Leibniz Universität Hannover

Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme
AG Intelligente Systeme - Data Mining group

Data Mining I

Summer semester 2019

Lecture 2: Getting to know your data

Lectures: Prof. Dr. Eirini Ntoutsi

TAs: Tai Le Quy, Vasileios Iosifidis, Maximilian Idahl, Shaheer Asghar, Wazed Ali

Recap from previous lecture

- KDD definition
- KDD process
- DM step
- Supervised (or predictive) vs Unsupervised (or descriptive) learning
- Main DM tasks
  - Clustering: partitioning in groups of similar objects
  - Classification: predict class attribute from input attributes, class is categorical
  - Regression: predict class attribute from input attributes, class is continuous
  - Association rules mining: find associations between attributes
  - Outlier detection: identify non-typical data

Warming up (5') – Learning from student data

- Continuing our example from last lecture regarding student data and what sort of knowledge one can extract upon such sort of data

- If students are the learning instances, what sort of features could I use to describe each of them?
- What could be the feedback/label for the learning model (if any)?

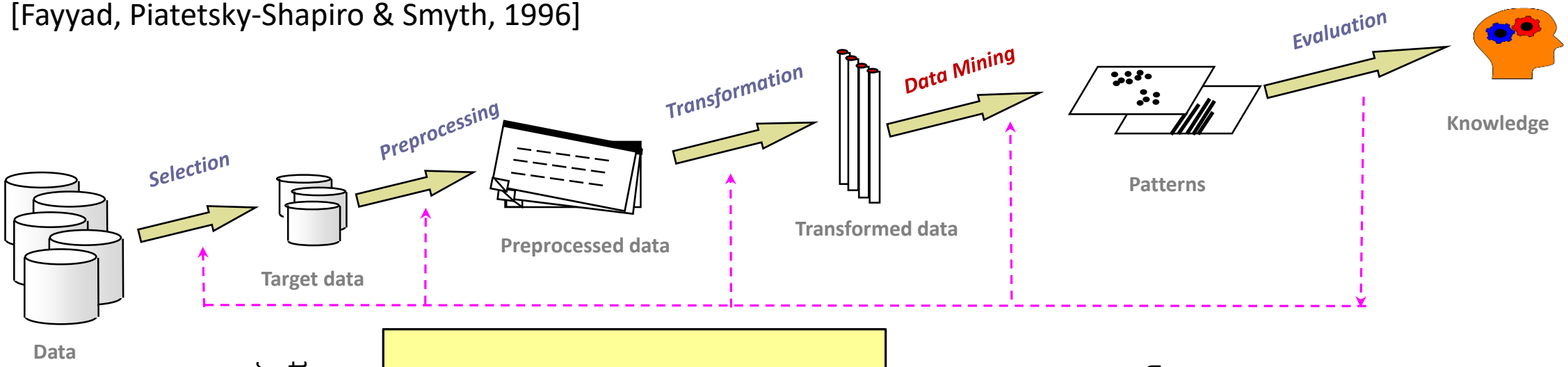
- What could be a supervised learning task?
- What could be an unsupervised learning task?
- What could be an outlier detection task?

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Recap: The KDD process and the Data Mining step

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Data Mining:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

Why data preprocessing?

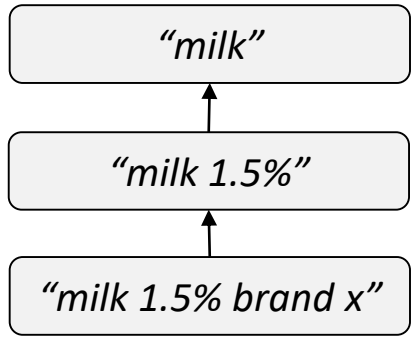
- Real world data is noisy, incomplete and inconsistent:
  - Noisy: errors/ outliers
    - erroneous values : e.g., *salary* = -10K
    - unexpected values: e.g., *salary* = 100K when the rest dataset lies in [30K-50K]
  - Incomplete: missing data
    - missing values: e.g., *occupation*=“ ”
    - missing attributes of interest: e.g., no information on occupation
  - Inconsistent: discrepancies in the data
    - e.g., *student grade* ranges between different universities might differ, in DE [1-5], in GR [0-10]
- “Dirty” data → poor mining results
- Data preprocessing is necessary for improving the quality of the mining results!
- Not a focus of this class!

“Garbage in, garbage out”

Your analysis is as good as your data.

Major tasks in data preprocessing

- Data integration:
  - Integration of multiple databases, data warehouses, or files (entity identification)
- Data cleaning:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data transformation:
  - Normalization in a given range, e.g., [0-1]
  - Generalization through some concept hierarchy
- Data reduction:
  - Duplicate elimination
  - Aggregation, e.g., from 12 monthly salaries to average salary per month.
  - Dimensionality reduction, through e.g., PCA, autoencoders



More on this on the "Data Mining II" course

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Datasets = instances + features

- Datasets consists of instances (also known as examples or objects or observations)
  - e.g., in a university database: *students*, *professors*, *courses*, *grades*,…
  - e.g., in a library database: *books*, *users*, *loans*, *publishers*, ….
  - e.g., in a movie database: *movies*, *actors*, *director*,…
- Instances are described through features (also known as attributes or variables or dimensions)
  - E.g. a course is described in terms of a *title*, *description*, *lecturer*, *teaching frequency* etc.
- The feedback feature (for supervised learning) is called the class attribute

Data matrix

- Data can often be represented or abstracted as an $D = n \times d$ data matrix
  - n rows corresponding to instances
  - d columns correspond to features, feature set F

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

- The number of instances n is referred to as the size or cardinality of the dataset, $n = |D|$
- The number of features d is referred to as the dimensionality of the dataset
- Subset of the data: $D' \subseteq D$
- Subspace $F' \subseteq F$
  - Subspace projection

An example from the iris dataset

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Iris Versicolor | Iris Setosa | Iris Virginica

Basic feature types

- Binary/ Dichotomous variables
- Categorical (qualitative)
  - Binary variables
  - Nominal variables
  - Ordinal variables
- Numeric variables (quantitative)
  - Interval-scale variables
  - Ratio-scaled variables

Binary/ Dichotomous variables

- The attribute can take two values, *{0,1}* or *{true,false}*
  - usually, 0 means absence, 1 means presence
  - e.g., smoker variable: 1→ smoker, 0→ non-smoker
  - e.g., *true (1), false (0)*
- Are both values equally important?
  - Symmetric binary: both outcomes are equally important
    - e.g., *gender* (*male*, *female*)
  - Asymmetric binary: outcomes are not equally important
    - e.g., *medical tests* (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., *HIV positive*)

Person	isSmoker
Eirini	0
Erich	1
Kostas	0
Jane	0
Emily	1
Markus	0

What are the binary variables in the example below?

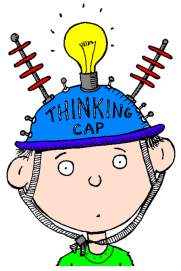
ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Categorical: Nominal variables

- The attribute can take values within a set of *M* categories/ states (binary variables are a special case)
  - *No ordering* in the categories/ states.
  - Only *distinctness relationships* apply, i.e.,
    - *equal (=)* and
    - *different* (≠)
  - Examples:
    - *Colors* = {*brown, green, blue,…,gray*},
    - *Occupation* = {engineer, doctor, teacher, …, driver}

Person	Occupation
Eirini	archaeologist
Erich	engineer
Kostas	doctor
Jane	engineer
Emily	teacher
Markus	driver

Operations that can be applied: $=, \neq$



What are the categorical variables in the example below?

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Categorical: Ordinal variables

- Similar to nominal variables, but the *M* states are ordered/ ranked in a meaningful way.
  - There is an *ordering* between the values.
  - Allows to apply *order relationships*, i.e., $>, \geq, <, \leq$
  - However, the *difference* and *ratio* between these values has no meaning.
    - E.g., 5*-3* is the same as 3*-1* or, 4* is 2 times better than 2*?
  - Examples:
    - *School grades*: *{A,B,C,D,F}*
    - *Movie ratings*: {hate, dislike, indifferent, like, love}
      - Also, *movie ratings*: {*, **, ***, ****, *****}
      - Also, *movie ratings*: {1, 2, 3, 4, 5}
    - *Medals* = {bronze, silver, gold}

Person	A beautiful mind	Titanic
Eirini	5*	3*
Erich	5*	1*
Kostas	3*	3*
Jane	1*	2*
Emily	2*	5*
Markus	4*	3*

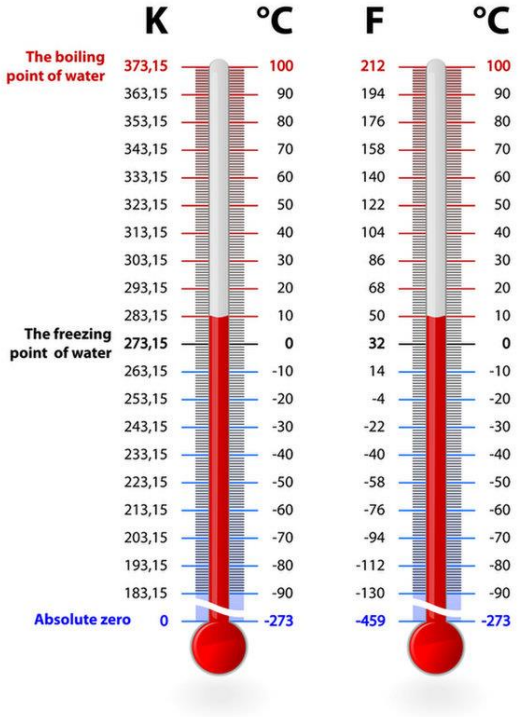
Operations that can be applied: $=, \neq, <, >$

What are the ordinal variables in the example below?

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Numeric: Interval-scale variables

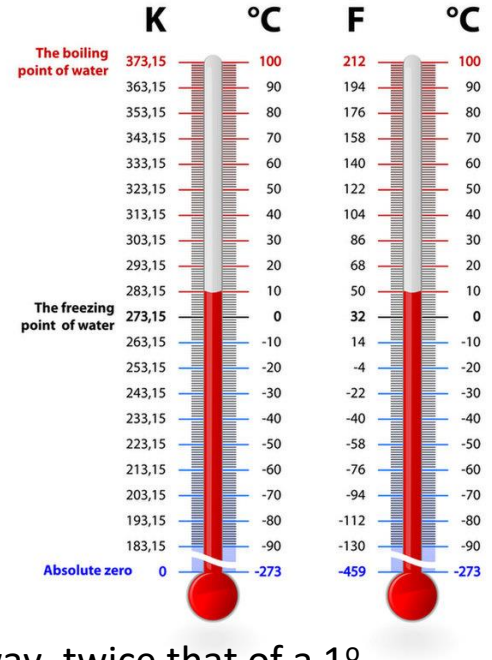
- Differences between values are meaningful
  - The difference between 90° and 100° *temperature* is the same as the difference between 40° and 50° *temperature*.
- Examples:
  - *Calendar dates* , *Temperature* in Farenheit or Celsius, ...
- Ratio still has no meaning
  - A *temperature* of 2° Celsius is not much different than a *temperature* of 1° Celsius.
  - The issue is that the 0° point of the Celsius scale is in a physical sense arbitrary and therefore the ratio of two *Celsius temperatures* is not physically meaningful.



Operations that can be applied: $=, \neq, <, >, +, -$

Numeric: Ratio-scale variables

- Both differences and ratios have a meaning
  - E.g., a 100 kgs person is twice heavy as a 50 kgs person.
  - E.g., a 50 years old person is twice old as a 25 years old person.
- Meaningful (unique and non-arbitrary) zero value
- Examples:
  - *age*, *weight*, *length*, *number of sales*
  - *temperature in Kelvin*
    - When measured on the Kelvin scale, a temperature of 2° is, in a physical meaningful way, twice that of a 1°.
      - The zero value is absolute 0, represents the complete absence of molecular motion



Operations that can be applied: $=, \neq, <, >, +, -, \times, \div$

What are the ratio-scale variables in the example below?

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Nominal, ordinal, interval-scale, ratio-scale variables: overview of operations

Table 1.1 ♦ Levels of Measurement, Arithmetic (

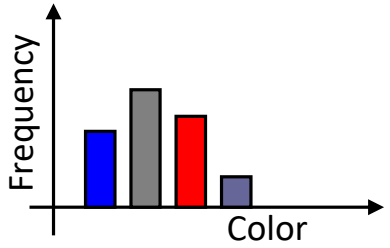
Stevens's Levels of Measurement	*Logical and Arithmetic Operations That Can Be Applied (According to Stevens)*
Nominal	=, ≠
Ordinal	=, ≠, <, >
Interval[b]	=, ≠, <, >, +, −
Ratio	=, ≠, <, >, +, −, ×, ÷

Source: https://www.sagepub.com/sites/default/files/upm-binaries/19708_6.pdf

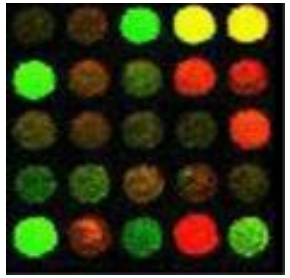
How do we extract features?

- In many cases, we are not given a feature description of the data, so we have to extract the features
- Feature extraction depends on the application

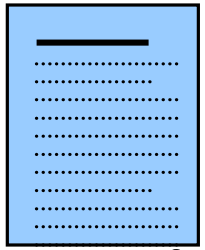
- Images:
  Color histograms: distribution of colors in the image



- Gene databases:
  gene expression level



- Text databases:
  Word frequencies

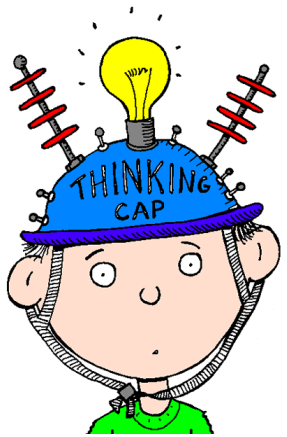


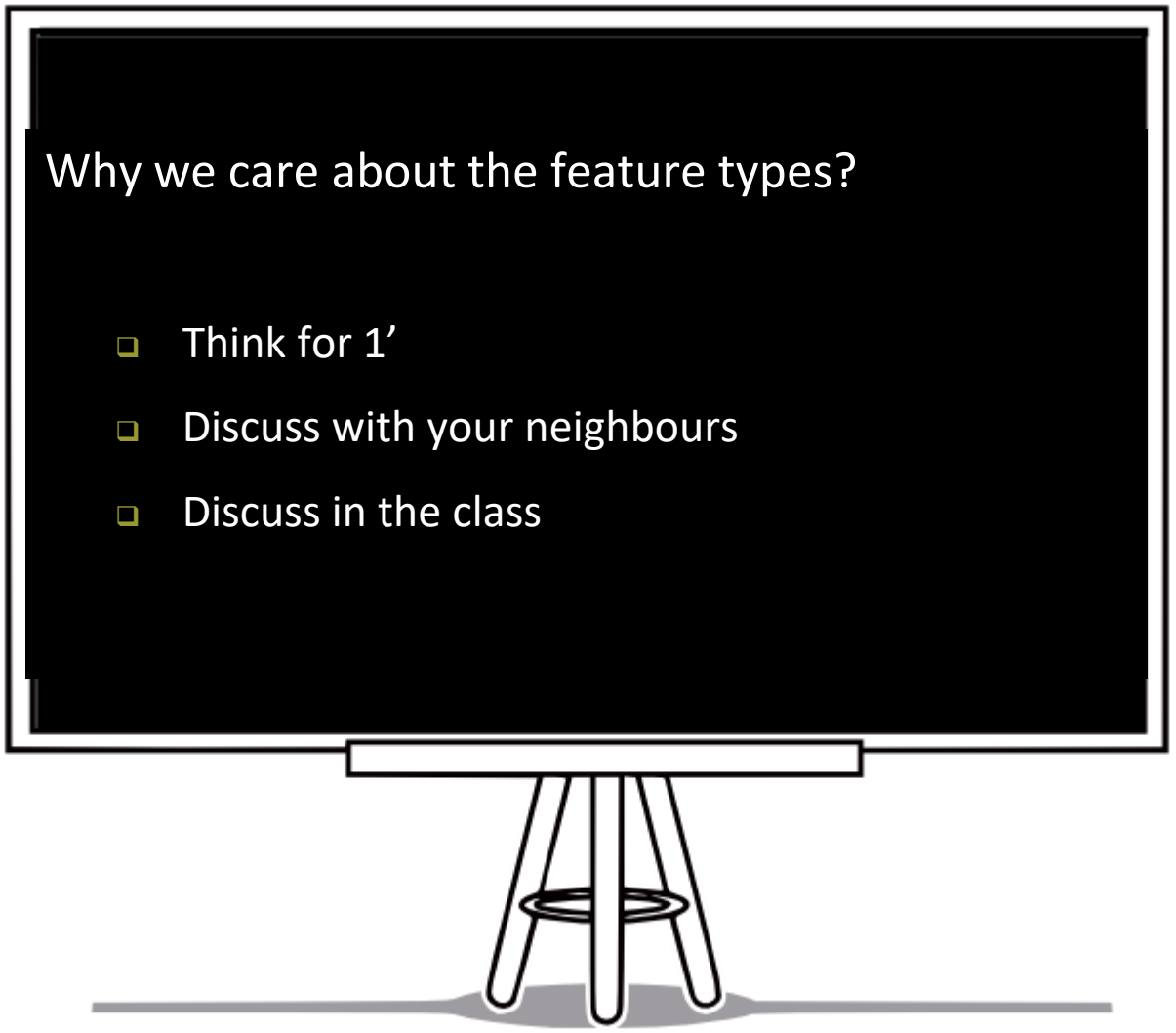
Data	25
Mining	15
Feature	12
...	

- Traditionally features were handcrafted.
- Nowadays, features can be also learned (e.g., through DNNs)
- Hybrid approaches also exist that combine handcrafted with learned features.

- But, the feature-based approach allows uniform treatment of instances from different applications.

Short break (5’)





Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Univariate vs bivariate vs multivariate analysis

- Univariate analysis: analysis of a single attribute
- Bivariate analysis: the simultaneous analysis of two attributes
- Multivariate analysis: the simultaneous analysis of more than two attributes

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Univariate descriptors: measures of central tendency

Let $x_1,\ldots,x_n$ be a random sample of an attribute X (the dataset projected w.rt. X). Measures of central tendency of X

include:

- (Arithmetic) mean/ center/ average:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- Weighted average:

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

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

What is the mean of:

3, 8, 3, 4, 3, 6, 4, 2, 3

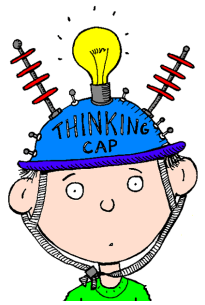
Univariate descriptors: measures of central tendency

- Mean is greatly influenced by outliers, a more robust measure is median

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- Median: the central element in ascending ordering
  - Middle value if odd number of values, or average of the middle two values otherwise

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

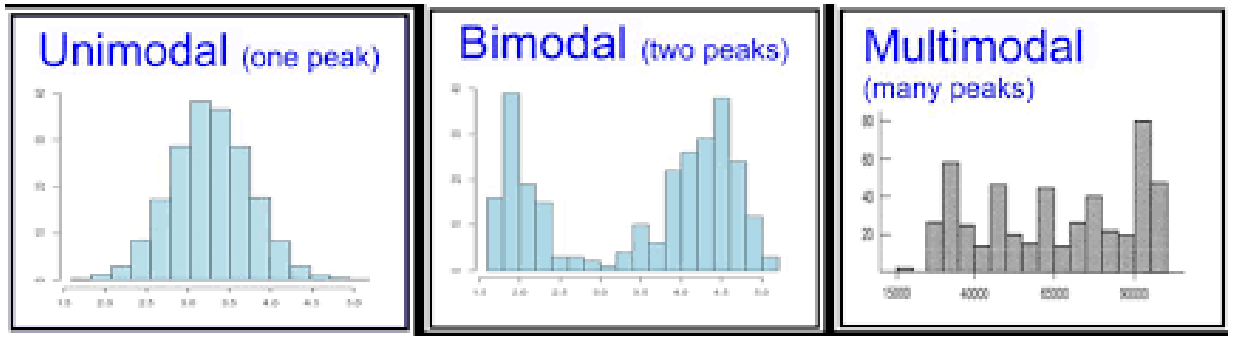


What is the median of:

3, 8, 3, 4, 3, 6, 4, 2, 3

Univariate descriptors: measures of central tendency

- Mode: the value that occurs most often in the data
  - Unimodal: 1 mode (peak)
  - Bimodal: 2 modes (peaks)
  - Multimodal: >2 modes (peaks)



$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

What is the mode of:

3, 8, 3, 4, 3, 6, 4, 2, 3

Unimodal – bimodal – multimodal distributions

- Bimodal: a distribution with two modes (peaks)
  - General term: Multimodal distributions

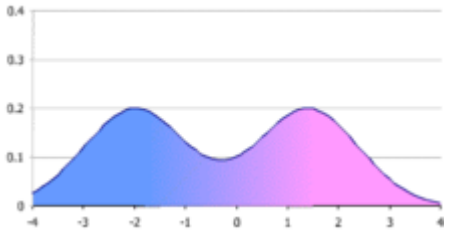


Figure 1. A simple bimodal distribution, in this case a mixture of two normal distributions with the same variance but different means. The figure shows the probability density function (p.d.f.), which is an equally-weighted average of the bell-shaped p.d.f.s of the two normal distributions. If the weights were not equal, the resulting distribution could still be bimodal but with peaks of different heights.

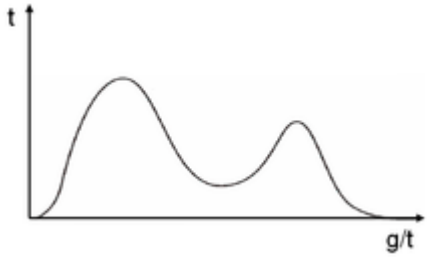
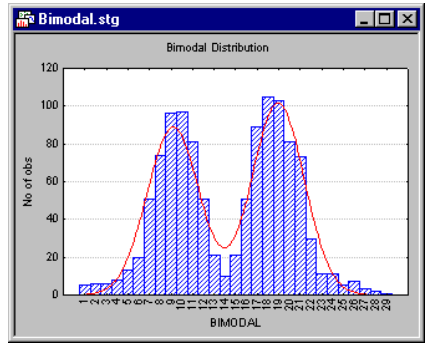


Figure 2. A bimodal distribution.



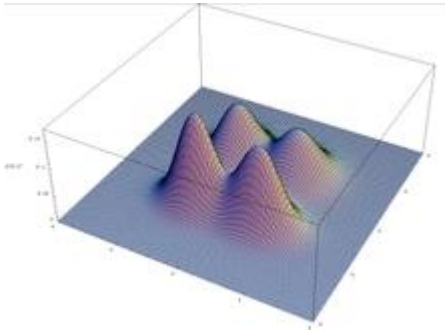


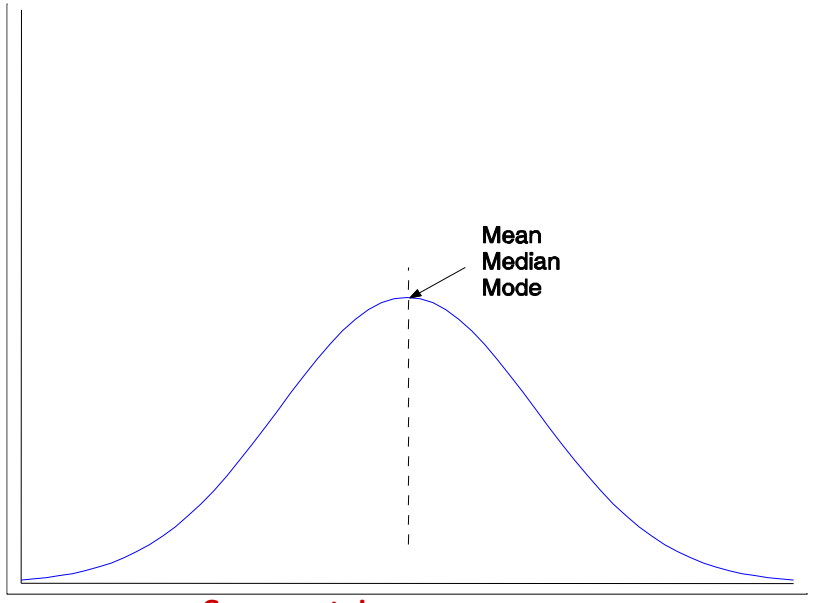
Figure 3. A bivariate, multimodal distribution

Bimodality of the distribution in a sample is often a strong indication that the distribution of the variable in population is not normal. *Bimodality* of the distribution may provide important information about the nature of the investigated variable (i.e., the measured quality). For example, if the variable represents a reported preference or attitude, then *bimodality* may indicate a polarization of opinions. Often, however, the *bimodality* may indicate that the sample is not homogenous and the observations come in fact from two or more "overlapping" distributions. Sometimes, *bimodality* of the distribution may indicate problems with the measurement instrument (e.g., "gage calibration problems" in natural sciences, or "response biases" in social sciences).

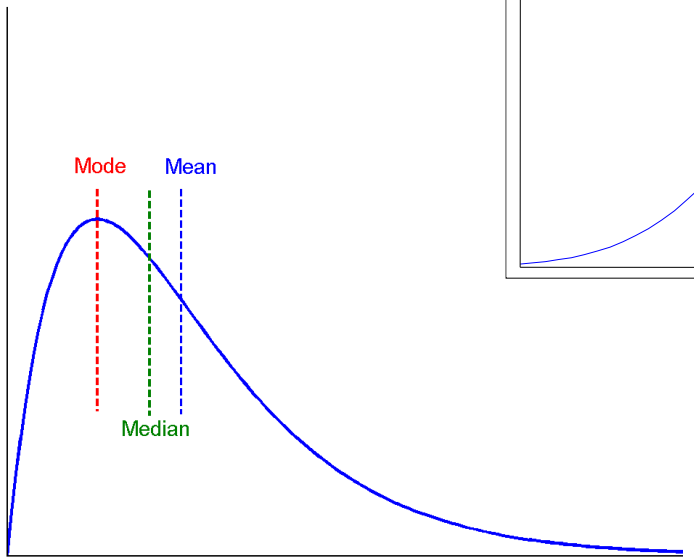
Source: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=glossary/GlossaryTwo/B/BimodalDistribution

Univariate descriptors: measures of central tendency

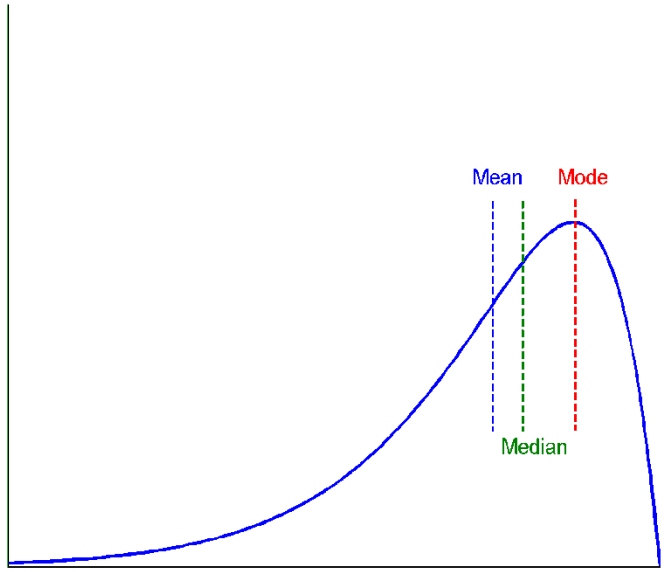
- Mean, median and mode in normal vs highly-skewed distributions



Symmetric



Positively skewed (or, skewed to the right)



Negatively skewed (or, skewed to the left)

Univariate descriptors: measures of spread

Let $x_1,\ldots,x_n$ be a random sample of an attribute X. The degree to which X values tend to spread is called dispersion or variance of X :

- Variance σ^2:

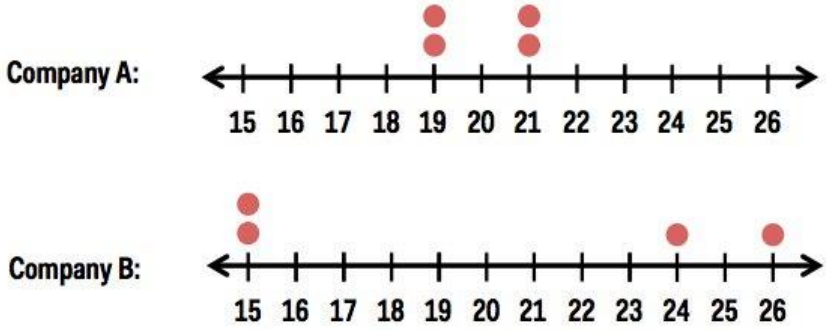
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation σ is the square root of the variance:

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

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Same mean (20), different spread



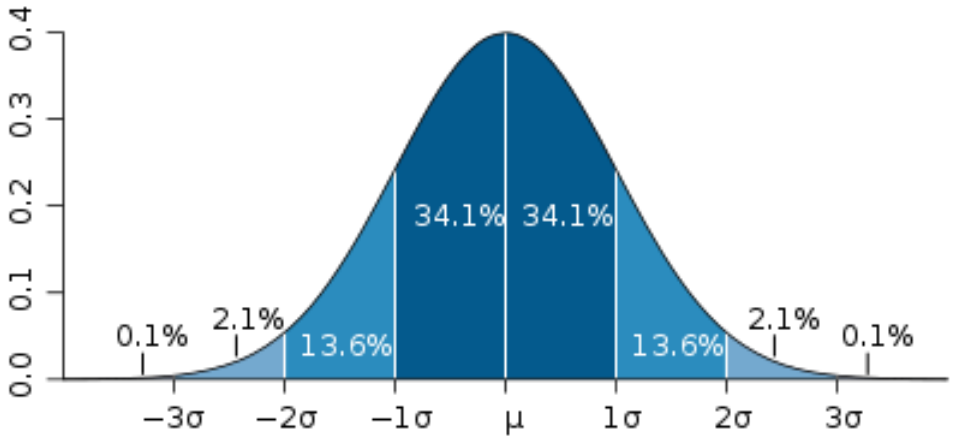
Source: http://www.businessinsider.com/standard-deviation-2014-12?IR=T

Univariate descriptors: measures of spread

Standard deviation appears as a parameter in a number of statistical and probabilistic formulas.

Example: for a normal distribution

- ~68% of values drawn from the distribution are within 1σ
- ~95% of the values lie within 2σ
- ~99.7% of the values lie within 3σ



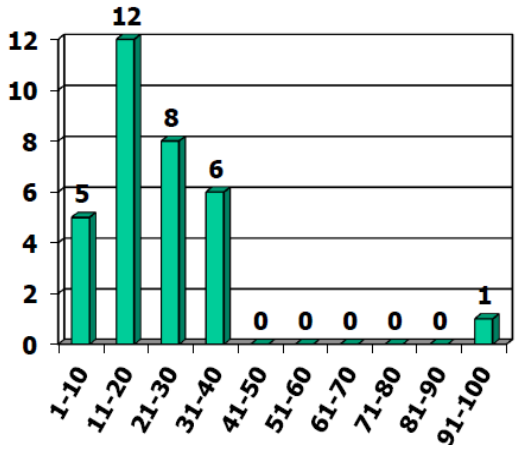
Source: http://en.wikipedia.org/wiki/Normal_distribution

Univariate descriptors: useful charts

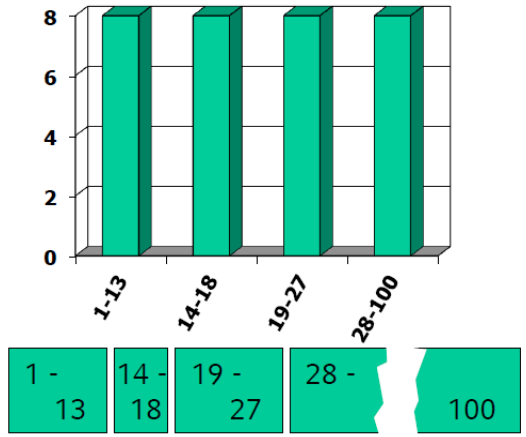
$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Let $x_1,\ldots,x_n$ be a random sample of an attribute X. For visual inspection of X, several types of charts are useful.

- Histograms:
  - Summarizes the distribution of X
  - X axis: attribute values, Y axis: frequencies
  - Absolute frequency: for each value a, # occurrences of a in the sample
  - Relative frequency: $f(a) = h(a)/n$
- Different types of histograms, e.g.:
  - Equal width:
    - It divides the range into N intervals of equal size
  - Equal frequency/ depth:
    - It divides the range into N intervals, each containing approximately same number of samples



Equal width histogram



Equal depth histogram

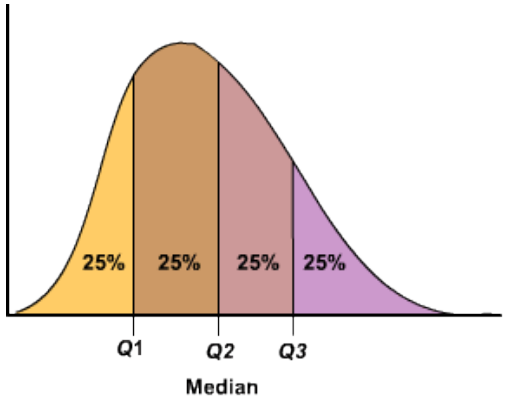
Source: http://www.dbs.ifi.lmu.de/Lehre/KDD/SS16/skript/2_DataRepresentation.pdf

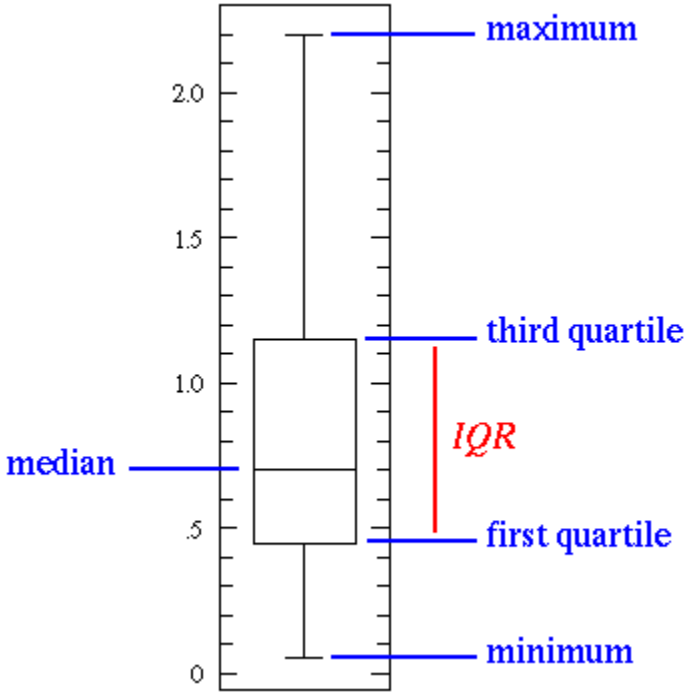
Univariate descriptors: useful charts

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

For visual inspection of an attribute X, several types of charts are useful.

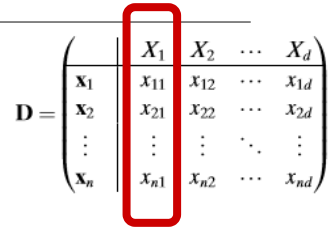
- Boxplots: a standardized way of displaying the distribution of data based on a 5 number summary:
  - *min*, Q_1, *median*, Q_3, *max*
    - Q_1 (25th percentile): 25% of the data follow below this percentile
    - Median (50th percentile): 50% of the data follow below this percentile
    - Q_3 (75th percentile): 75% of the data follow below this percentile
    - Range: max value –min value
    - The whiskers go from each quartile to min or max





Source: http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

Univariate descriptors: An example



- Sample: The *Annual Salaries* ($) for 20 Selected Employees at a Local Company, already sorted

30000 32000 32000 33000 33000 34000 34000 38000 38000 38000 42000
43000 45000 45000 48000 50000 55000 55000 65000 110000

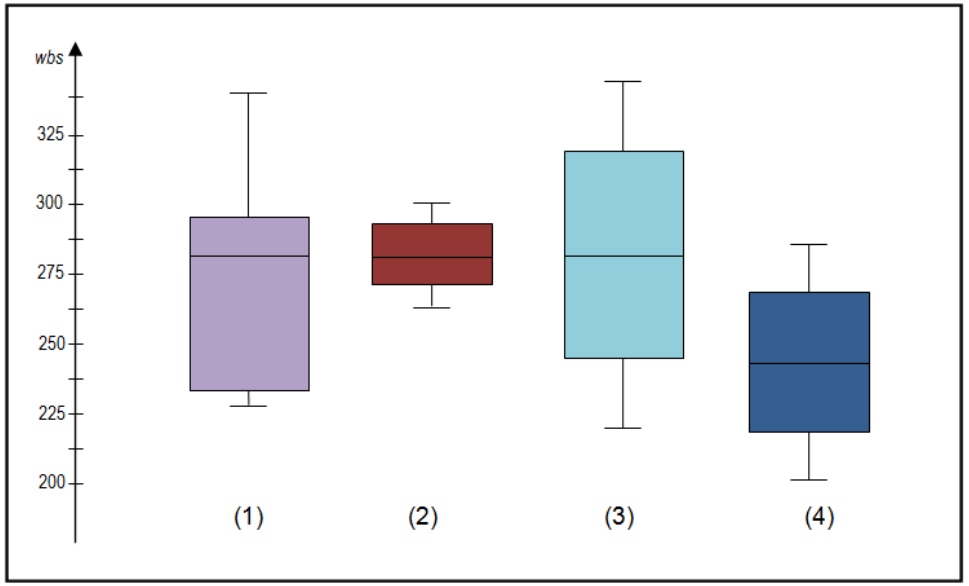
- How to compute the boxplot? (Recall a boxplot is a 5 number summary: *min*, Q_1, *median*, Q_3, *max)*
  - Order the data from smallest to largest
  - Find the median
  - Find the quartiles
    - Q1 is the median of the data points to the left of the median
    - Q3 is the median of the data points to the right of the median
  - Find min and max

Lowest	Lower Quartile (QL)	Median	Upper Quartile (QU)	Highest
$30,000	$33,500	$40,000	$49,000	$110,000

Univariate descriptors: useful charts

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

- Box plots are used to show overall patterns of response for a group. They provide a useful way to visualize the range and other characteristics of responses for a large group.



Source: http://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots

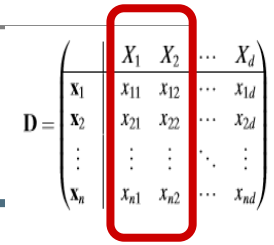
- Boxplot 2 is comparatively short: similar values
- Boxplots 1 and 3 are comparatively tall: quite different values
- …

Bivariate descriptors

- Given two attributes *X*, *Y* one can measure how strongly they are correlated
  - For numerical data → correlation coefficient
  - For categorical data → χ^2 (chi-square)

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Bivariate descriptors: for numerical features

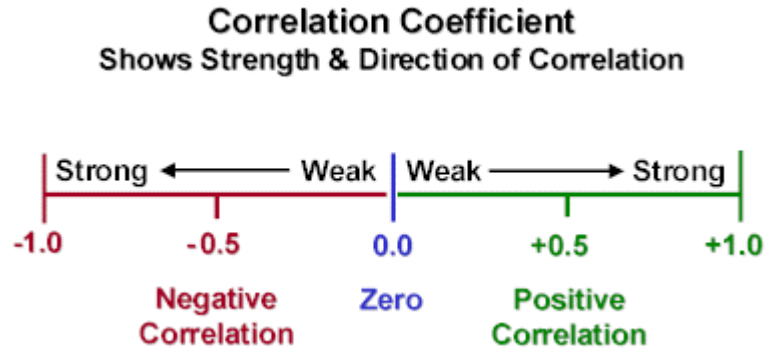


- Correlation coefficient (also called Pearson's correlation coefficient) measures the *linear* association between *X*, *Y*:

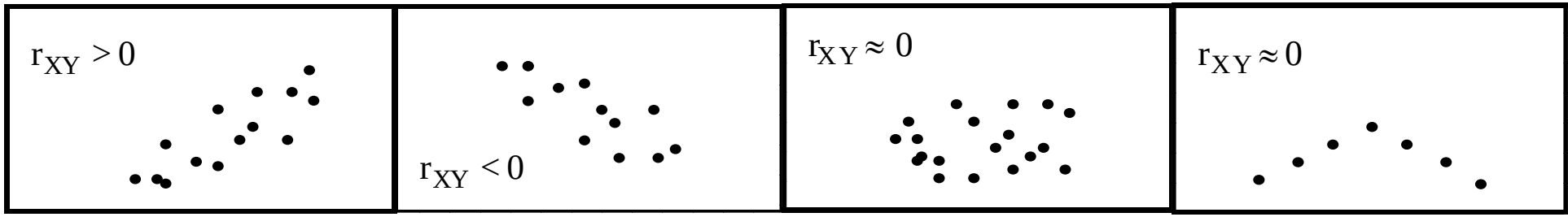
$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_X \sigma_Y}$$

  - x_i, y_i: the values in the ith tuple for *X*, *Y*

- value range: $-1 \leq r_{XY} \leq 1$
- the higher r_{XY} the stronger the correlation
  - $r_{XY} > 0$ positive correlation
  - $r_{XY} < 0$ negative correlation
  - $r_{XY} \sim 0$ no correlation/ independent



Source: https://psychlopedia.wikispaces.com/Correlation+Coefficien



Bivariate descriptors: for numerical features

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

- Visual inspection of correlation

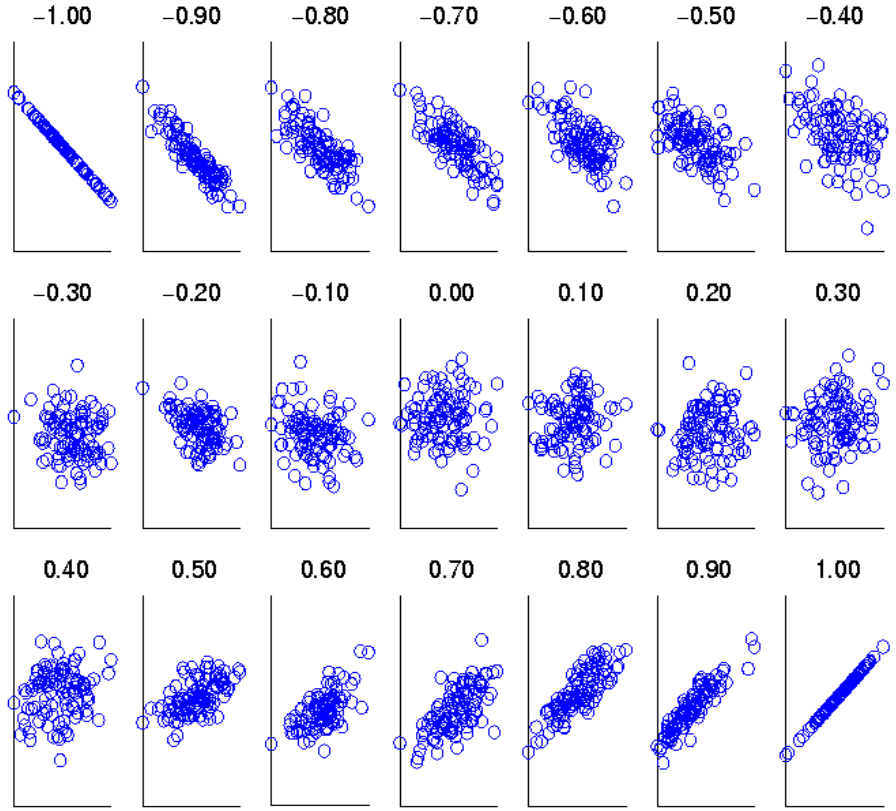
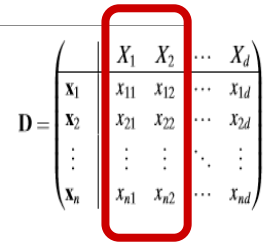


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

Bivariate descriptors: for categorical features



- The chi-square (χ^2) test tests whether two categorical variables $X=\{x_1, \ldots, x_c\}$, $Y=\{y_1, \ldots, y_r\}$ are independent (no relationship)
- How to compute the chi-square statistic? → use a contingency table
  - Represents the absolute frequency h_{ij} of each combination of values (x_i, y_j) and the marginal frequencies h_i, h_j of X, Y.

Attribute Y

Attribute X	*Medium-term unemployment*	*Long-term unemployment*	*Total*
No education	19	18	37
Teaching	43	20	63
Total	62	38	100

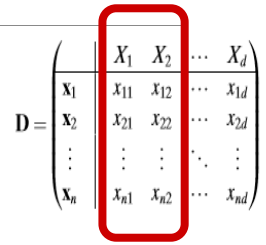
- Chi-square χ^2 test

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

o_{ij}: observed frequency
e_{ij}: expected frequency

$$e_{ij} = \frac{h_i h_j}{n}$$

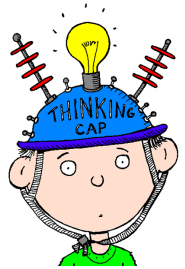
Bivariate descriptors: for categorical features



- Chi-square example
  - (numbers in parenthesis are the expected counts)

Attribute *Y*

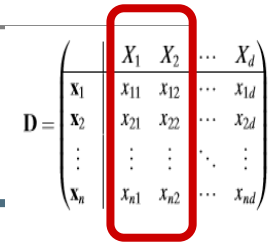
Attribute *X*	Play chess	Not play chess	Sum (row)
Like science fiction	250 (???)	200 (???)	450
Not like science fiction	50 (???)	1000 (???)	1050
Sum(col.)	300	1200	1500



What are the expected values?

Recall: $e_{ij} = \frac{h_i h_j}{n}$

Bivariate descriptors: for categorical features

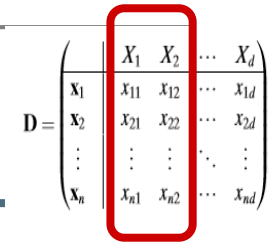


- Chi-square example
  - (numbers in parenthesis are the expected counts)

Attribute *Y*

Attribute *X*	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

Bivariate descriptors: for categorical features



- Chi-square example

Attribute *Y*

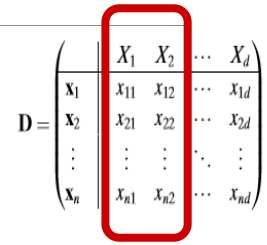
Attribute *X*	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation $\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- How do we interpret this value?
  - Using the table of critical values

Table of critical values

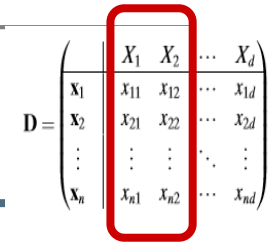


- Based on your desired confidence level (e.g., 95% → $p = 0.05$)
- Based on the degrees of freedom (e.g., 1 degrees of freedom)
- Check if your value is significant or nonsignificant

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Source: http://www.ox.ac.uk/media/global/wwwoxacuk/localsites/uasconference/presentations/P8_Is_it_statistically_significant.pdf

Bivariate descriptors: for categorical features



- Chi-square example

Attribute *Y*

Attribute *X*	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation $\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Look up the critical chi-square statistic value for e.g., p = 0.05 (95% confidence level) and
  - 1 degree of freedom (2-1)*(2-1)=1

Table of critical values

- Look up the critical chi-square statistic value for e.g., p = 0.05 (95% confidence level) with 1 degree of freedom ((2-1)*(2-1)=1) ➔ 3,84 < 507,93 so reject the hypothesis that they are not correlated

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Source: http://www.ox.ac.uk/media/global/wwwoxacuk/localsites/uasconference/presentations/P8_Is_it_statistically_significant.pdf

Outline

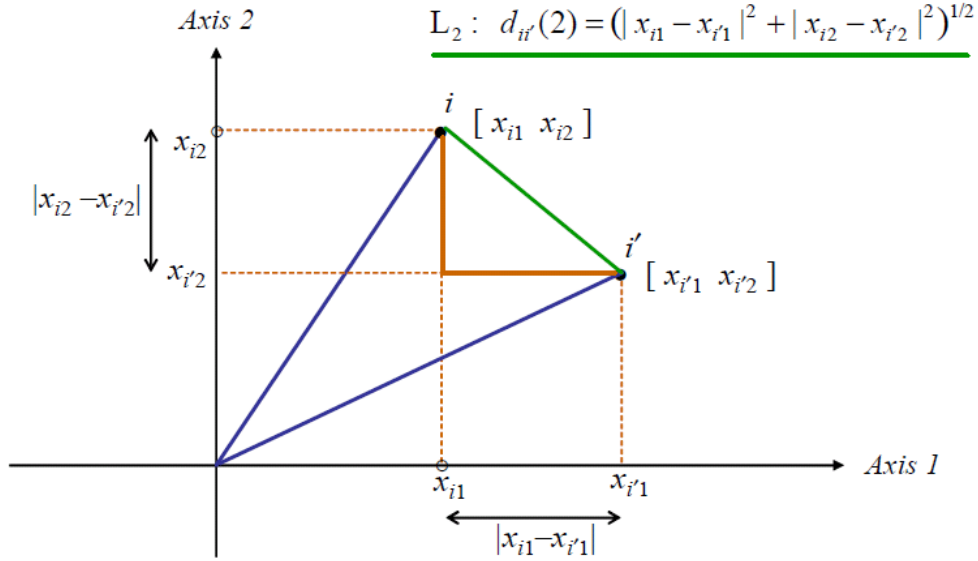
- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Proximity measures for numerical attributes

- Manhattan distance or City-block distance (L_1 norm)
  - $dist_1 = |p_1-q_1|+|p_2-q_2|+...+|p_d-q_d|$
  - The sum of the absolute differences of the *p,q* coordinates
- Euclidean distance (L_2 norm)
  - $dist_2 = ((p_1-q_1)^2+(p_2-q_2)^2+...+(p_d-q_d)^2)^{1/2}$
  - The length of the line segment connecting *p* and *q*
- Supremum distance (L_{max} norm or L_∞ norm)
  - $dist_\infty = \max\{|p_1-q_1|, |p_2-q_2|,...,|p_d-q_d|\}$
  - The max difference between any attributes of the objects.
- Minkowski Distance (Generalization of L_p-distance)
  - $dist_p = (|p_1-q_1|^p + |p_2-q_2|^p + ...+ |p_d-q_d|^p)^{1/p}$

$L_1: \; d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$

$L_2: \; d_{ii'}(2) = (|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2)^{1/2}$



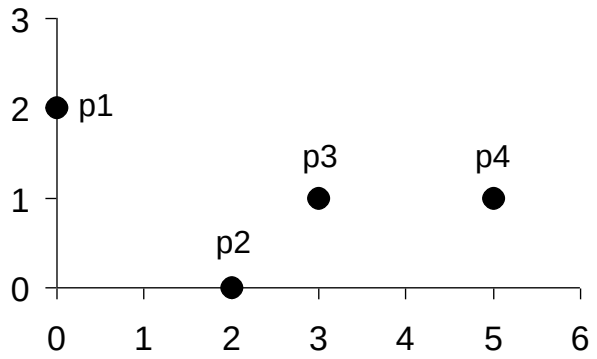
Source: http://www.econ.upf.edu/~michael/stanford/maeb5.pdf

Proximity measures for numerical attributes: examples

- Example

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

Point coordinates



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L1 distance matrix

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L2 distance matrix

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

L_∞ distance matrix

Normalization

- Attributes with large ranges outweigh ones with small ranges
  - e.g. *income* [10.000-100.000]; *age* [10-100]
- To balance the "contribution" of an attribute *A* in the resulting distance, the attributes are scaled to fall within a small, specified range.
- min-max normalization: Transform the feature from measured units to a new interval [new_min_A, new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A$$

  - v is the current feature value

Normalize *age* = 30 in the [0-1] range, given min_{age}=10, max_{age}=100

new_age=((30-10)/(100-10))*(1-0)+0=2/9

Normalization

- z-score normalization also called zero-mean normalization or standardization: Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one.
  - After zero-mean normalization, each feature will have a mean value of 0

$$v' = \frac{v - mean_A}{stand_dev_A}$$

    - where $mean_A$, $stand_dev_A$ are the mean and standard deviation of the feature

Normalize *income = 70,000* if $mean_{income}$=50,000, $stand_dev_{income}$ =15,000

new_value = (70,000-50,000)/15,000=1.33

Proximity measures for binary attributes 1/2

- A binary attribute has only two states: 0 (absence), 1 (presence)
- A contingency table for binary data

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	0

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

q = the number of attributes where i was 1 and j was 1
t = the number of attributes where i was 0 and j was 0

s = the number of attributes where i was 0 and j was 1
r = the number of attributes where i was 1 and j was 0

- Simple matching coefficient

  (for symmetric binary variables)

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

- for asymmetric binary variables:

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

- Jaccard coefficient

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

(for *asymmetric* binary variables)

Proximity measures for binary attributes 2/2

- Example:

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	0

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

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

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

(from previous slide)

q = the number of attributes where *i* was 1 and *j* was 1
t = the number of attributes where *i* was 0 and *j* was 0

s = the number of attributes where *i* was 0 and *j* was 1
r = the number of attributes where *i* was 1 and *j* was 0

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

Proximity measures for categorical attributes

- A nominal attribute has >2 states (generalization of a binary attribute)
  - e.g. *color* = {*red, blue, green*}
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

Name	Hair color	Occupation
Jack	Brown	Student
Mary	Blond	Student
Jim	Brown	Architect

- Method 2: Map it to binary variables
  - create a new binary attribute for each of the *M* nominal states of the attribute

Name	Brown hair	Blond hair	IsStudent	IsArchitect
Jack	1	0	1	0
Mary	0	1	1	0
Jim	1	0	0	1

Selecting the right proximity measure

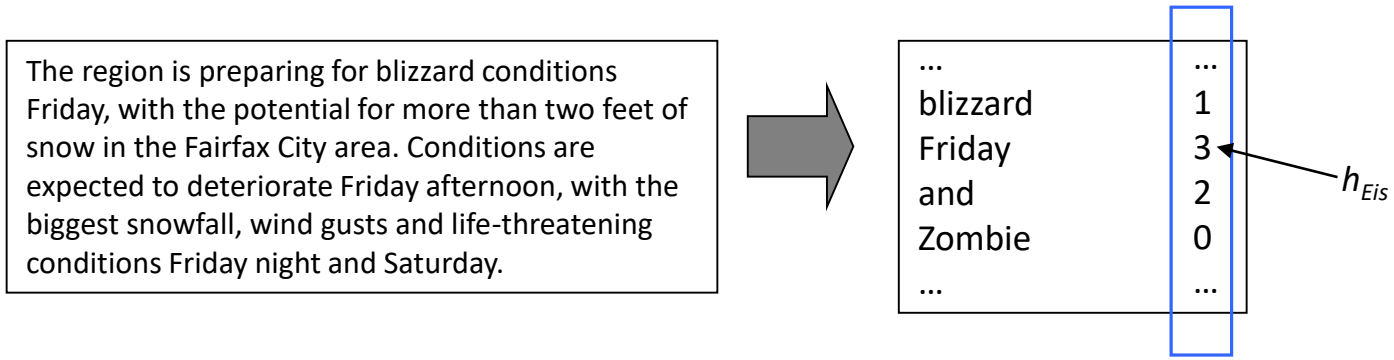
- The proximity function should fit the type of data
  - For dense continuous data, metric distance functions like Euclidean are often used.
  - For sparse data, typically measures that ignore 0-0 matches are employed
    - We care about characteristics that objects share, not about those that both lack
- Domain expertise is important, maybe there is already a state-of-the-art proximity function in a specific domain and we don't need to answer that question again.
- In general, choosing the right proximity measure can be a very time consuming task
- Other important aspects: How to combine proximities for heterogenous attributes (binary and numeric and nominal etc.)

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Feature transformations for text data 1/6

- Text represented as a set of terms ("Bag-Of-Words" model)
  - Terms:
    - Single words ("cluster", "analysis"..)
      or
    - bigrams, trigrams, …, *n*-grams ("cluster analysis",..)
  - Transformation of a document d in a vector $r(d) = (h_1, ..., h_d)$, $h_i \geq 0$: the frequency of term t_i in d



Feature transformations for text data 2/6

- Challenges/Problems in Text Mining:
  1. Common words (“e.g.”, “the”, “and”, “for”, “me”)
  2. Words with the same root (“fish”, “fisher”, “fishing”,…)
  3. Very high-dimensional space (dimensionality $d > 10.000$)
  4. Not all terms are equally important
  5. Most term frequencies $h_i = 0$ (“sparse feature space“)
- More challenges due to language:
  - Different words have same meaning (synonyms)
    - “freedom” – “liberty”
  - Words have more than one meanings
    - e.g. “java”, “mouse”

Feature transformations for text data 3/6

- Problem 1: Common words (“e.g.”, “the”, “and”, “for”, “me”)
  - Solution: ignore these terms (stopwords)

    There are stopwords list for all languages in WWW.

- Problem 2: Words with the same root (“fish”, “fisher”, “fishing”,…)
  - Solution: Stemming

    Map the words to their root

    - "fishing", "fished", "fish", and "fisher" to the root word, "fish"

    For English, the Porter stemmer is widely used.
    (Porters Stemming Algorithms: http://tartarus.org/~martin/PorterStemmer/index.html)

    Stemming solutions exist for other languages also.

    The root of the words is the output of stemming.

Feature transformations for text data 4/6

- Problem 3: Too many features/ terms (Very high-dimensional space)
  - Solution: Select the most important features ("Feature Selection")
  - Example: average document frequency for a term
    - Very frequent items appear in almost all documents
    - Very rare terms appear in only a few documents

Ranking procedure:

1. Compute document frequency for all terms t_i :
2. Sort terms w.r.t. $DF(t_i)$ and get $rank(t_i)$
3. Sort terms by $score(t_i) = DF(t_i) \cdot \text{rank}(t_i)$
   e.g. $score(t_{23}) = 0.82 \cdot 1 = 0.82$
   $score(t_{17}) = 0.65 \cdot 2 = 1.3$
4. Select the k terms with the largest $score(t_i)$

$$DF(t_i) = \frac{\#Docs\ containing\ t_i}{\#All\ documents}$$

Rank	Term	DF
1.	t_{23}	*0.82*
2.	t_{17}	*0.65*
3.	t_{14}	*0.52*
4.	…	…

Feature transformations for text data 5/6

- Problem 4: Not all terms are equally important
  - Idea: Very frequent terms are less informative than less frequent words. Define such a term weighting schema.
  - Solution: TF-IDF (Term Frequency · Inverse Document Frequency)

    Consider both the importance of the term in the document and in the whole collection of documents.

$$TF(t,d) = \frac{n(t,d)}{\sum_{w \in d} n(w,d)}$$

The frequency of term *t* in *d*

$$IDF(t) = \log\left(\frac{|DB|}{|\{d \mid d \in DB \wedge t \in d\}|}\right)$$

Inverse frequency of term t in all DB

$$TF \times IDF = TF(t,d)\,IDF(t)$$

Feature vector with TF-IDF : $r(d) = (TF(t_1,d) \cdot IDF(t_1), \ldots, TF(t_n,d) \cdot IDF(t_n))$

Feature transformations for text data 6/6

- Problem 5: for most of the terms $h_i = 0$
  - Euclidean distance is not a good idea: it is influenced by vectors lengths
  - Idea: use more appropriate distance measures

Jaccard Coefficient: Ignore terms absent in both documents

$$d_{Jaccard}(d_1, d_2) = 1 - \frac{|d_1 \cap d_2|}{|d_1 \cup d_2|} = \frac{|\{t|t \in d_1 \wedge t \in d_2\}|}{|\{t|t \in d_1 \vee t \in d_2\}|}$$

Cosine Coefficient: Consider term values (e.g. TFIDF values)

$$d_{cosinus}(d_1, d_2) = 1 - \frac{\langle d_1, d_2 \rangle}{\|d_1\| \cdot \|d_2\|} = 1 - \frac{\sum_{i=0}^{n} (d_{1,i} \cdot d_{2,i})}{\sqrt{\sum_{i=0}^{n} d_{1,i}^2} \cdot \sqrt{\sum_{i=0}^{n} d_{2,i}^2}}$$

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Data Visualization

A variety of techniques

- Parallel Coordinates
- Quantile Plot (✓)
- Scatter Plot Matrix (✓)
- Loess Curve
- Spiderweb model
- Chernoff faces

Parallel Coordinates

Method to visualize high dimensional data sets in parallel axes of feature spaces

Abstract Idea: Represent a *d*-dimensional feature space with a system of *d*-parallel axes

Instance Representation: a polygonal line crossing each parallel axis to its corresponding feature value of the instance

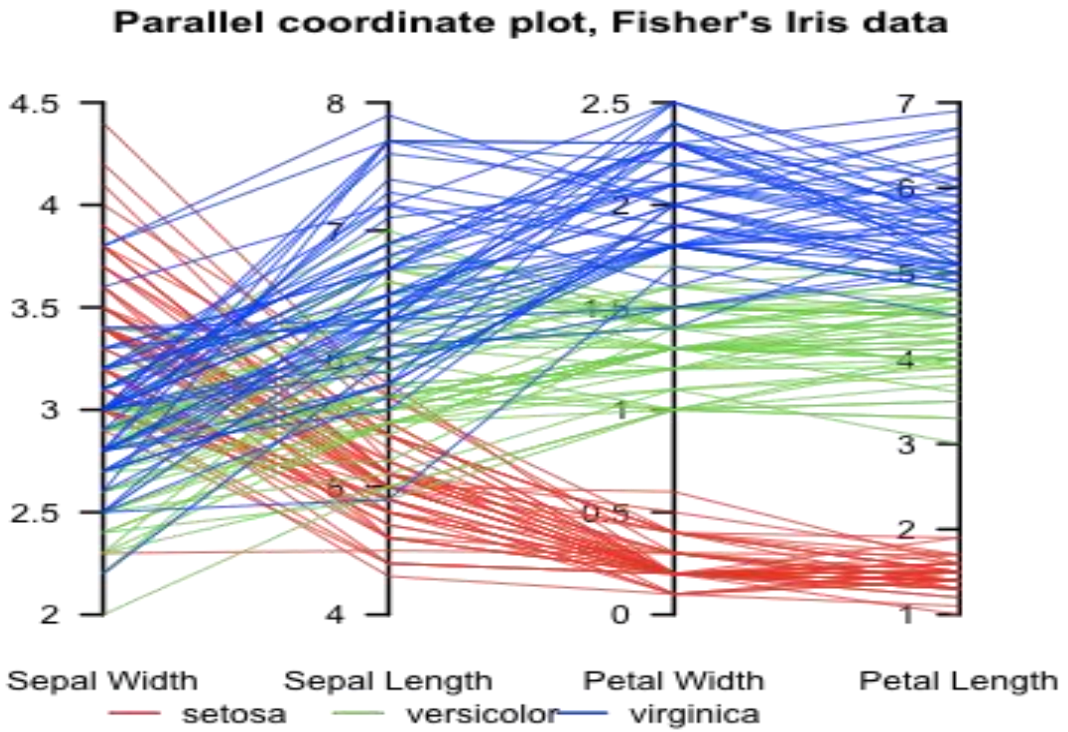


Image source: https://en.wikipedia.org/wiki/Parallel_coordinates

Slide adopted from http://www.dbs.ifi.lmu.de/Lehre/KDD/SS16/skript/2_DataRepresentation.pdf

Spiderweb Model

Method to visualize multivariate data sets using a spider net where each *radii* of net represent one feature of the data set

Abstract Idea: "Spider chart or radar chart is a visualization technique for multivariate data in the form of two-dimensional chart represented on axes,radii, starting from the same point"

** Notation:

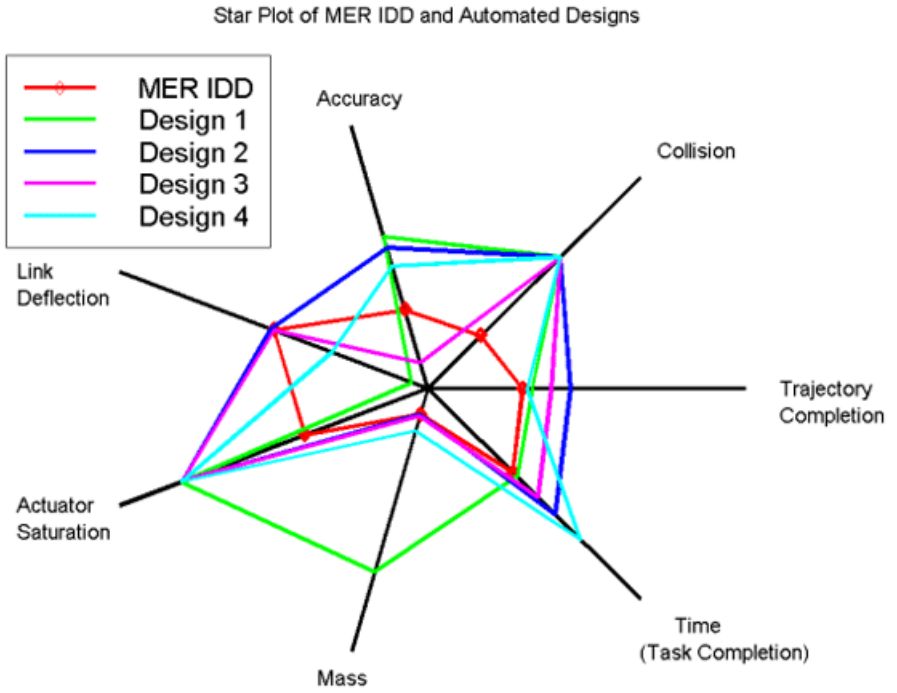
radii= equi-angular spokes ≈ spider-net axes

Instance Representation:

Polyline that intersects each spider-net axis

Motivation: Sets all instances to a common origin

But does not help if number of instances is big (Big Data)



Example spiderweb visualization from NASA.
The most desirable design results can be found in the center of the net.
Image source: https://en.wikipedia.org/wiki/Radar_chart

Slide after https://en.wikipedia.org/wiki/Radar_chart

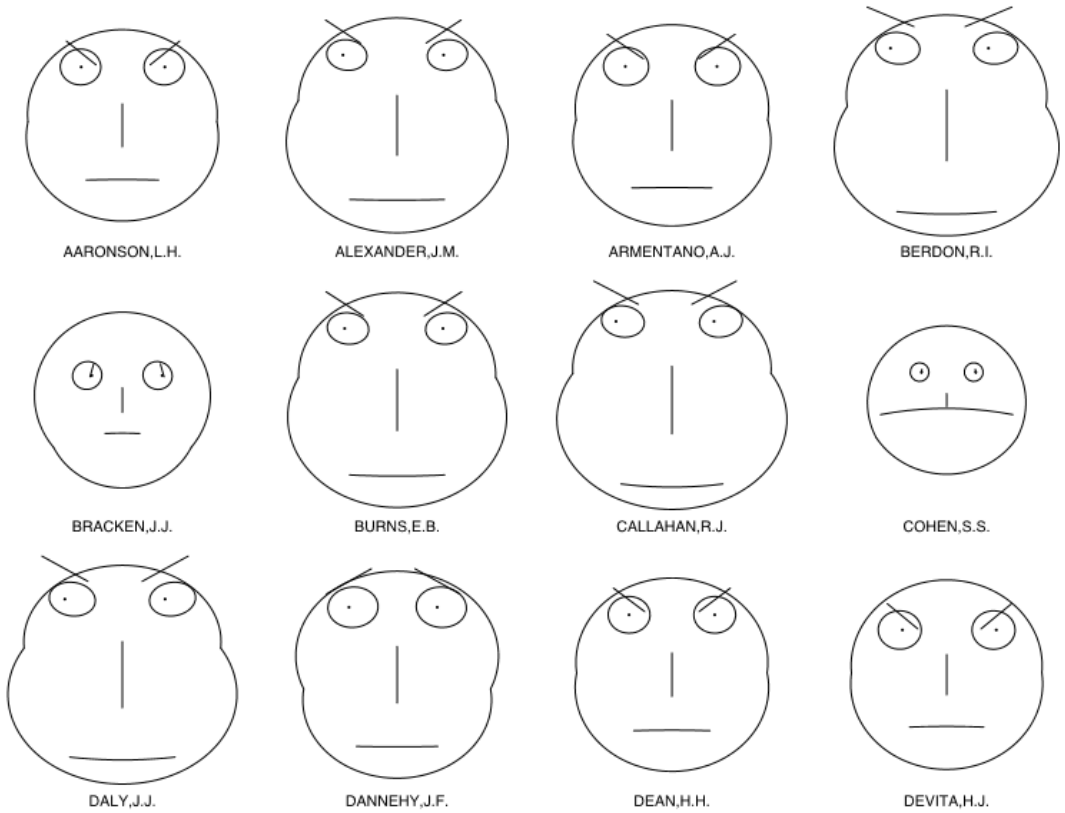
Chernoff Faces

Visualize multivariate data
in the shape of human face

Motivation: Humans can easily perceive faces
and notice small variations on them

Method: Each individual parts of face,
e.g. *eyes*, *nose*,.. represent one feature
and the shape the corresponding instance's value

But applicable with at a certain number of
feature dimensions (dimensional reduction)

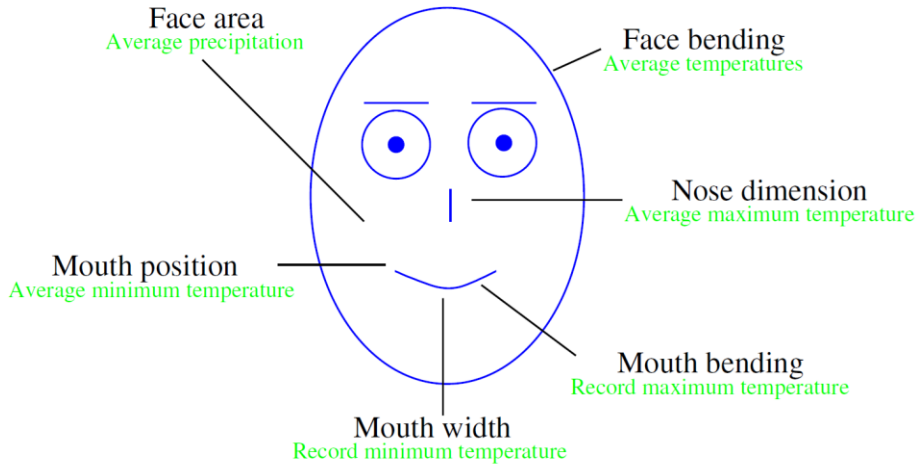


Chernoff faces for laywers' ratings for twelve judges.
Image source: https://en.wikipedia.org/wiki/Chernoff_face

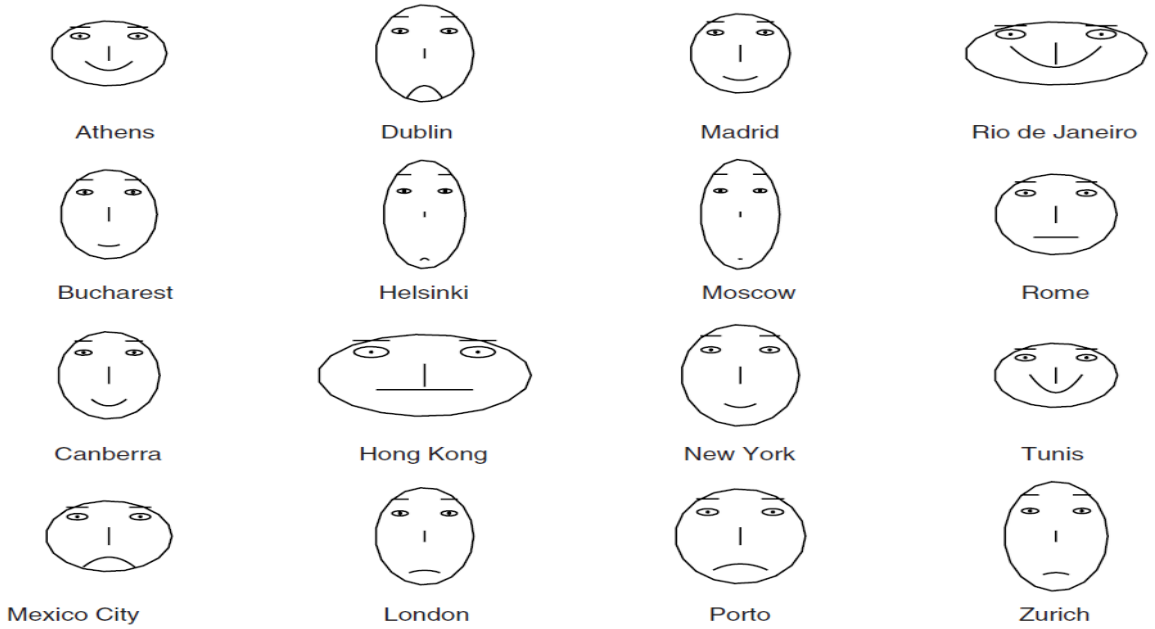
Slide after https://en.wikipedia.org/wiki/Chernoff_face

[Chef 85] Chernoff, H: The use of faces to represent points in k-dimensional space graphically. Journal of American Statistical Association, Vol. 68, pp. 361-368, 1973.

Chernoff Faces



Example of mapping of face parts to climate features.



Using the left mapping, Climatic data of some cities represented by Chernoff faces.

Fig.4.11(left), Fig.412 (right) of R. Maza “Introduction to Information Visualization” Springer 2009.

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Proximity (similarity, distance) measures
- ~~Feature transformation for text data~~
- ~~Data Visualization~~
- Homework/ Tutorial
- Things you should know from this lecture

Homework/ tutorial

- 2nd tutorial follows next week
  - No tutorials on Monday, but please come on Tuesday (it might be more crowded)
- Homework
  - Investigate a dataset (e.g., the iris dataset) using Python. What can you see?
- Readings:
  - Tan P.-N., Steinbach M., Kumar V book, Chapter 2.
  - Zaki and Meira book, Chapter 1
  - Han J., Kamber M. Pei J. *Data Mining: Concepts and Techniques 3rd ed., Morgan Kaufmann, 2011* (Chapter 2)

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data
- Data Visualization
- Homework/ Tutorial
- Things you should know from this lecture

Things you should know from this lecture

- Basics of data preprocessing
- Basic feature types
- Proximity measures

Acknowledgement

- The slides are based on
  - KDD I lecture at LMU Munich (Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Eirini Ntoutsi, Jörg Sander, Matthias Schubert, Arthur Zimek, Andreas Züfle)
  - Introduction to Data Mining book slides at http://www-users.cs.umn.edu/~kumar/dmbook/
  - Pedro Domingos Machine Lecture course slides at the University of Washington
  - Machine Learning book by T. Mitchel slides at http://www.cs.cmu.edu/~tom/mlbook-chapter-slides.html
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran