Lecture: Machine Learning for Data Science

Winter semester 2021/22

Lecture 2: Getting to know your data

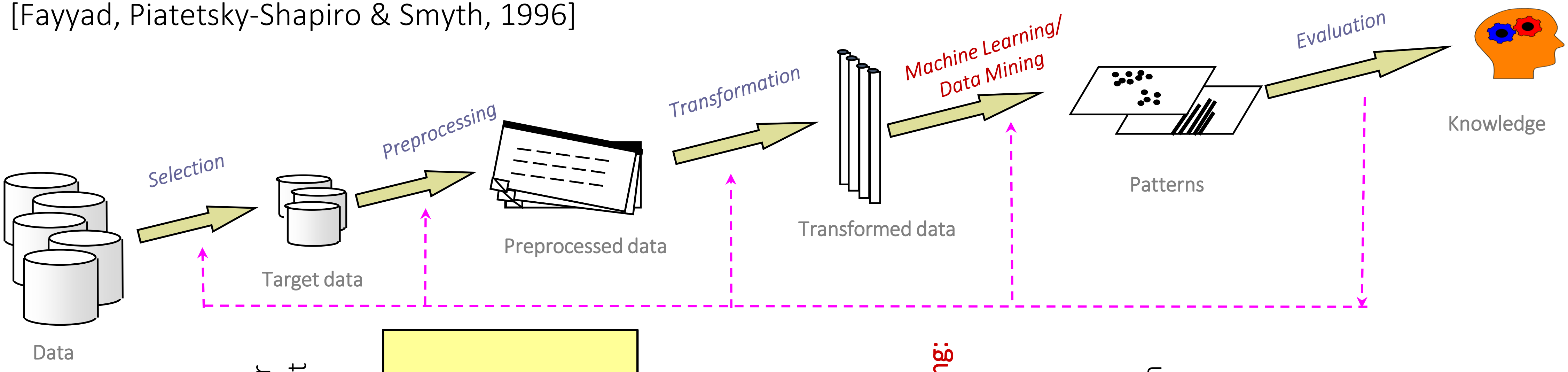
Prof. Dr. Eirini Ntoutsi

Outline

- Data preprocessing and data transformation
- Features
- Basic data descriptors
- Feature space and Distance function
- Things you should know from this lecture & reading material

The KDD process

[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

Machine Learning/ 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 = 200K when the rest dataset lies in [30K-50K]
    - Irrelevant information
  - 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 [1-10]
- “Dirty” data → poor learning
- Data preprocessing is necessary for improving the quality of learning!

“Garbage in, garbage out”

Your analysis is as good as your data.

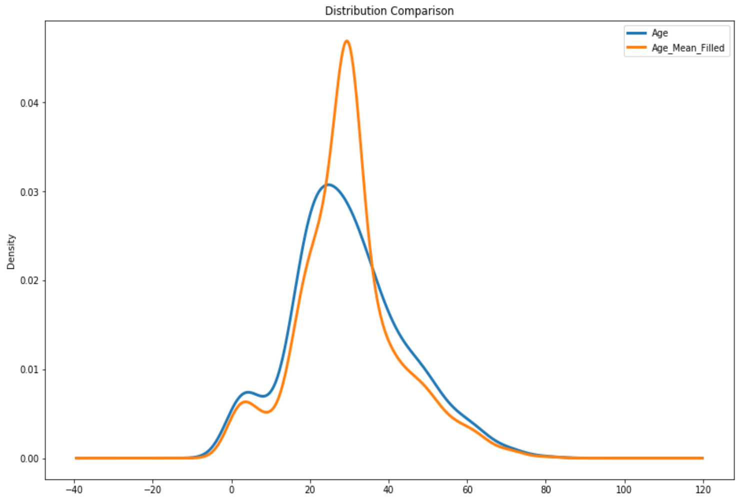
Typical tasks in data preprocessing

- Data integration:
  - Integration of multiple databases, data warehouses, or files (entity identification, value resolution)
- Data cleaning:
  - Fill in missing values
  - Smooth noisy data
  - Identify or remove outliers
  - Resolve inconsistencies
- Data reduction:
  - Duplicate elimination

There exist dedicated lectures on these topics. Also, nowadays many of these tasks rely on AI/ML

Mind the preprocessing decisions/assumptions

- Many of the preprocessing operations do actually change the data → Beware of side effects
- An example on the effect of mean imputation (replacing missing values with average feature values)



Source: https://towardsdatascience.com/stop-using-mean-to-fill-missing-data-678c0d396e22

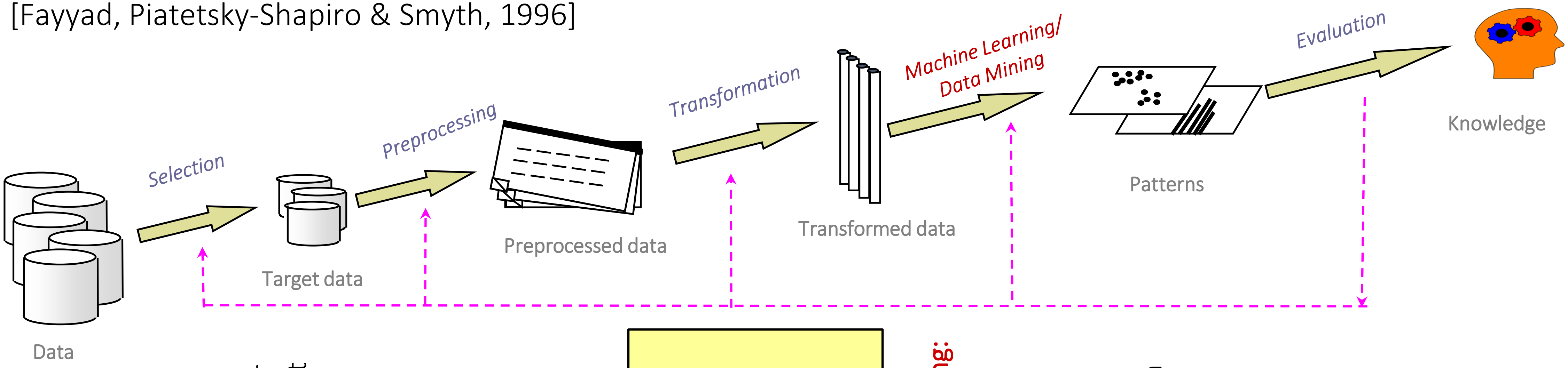
- Libraries/Tools we use might make such decisions for us
  - e.g., in case of algorithms not able to cope with missing values, non-numerical features, multi-class problems,…

Mind the modeling assumptions

- E.g., modeling gender as a binary variable {Male, Female} might lead to discrimination against non-binary people
  - "*Computers are binary, people are not: how AI systems undermine LGBTQ identity*"
- E.g., modeling race as {white, non-white} might lead to race discrimination
  - There are more race categories

The KDD process

[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

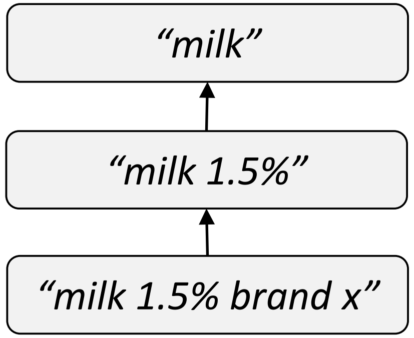
Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Machine Learning/ Data Mining:
- Search for patterns of interest

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

Typical tasks in data transformation

- Transformation
  - Normalization in a given range, e.g., [0-1]
  - Generalization through some concept hierarchy
  - Discretization (convert continuous data into discrete ones)



- Data reduction:
  - Aggregation, e.g., from 12 monthly salaries to average salary per month.
  - Feature selection
  - Dimensionality reduction, through e.g., PCA.

Outline

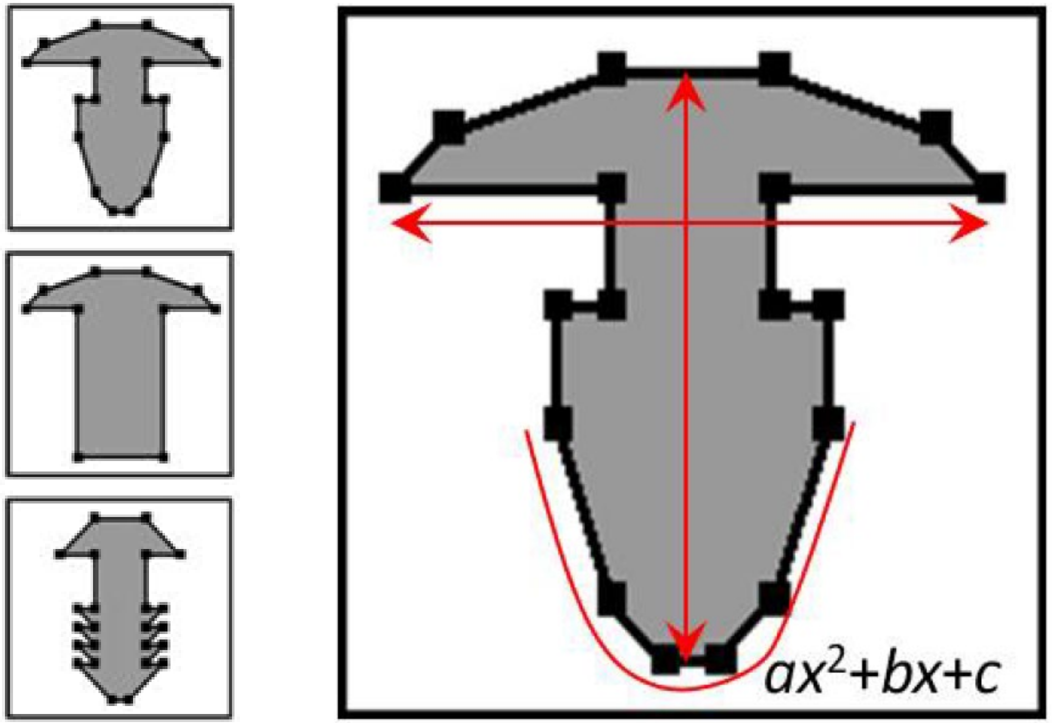
- Data preprocessing and data transformation
- Features
- Basic data descriptors
- Feature space and Distance function
- Things you should know from this lecture & reading material

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

Deriving features from complex objects

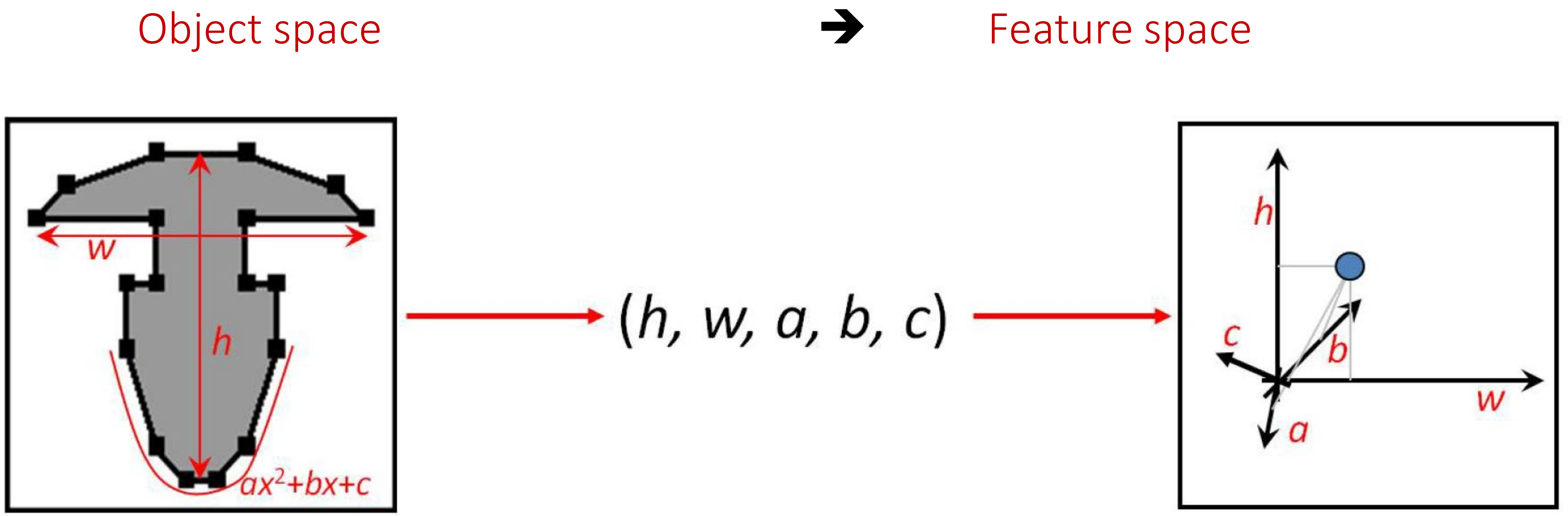
- In many cases, we are not given a feature description of the data, so we have to extract the features
- Example: CAD objects
- Possible features
  - Width
  - Height
  - Curvature parameters *(a,b,c)*



Slide: from Arthur Zimek

Deriving features from complex objects

- Transformation



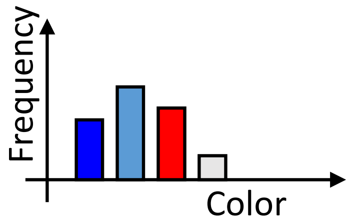
- Features are combined to feature vectors
- Often high-dim feature spaces (here only 5-*d*)
- Statistical context: features are called variables

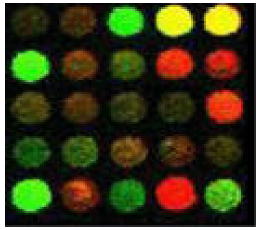
Slide: from Arthur Zimek

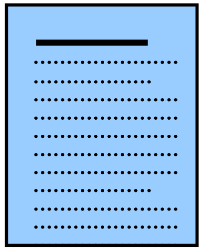
Deriving features from complex objects

- Feature extraction depends on the application
- Images
  - E.g., color histograms (the distribution of colors, e.g., in the RGB space, over the pixels of an image)
- Gene databases
  - E.g., gene expression levels
- Text databases
  - E.g., word counts

- ML methods work on the given/extracted feature representation thereafter
  - The extraction of meaningful features is very important







Machine	25
Learning	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.

Feature extraction for text data

- Text can be represented as a set of terms (Bag-Of-Words (Bow) model)
  - Terms can be:
    - Unigrams (“cluster“, “analysis“..)
    - Bigrams (“cluster analysis“, “Angela Merkel”, ...)
    - *n*-grams
- Typical feature extraction from text: transform a text/document *d* into a vector of term frequencies

$$d \mapsto (f_{t_1 d}, f_{t_2 d}, \ldots, f_{t_n d})$$

  - Where $f_{t_i d}$ is the frequency of term t_i in document *d*.

The region is preparing for blizzard conditions Friday, with the potential for more than two feet of snow in the Fairfax City area. Conditions are expected to deteriorate Friday afternoon, with the biggest snowfall, wind gusts and life-threatening conditions Friday night and Saturday.

…	…
blizzard	1
Friday	3
and	2
Zombie	0
…	…

Feature extraction for text data

- Challenges/Problems for learning:
  - Common words (“e.g.”, “the”, “and”, “for”, “me”)
  - Words with the same root (“fish”, “fisher”, “fishing”,…)
  - Very high-dimensional space (dimensionality d > 10.000)
  - Not all terms are equally important
  - 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 extraction for text data

- Problem 1: Common words (“e.g.”, “the”, “and”, “for”, “me”)
  - Solution: ignore these terms (stopword removal)
    - There are stopwords list available for all (?) languages
- Problem 2: Words with the same root (“fish”, “fisher”, “fishing”,…)
  - Solution: Reduction → 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)
      - The root of the words is the output of stemming.

Feature extraction for text data

- Problem 3: Too many features/ terms (Very high-dimensional space)
  - Solution: Feature Selection (select the most important features)
    - Find document frequency for all terms

$$DF(t_i) = \frac{|\{d|t_i \in d\}|}{|\mathcal{D}|}$$

    - Sort terms according to $DF(t_i)$

Rank	Term	DF
1.	t_{23}	0.82
2.	t_{17}	0.65
3.	t_{14}	0.52
4.	$\ldots$	$\ldots$

    - Sort terms according to $score(t_i)=DF(t_i)*rank(t_i)$

$$\text{score}(t_{23}) = 0.82 \cdot 1$$
$$\text{score}(t_{17}) = 0.65 \cdot 2$$

    - Choose the k terms with the largest scores

Feature extraction for text data

- Problem 4: Not all terms are equally important
  - Solution: TF-IDF (Term Frequency · Inverse Document Frequency)
    - Consider both the importance of a term *d* in the document (*TF*) and in the whole collection of documents (*IDF*).
    - Higher weights for rare words
    - Higher weights for terms that are more frequent than others in some document
  - *TF* is the relative term frequency in some document *d*:

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

  - *IDF* is the the inverse document frequency of *t* for all documents *D*:

$$IDF(t) = \frac{|\mathcal{D}|}{|\{d \mid d \in \mathcal{D} \wedge t \in d\}|}$$

  - Feature vector for document *d*:

$$d = \begin{pmatrix} TF(t_1,d) \cdot IDF(t_1) \\ TF(t_2,d) \cdot IDF(t_2) \\ \vdots \\ TF(t_k,d) \cdot IDF(t_k) \end{pmatrix}$$

Feature extraction for text data and beyond

- Many ways to extract information from text data nowadays
  - The old-fashioned Bag-of-Words with TF-IDF
  - Word embeddings (like Word2Vec)
  - Language models (like BERT)
- Dedicated field: NLP (Natural Language Processing)
- Likewise for other applications
  - Images → Computer Vision field
  - …

Data matrix

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

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- The number of instances n is referred to as the size or cardinality of the dataset, *n=lDl*
- 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 of the data $D_{F'}$

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	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): discrete values
  - Binary variables
  - Nominal variables
  - Ordinal variables
- Numerical variables (quantitative): values can be discrete or continuous
  - Interval-scale variables
  - Ratio-scaled variables

Binary/ Dichotomous variables

- The attribute can take only 2 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?

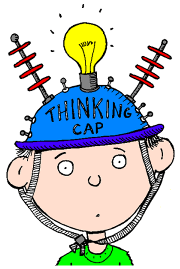
ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Categorical: Nominal variables

- The attribute can take values within a set of *M* categories/ states (binary variables are a special case)
  - No ordering (better, more, …) 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 (better/worse, more/less, …) 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*

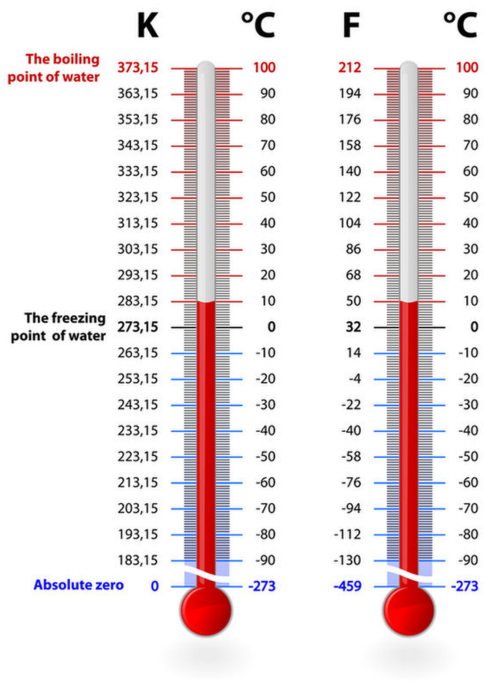
Operations that can be applied: $=, \neq, <, >$

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

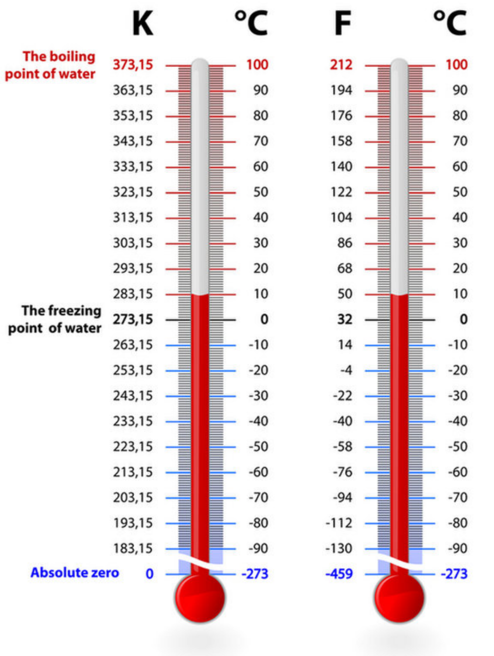
Numerical features: Interval-scale variables

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



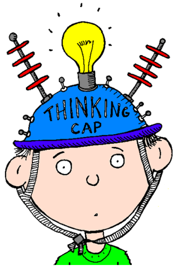
Operations that can be applied: $=, \neq, <, >, +, -$

Numerical features: 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	O	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	$=, \neq$
Ordinal	$=, \neq, <, >$
Interval[b]	$=, \neq, <, >, +, -$
Ratio	$=, \neq, <, >, +, -, \times, \div$

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

Outline

- Data preprocessing and data transformation
- Features
- Basic data descriptors
- Feature space and Distance function
- Things you should know from this lecture & reading material

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

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Univariate descriptors: measures of central tendency

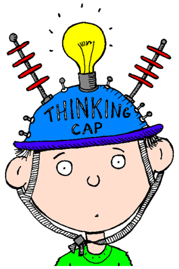
- For a numerical feature X we have a sample $x_1,\ldots,x_n$ (i.e., the dataset projected w.rt. X)
- Measures of central tendency of X include:
  - (Arithmetic) mean/ center/ average:
    - We use the notation x-bar

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

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

  - Weighted average:

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



What is the mean of:

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

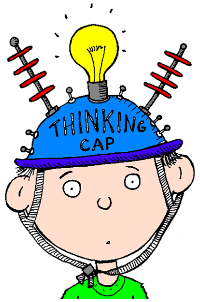
Univariate descriptors: measures of central tendency

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

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

- (For at least ordinal variables) Median: the central element in ascending ordering
  - Middle value if odd number of values, or average of the middle two values otherwise.

$$\mathbf{D} = \left(\begin{array}{c|cccc} & \boxed{X_1} & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$



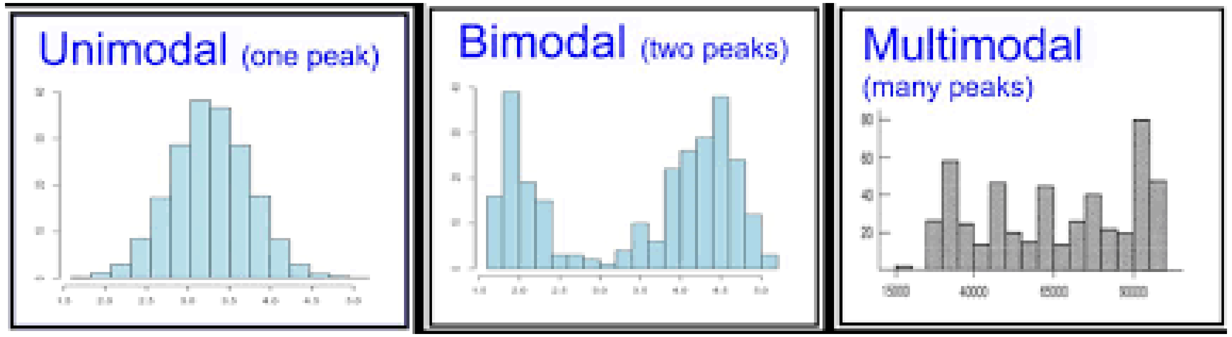
What is the median of:

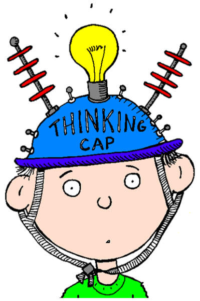
3, 8, 3, 4, 3, 6, 4, 2, 3

Univariate descriptors: measures of central tendency

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

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$



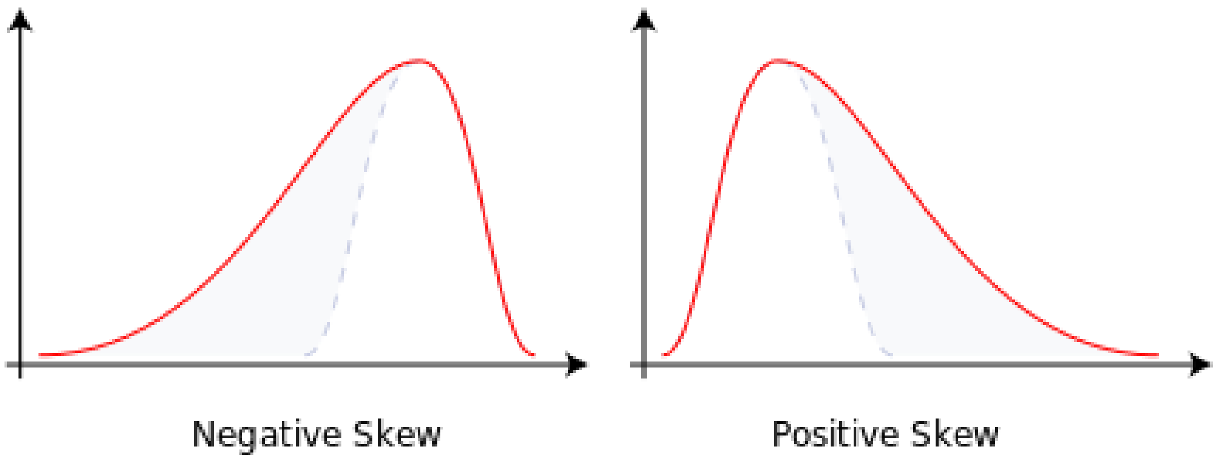


What is the mode of:

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

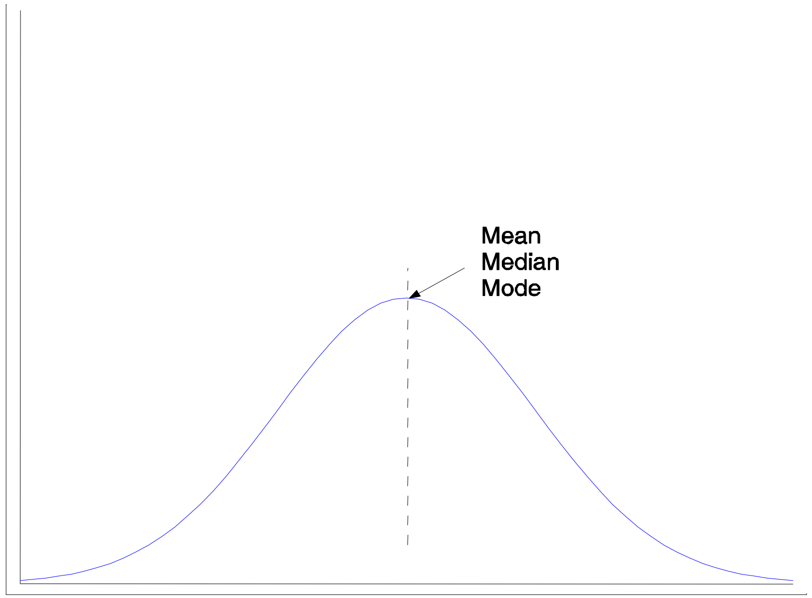
Univariate descriptors: measures of central tendency

- Skewness: a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
  - Symmetric
  - Positively skewed (or, skewed to the right)
  - Negatively skewed (or, skewed to the left)

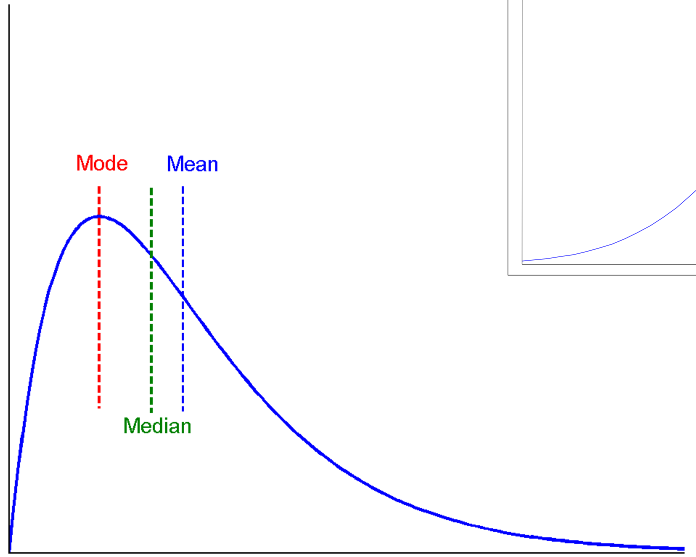


Univariate descriptors: measures of central tendency

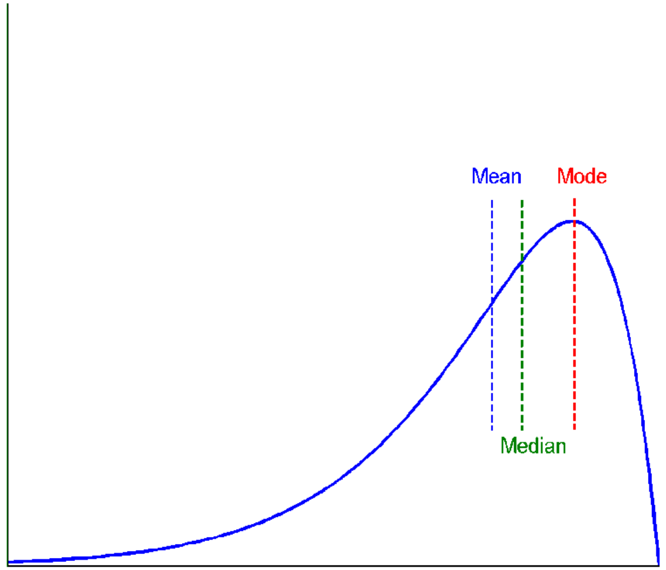
- Mean, median and mode in normal vs highly-skewed distributions



Symmetric



Positively skewed (or, skewed to the right)



Negatively skewed (or, skewed to the left)

Univariate descriptors: measures of spread

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

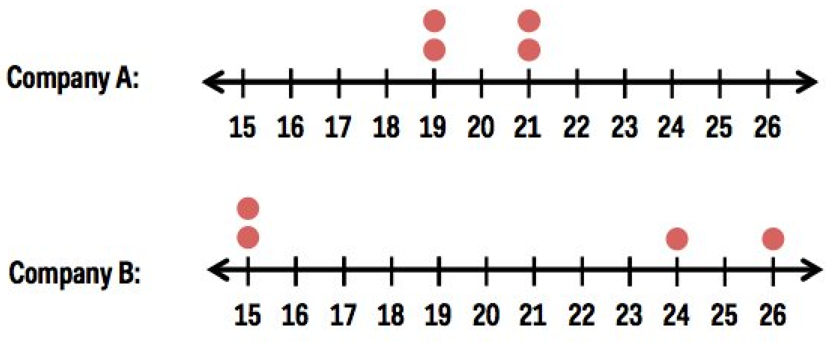
- For a feature X we have a sample $x_1,\ldots,x_n$ (i.e., the dataset projected w.rt. X)
- The degree to which X values tend to spread is called dispersion or variance of X and is denoted by σ^2:

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

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

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

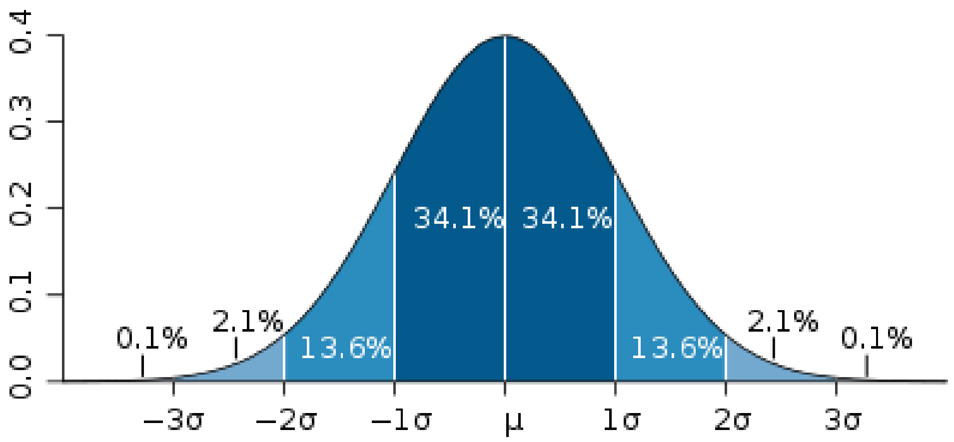
Same mean (20), different spread



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: the 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

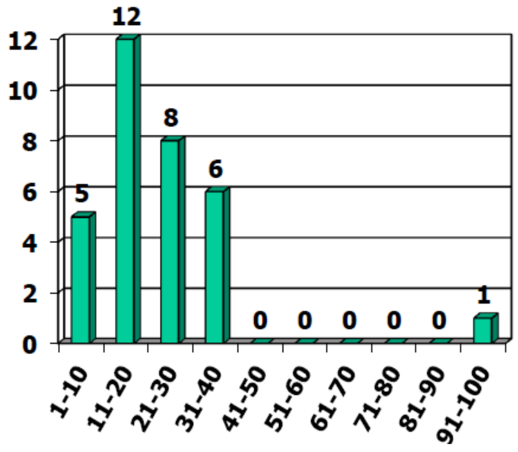


Source: http://en.wikipedia.org/wiki/Normal_distribution

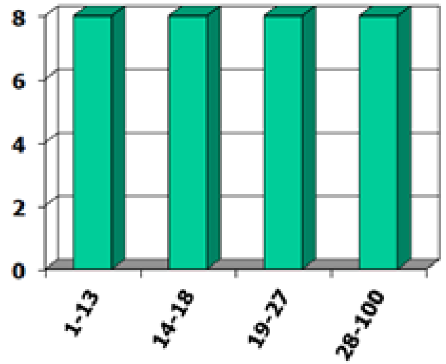
Univariate descriptors: useful charts

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

- For visual inspection of an attribute *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*, *h(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 sam



Equal width histogram

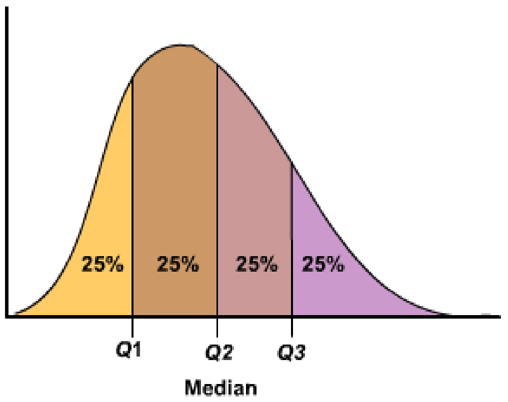


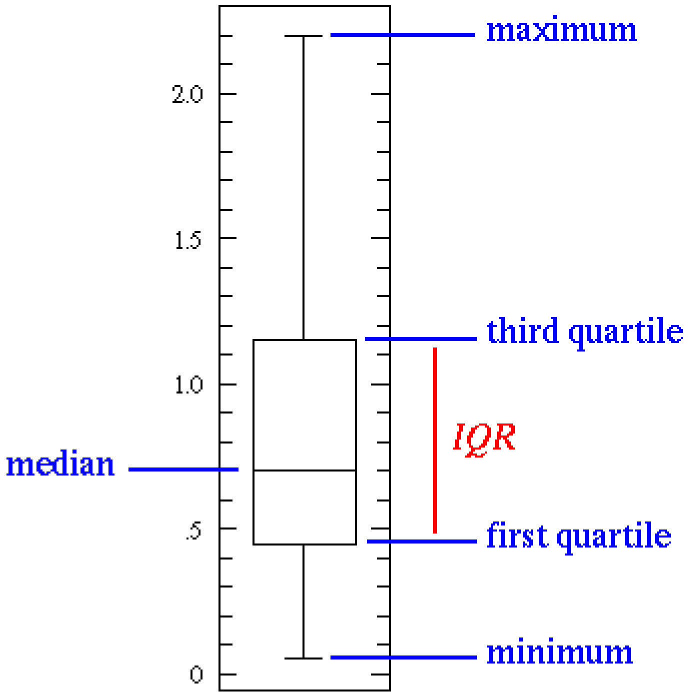
Equal depth histogram

Univariate descriptors: useful charts

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

- Boxplots: a standardized way of displaying the distribution of data based on a 5 number summary:
  - min, Q1, median, Q3, max
    - Q1 (25th percentile): 25% of the data follow below this percentile
    - Median (50th percentile): 50% of the data follow below this percentile
    - Q3 (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





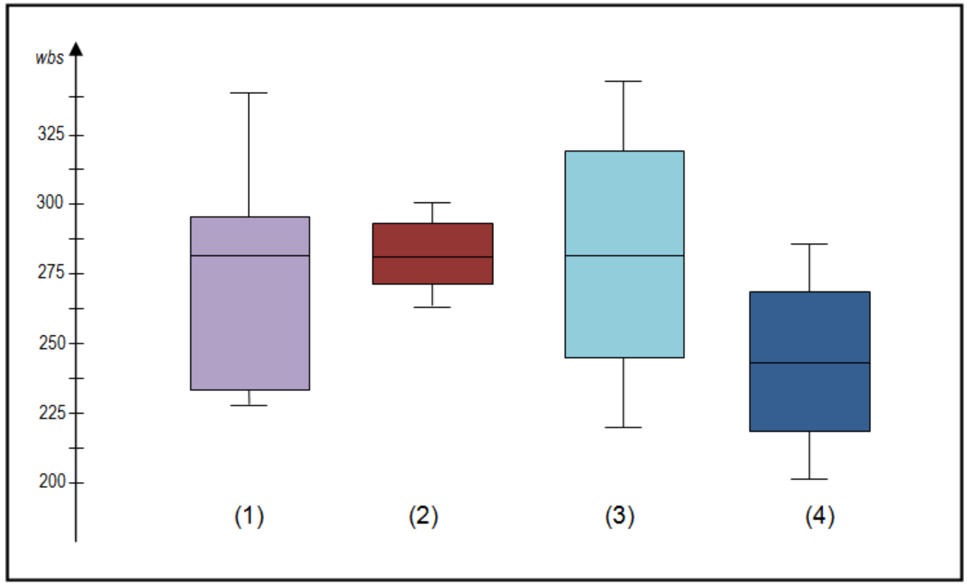
Source: http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

Univariate descriptors: Boxplot example

- Sample: 27, 2, 5, 19, 7, 9, 12, 6, 15, 18, 1.
- How to compute the boxplot? (Recall a boxplot is a 5 number summary: min, Q1, median, Q3, max)
- Order the data from smallest to largest 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Find the median → Q2=9
- Find the quartiles
  - Q1 is the median of the data points to the left of the median→ Q1=5
  - Q3 is the median of the data points to the right of the median → Q3=18
- Find min (min=1) and max (max=27)

Univariate descriptors: useful charts

- 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)

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

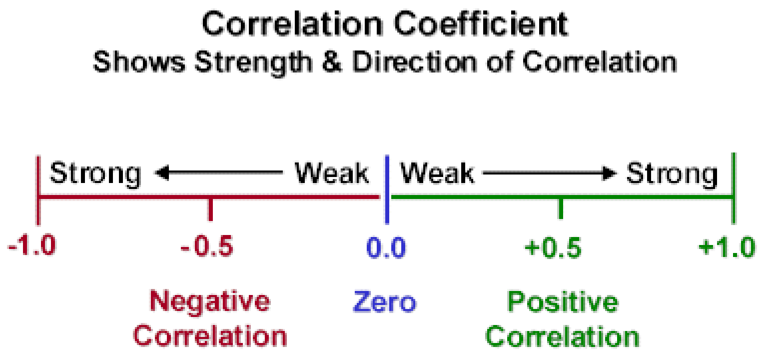
Bivariate descriptors: for numerical features

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

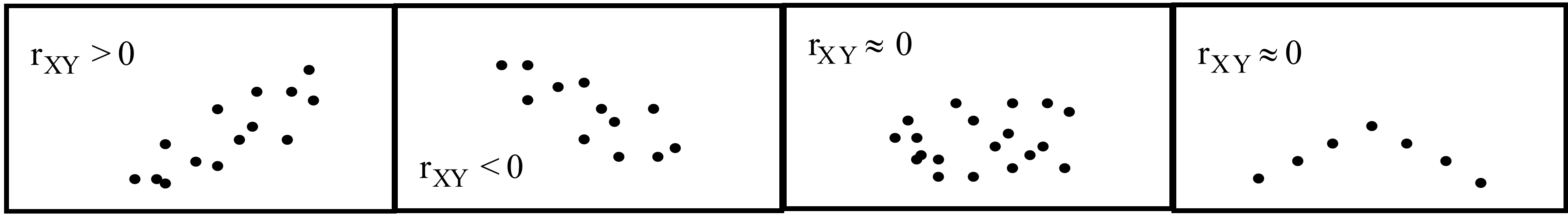
- Correlation coefficient (also called Pearson's correlation coefficient) measures the linear association between features X, Y:

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

  - x_i, y_i: the values in the i^{th} 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+Coefficient



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

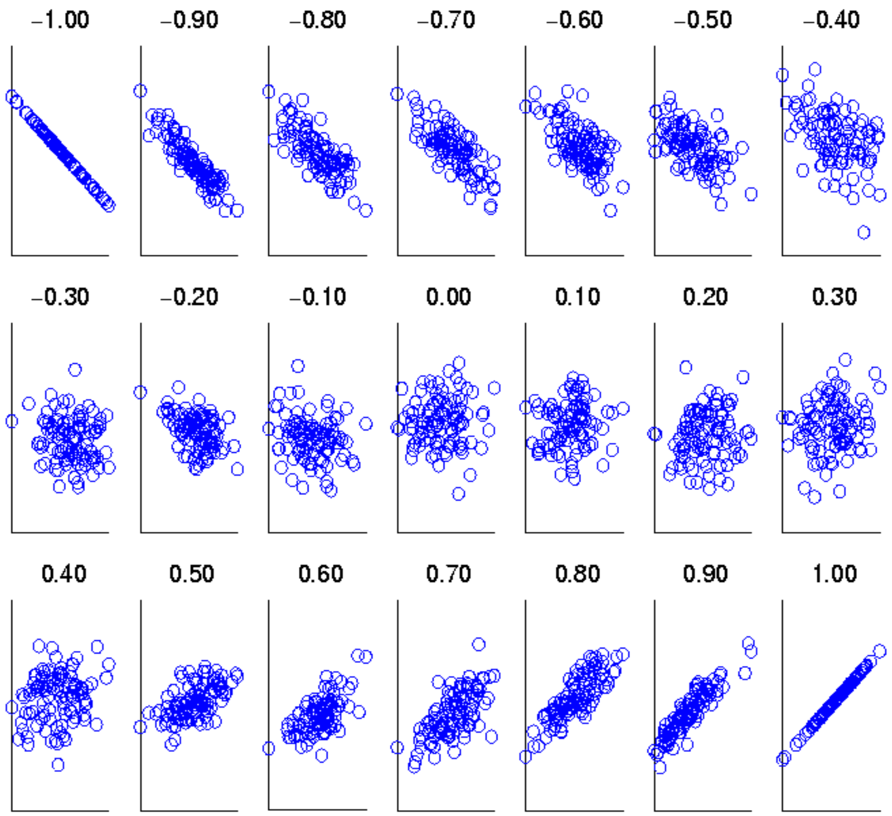


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

Bivariate descriptors: for categorical features

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

- 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 marginal frequencies h_i, h_j of X, Y.

Attribute X		Attribute Y		
		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: Chi-square example

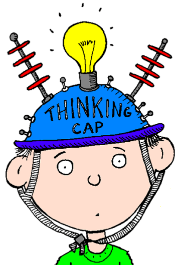
$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

- 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: Chi-square example

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- 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: Chi-square example

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

- 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

- χ^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

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

- Based on your desired confidence level (e.g., 95% → p = 0.05)
- Based on the degrees of freedom
  - (r-1)(c-1) degrees of freedom, where r represents the number of rows in the two-way table and c represents the number of columns.
- Check if your value is significant or non-significant

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: Chi-square example

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

- 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

- χ^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 and data transformation
- Features
- Basic data descriptors
- Feature space and Distance function
- Things you should know from this lecture & reading material

Feature spaces and distance functions

A feature space is a domain with a distance function

F = (dom, dist)

- *dom* is a sorted set of features

- $\mathrm{dist} : \mathrm{dom} \times \mathrm{dom} \rightarrow \mathbb{R}_0^+$ is a distance function

with the following properties

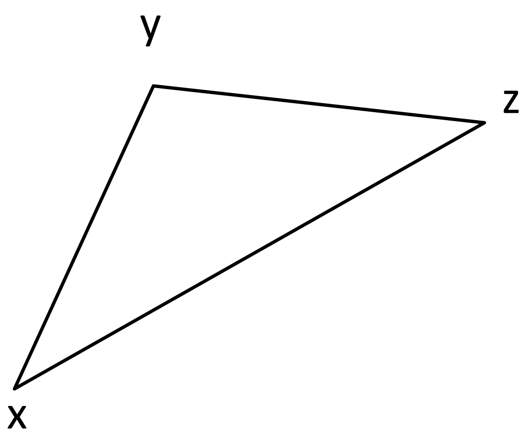
- Strickness: $\forall p, q \in \mathrm{dom}, p \neq q : \mathrm{dist}(p,q) > 0$
- Reflexivity: $\forall o \in \mathrm{dom} : \mathrm{dist}(o,o) = 0$
- Symmetry: $\forall p, q \in \mathrm{dom} : \mathrm{dist}(p,q) = \mathrm{dist}(q,p)$

Metric space

- *M = (dom, dist)* is a metric space if, the following properties hold
  - *M* is a feature space
  - The triangle inequality holds

$$\forall o, p, q \in \text{dom} : \text{dist}(o, p) \leq \text{dist}(o, q) + \text{dist}(q, p)$$

- Most common example: Euclidean vector space



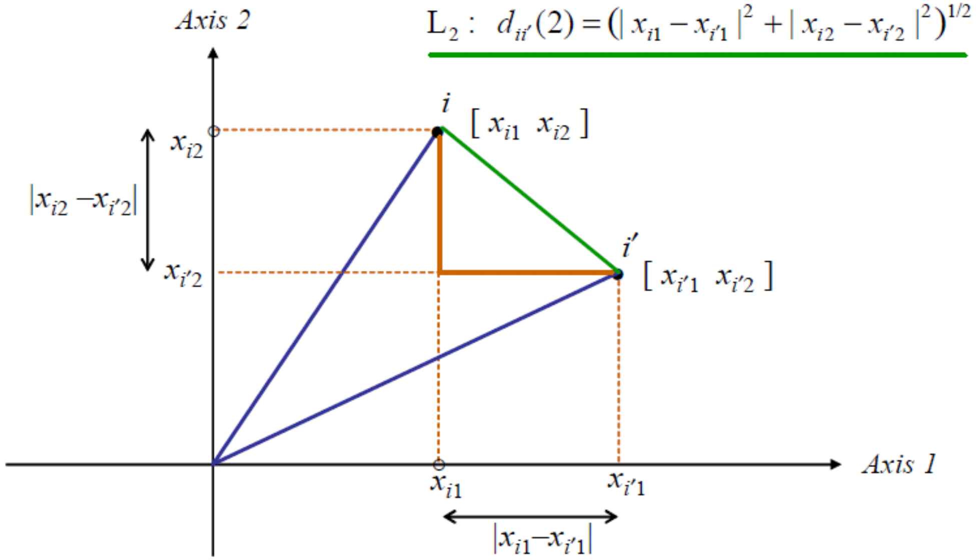
Common distance measure for (Euclidean) feature vectors

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

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

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

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



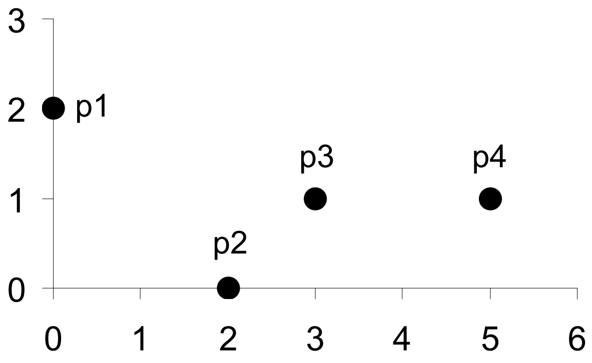
Source: http://www.econ.upf.edu/~michael/stanford/maeb5.pdf

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

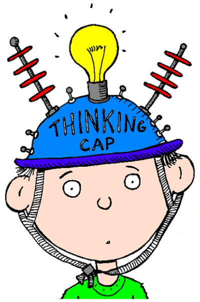
L_∞ distance matrix

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_minA* , *new_maxA*]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - v is the current feature value



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

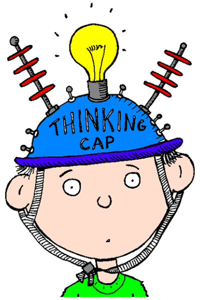
new_age=((30-10)/(100-10))*(1-0)+0=2/9

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

- 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

Instance i		1	0	sum
	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
- (for asymmetric binary variables)

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

Proximity measures for binary attributes: example

- 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 (nominal) 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 and data transformation
- Features
- Basic data descriptors
- Feature space and Distance function
- Things you should know from this lecture & reading material

Overview and Reading

- Overview
  - Data: instances & features
  - Feature types
  - Basic descriptors
  - Feature spaces and proximity measures

- Reading
  - Part 1: Data Analysis Foundations from the book by Meira and Zaki

Thank you

Questions/Feedback/Wishes?

Acknowledgements

- 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
  - Arthur Zimek DMLML lecture at SDU.
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran