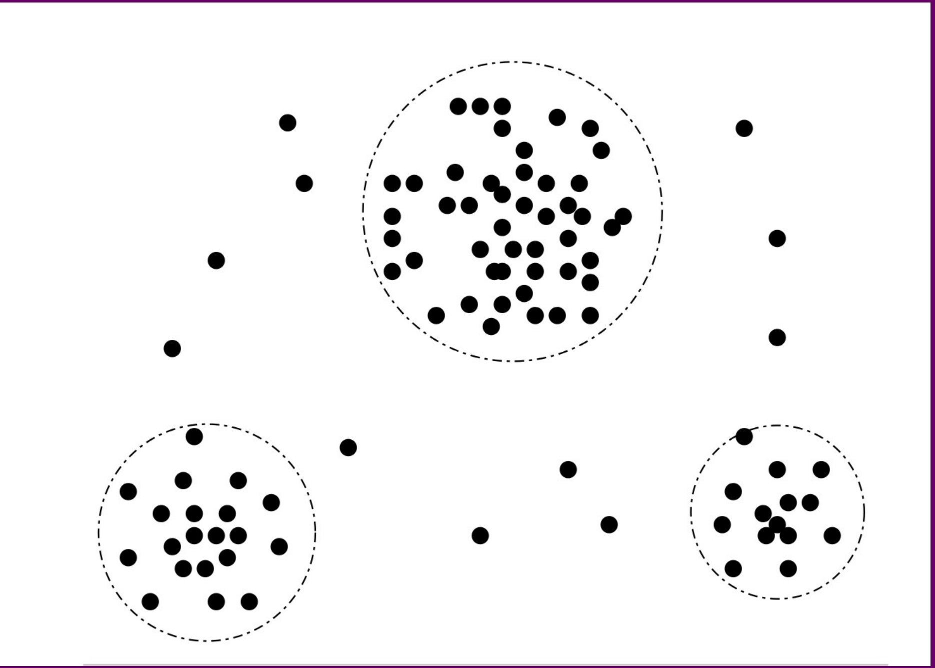
- Data smoothing is used to remove ***noise*** from the data.
  - Noise is a random error or variance in a measured variable
- Binning i.e. smoothing by looking at your neighbors
  - Values are sorted and then distributed into a number of equal-size buckets or bins
  - Smoothing by bin means
    - All values in each bin replaced with bin mean
  - Smoothing by bin medians
    - All values in each bin replaced with bin median
  - Smoothing by bin boundaries
    - Each bin value replaced with closest boundary value in the bin

Data Smoothing

- Data smoothing is used to remove ***noise*** from the data.
  - Noise is a random error or variance in a measured variable
- Binning i.e. smoothing by looking at your neighbors
  - Values are sorted and then distributed into a number of equal-size buckets or bins
  - Smoothing by bin means
    - All values in each bin replaced with bin mean
  - Smoothing by bin medians
    - All values in each bin replaced with bin median
  - Smoothing by bin boundaries
    - Each bin value replaced with closest boundary value in the bin
- Regression: Smooth by fitting the data into regression functions

Outlier Analysis

- Outliers found via cluster analysis
- Data is divided into clusters, based on how close each data point is to each other
- Data points found to be outside some cluster range are considered an outlier
- More on clusters later in the course



Data Redundancy

- An attribute is redundant if it can be derived from one or more other attributes
  - Example: *area, width, height*
- Can be detected using visual means like scatter plots
- Can be detected using correlation analysis (chapter 3.3.2)
  - Nominal data: chi-square test
  - Numerical data: Pearson's correlation coefficient
    - If found correlation coefficient is 0, then no correlation. If larger than 0, then positive correlation. If less than 0, then negative correlation

Reducing data

Data Reduction

- Data mining using huge data sets can take a long time, which may make it impossible/infeasible to complete
- Data reduction investigates whether it is possible to reduce the data set while still retaining (or almost retaining) all the characteristics of original data set

Data Reduction

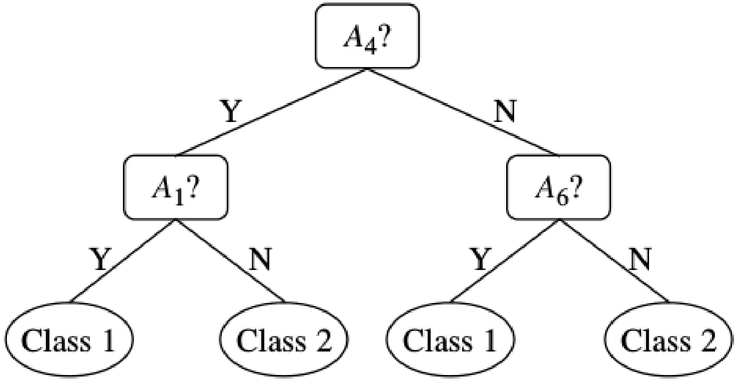
- Reduction strategies
  - Dimensionality reduction: reducing the number of attributes under consideration
    - Wavelet transform (3.4.2); Principal Components Analysis (3.4.3); Attribute Subset Selection (3.4.4)
  - Numerosity reduction: replacing data with a smaller-size representation
    - Parametric methods create a model to estimate data. Data parameters are stored instead of actual data. Example: regression
    - Non-parametric methods store reduced representation of actual data set. Examples: Histograms, clustering and sampling
  - Data compression: Data is transformed into reduced representation. Lossless if original data can be recreated from reduced representation. Lossy if only approximation can be recreated.

Attribute Subset Selection

- Based on the data mining task at hand we may be able to remove attributes that we deem to be irrelevant
  - Domain experts may be able to do this, but can be difficult and time-consuming
  - If we accidentally remove relevant attributes, data mining results will suffer
- Goal of attribute subset selection algorithms is to produce smallest set of relevant attributes

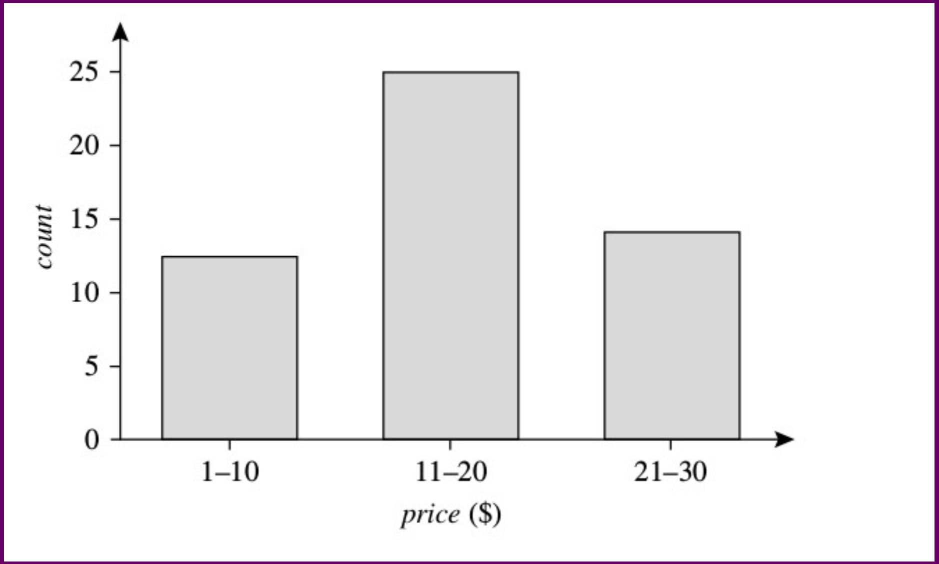
Attribute Subset Selection Strategies

- Measure of "best" attribute needed
- Usually statistical significance or other measure like *information gain* (more on this later in course)

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

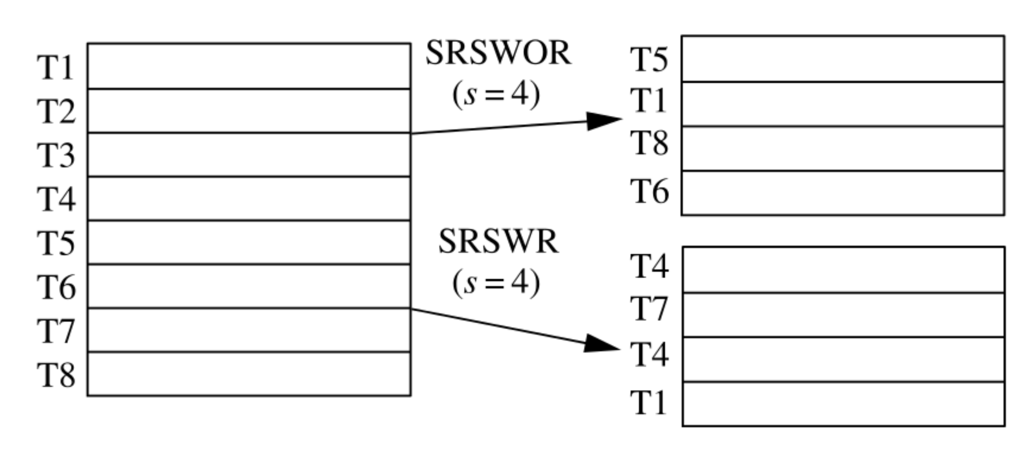
Histograms

- Histograms can be used to transform numerical data into nominal by partitioning numerical data into bins
- Equal-width bins
  - Range of buckets is same
- Equal-frequency bins
  - Each bucket contains roughly the same number of data samples



Sampling

- Sampling can be used to create a smaller sized version of the data set
- Smaller sized data set is created by randomly selecting data points in original data set
- Different sampling strategies in book (3.4.8), two examples:
  - SRSWOR: Simple random sample without replacement
  - SRSWR: Simple random sample with replacement



Transforming Data

Data Transformation Overview

- Smoothing
- Attribute construction
  - Making new attributes based on other attributes (e.g. width/height => area)
- Aggregation
  - Summary or aggregation operations applied on data, e.g. converting daily sales to monthly sales etc.
- Normalization
  - Scaling numerical values to fall inside a smaller range, e.g. [0;1]
- Discretization
  - Converting numerical data into categories (i.e. nominal data), e.g. 0-10, 10-20 or young/old. Done by for instance using binning or histograms
- Concept hierarchy generation for nominal data
  - Converting nominal data into higher level labes, e.g. street converted city.

Normalization

- Problem: Numerical attribute can affect data mining results
  - Centimeters (50 cm) vs meters (0.5 m)
- Can make attributes carry more weight in results
- Therefore we use normalization to standardize numerical values into a common range, e.g. [0;1]
- Different normalization techniques
  - Min-max: Maps values based on minimum and maximum values of the attribute
  - Z-score: Maps values based on attribute mean. Useful when minimum and maximum is unknown
  - Decimal scaling: Normalization done by moving decimal point, e.g 30 becomes 0.3

Min-Max Normalization

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v'_i in the range $[new\_min_A, new\_max_A]$ by computing

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \tag{3.8}$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

Conclusion

Why visualization and descriptive statistics?

- Data is too complex to evaluate it by simply look at it
- Visualization and data statistics help us understand the data and the results, and identify problems

Why cleaning and preprocessing?

- Garbage in, garbage out!
  - Cleaning and preprocessing is need because of dirty data in the real world
    - Needed in order to ensure optimal data mining results
  - Many problems can be avoided with better questionnaire and data collection design
    - As you will find out in today's lab...
- Create the dataset that you need from the data that you have

Food for thought: Getting what you ask for

- One thing is poorly designed questionnaires that yield data of poor quality
- Another thing is tailoring or “framing” your questionnaire questions to get the answer you want
- Example: Yes, Minister (BBC comedy series): https://www.youtube.com/watch?v=G0ZZJXw4MTA
- Were any of the questions in last weeks questionnaire framed?
- What about opinion polls in the national debate?

Thanks for listening!

How did I do? Send questions or feedback to andershh@itu.dk