Data Analytics Foundations

Chapter 1 – Hands on Data Analytics for Everyone

September 2, 2022

北京师范大学-香港浸会大学联合国际学院
United International College

Contents

What is Data Science?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights** from structured and unstructured data.

[Wikipedia quoting Dhar 13, Leek 13]

Knowledge discovery in databases (*KDD*) is the process of (semi-)automatic **extraction of knowledge** from databases which is *valid*, *previously unknown*, and *potentially useful.*

[Fayyad, Piatetsky-Shapiro & Smyth 96]

Data

- refer to single instances (single objects, people, events, points in time, etc.)
- describe individual properties
- are often available in large amounts (databases, archives)
- are often easy to collect or to obtain (e.g., scanner cashiers in supermarkets, Internet)
- do not allow us to make predictions or forecasts

Knowledge

- refers to *classes* of instances (*sets* of objects, people, events, points in time, etc.)
- describes general patterns, structures, laws, principles, etc.
- consists of as few statements as possible
- is often difficult and time consuming to find or to obtain (e.g., natural laws, education)
- allows us to make predictions and forecasts

Main Industries for Data Analytics According to Experts

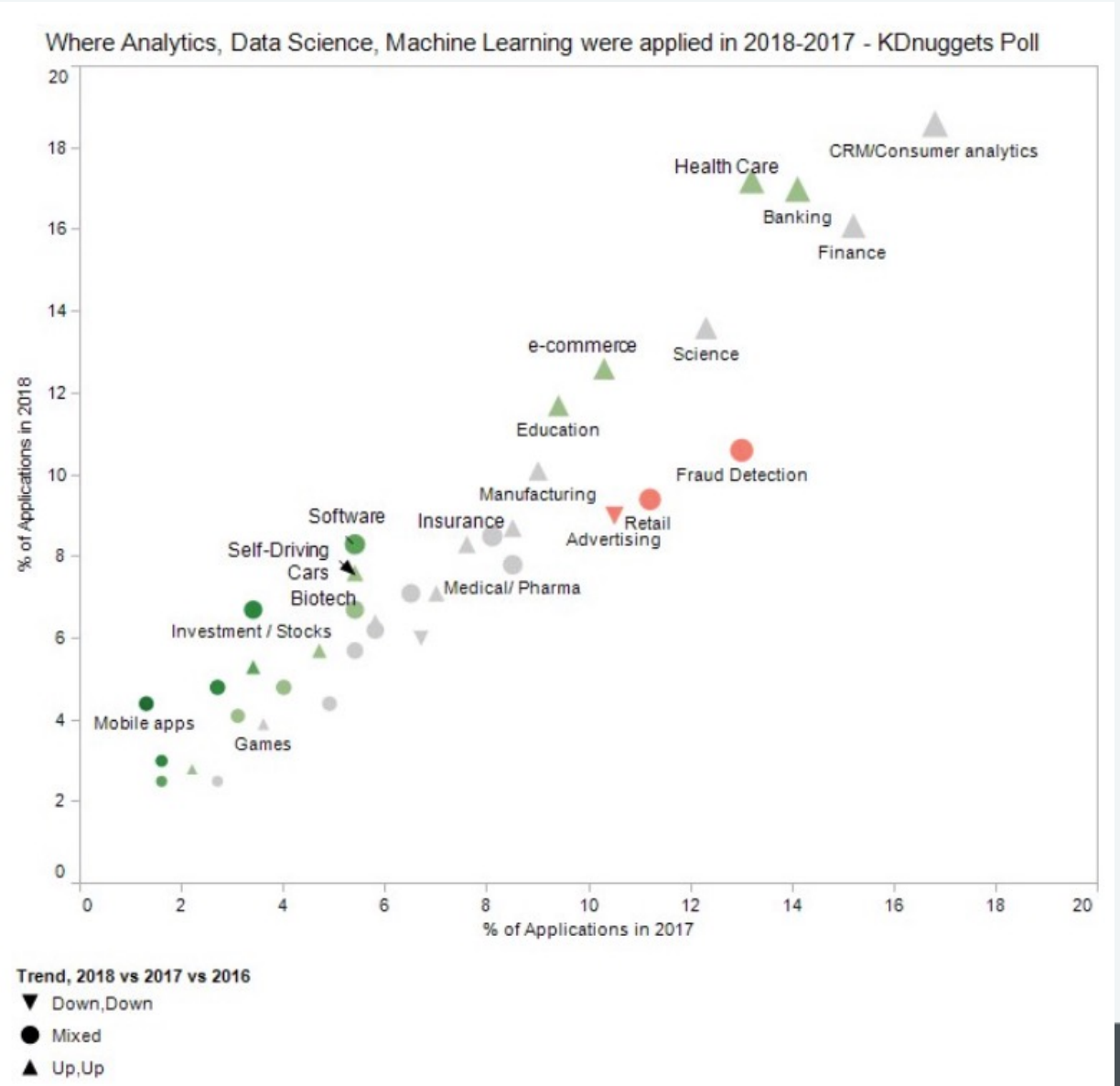


Fig. 1: Where Analytics, Data Science, Machine Learning were applied in 2017, 2018 - KDnuggets Polls

Source: Kdnuggets
https://www.kdnuggets.com/2019/03/poll-analytics-data-science-ml-applied-2018.html

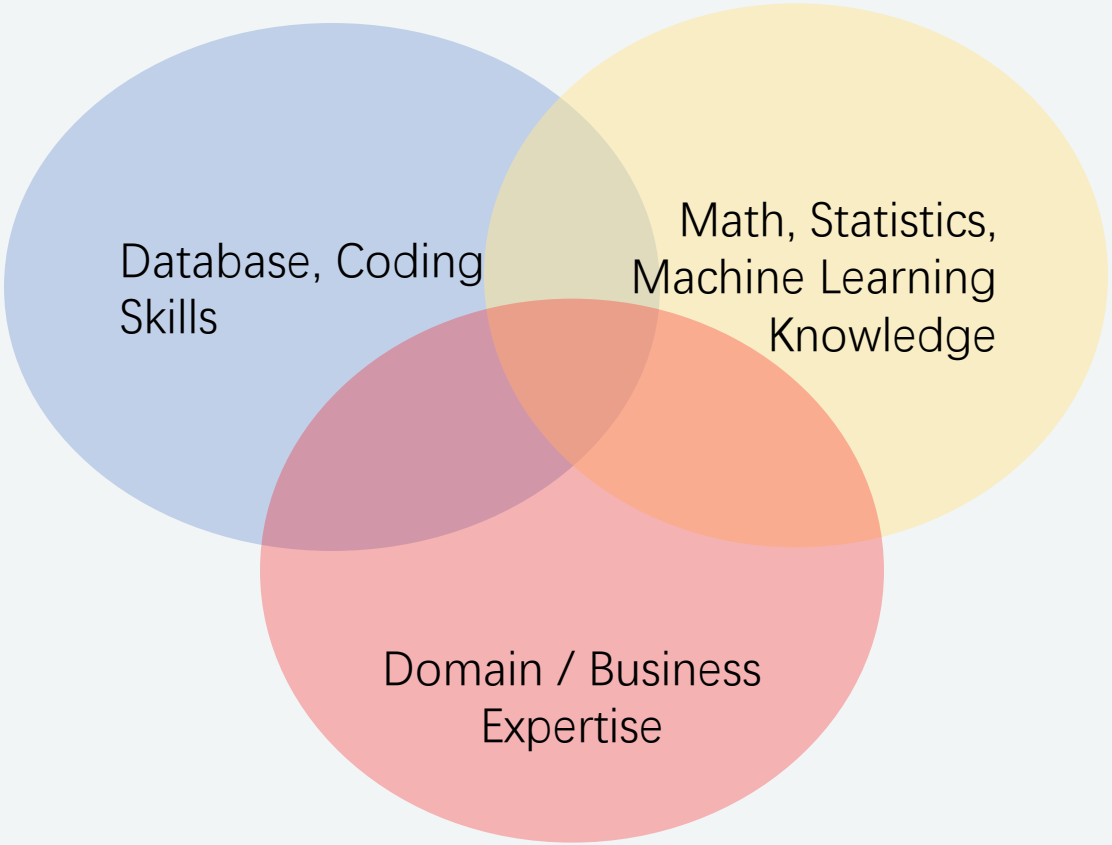
Readers of KDnuggets were asked:

What are the Industries / Fields where you applied Analytics, Data Science, Machine Learning in 2018?

The top 5 industries were:

1. **CRM/Consumer analytics**, 18.6% (number 1 for the last 5 years)
2. **Health care**, 17.2% (jump to number 2 from number 4)
3. **Banking**, 17.0%
4. **Finance**, 16.1%
5. **Science**, 13.6%

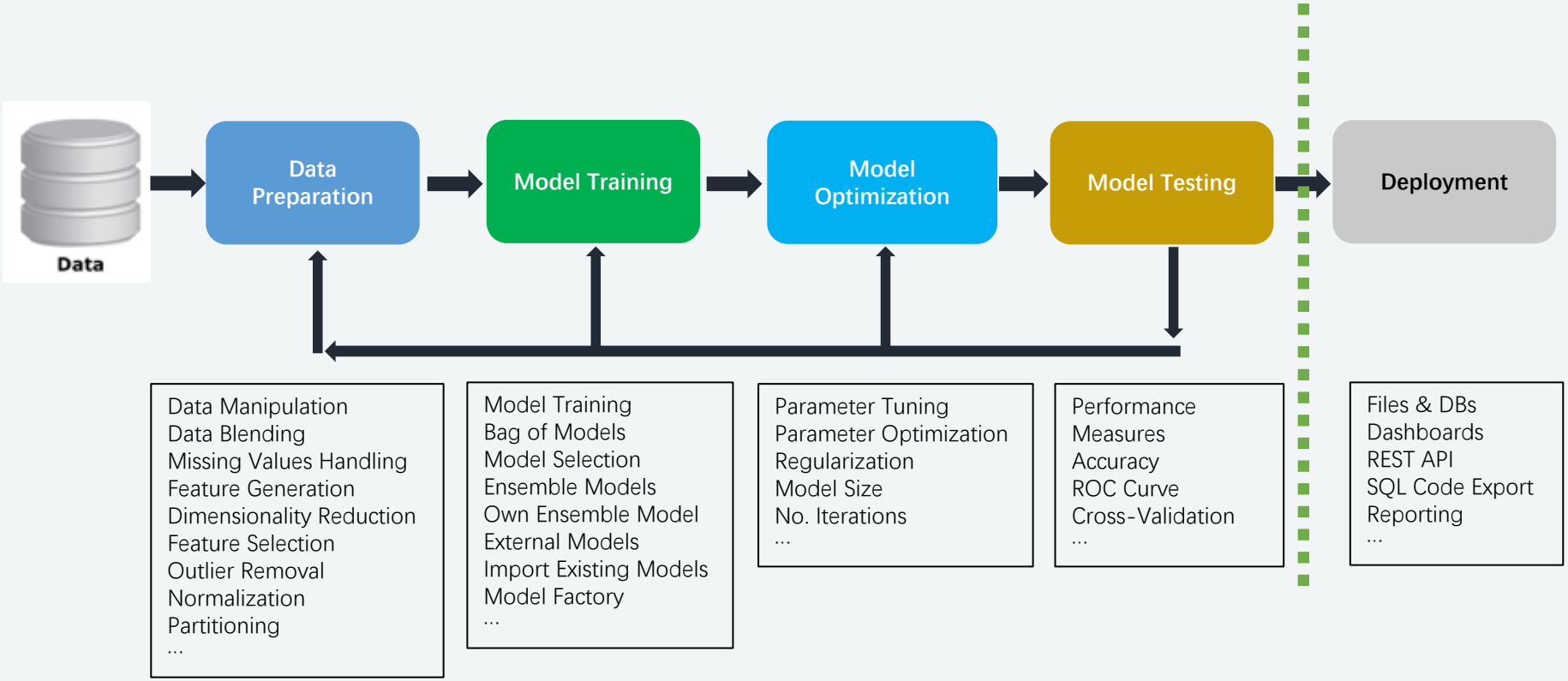
Data Science Skills



Contents

Process Flow of a Data Science Project

It always starts with some data ···



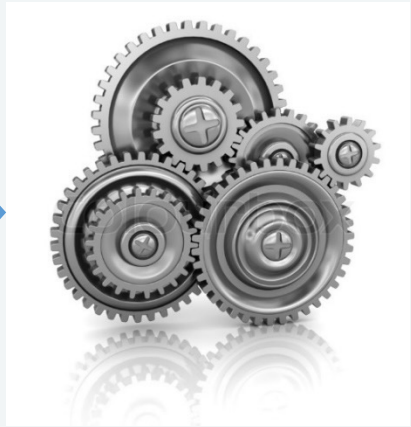
Contents

Example 1: Customer Segmentation

Business question: Which groups of customers am I serving?

CRM System
Data about your customer

- Demographics
- Behavior
- Revenues



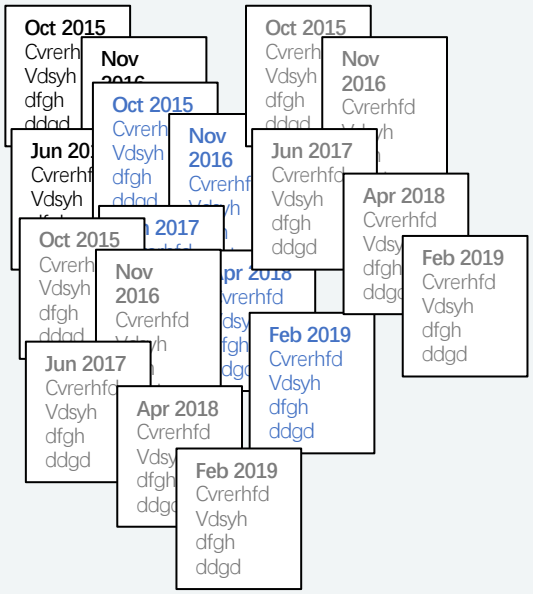
Model

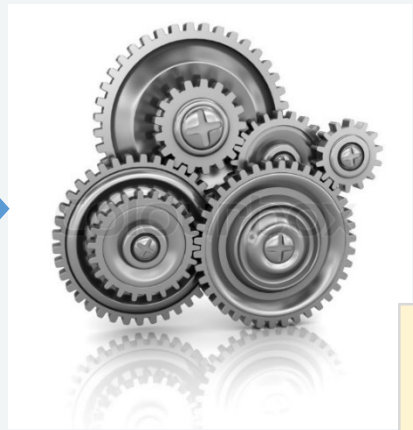
- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- Customer Segmentation
- …

Example 2: Risk Assessment

Business question: Is this person going to repay the loan?

Customer History





Model

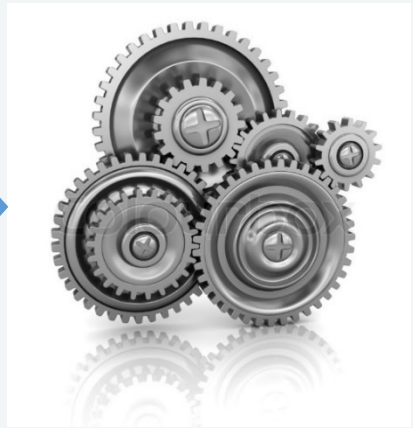
Risk Prognosis

- High Risk
- Low Risk
- High Risk
- Very High Risk
- Very Low Risk
- Medium Risk
- ...

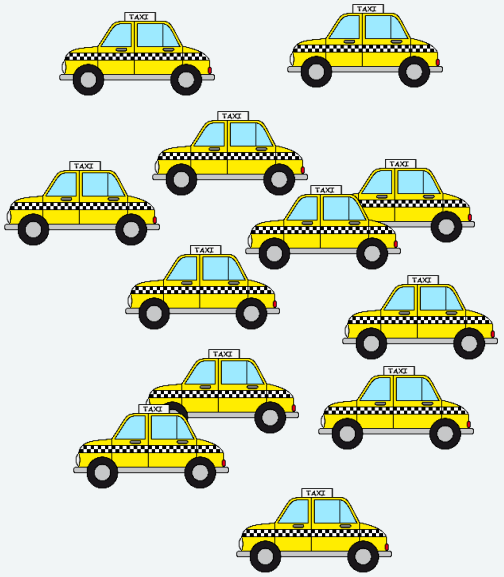
Example 3: Demand Prediction

Business questions:

- How many taxis do I need in NYC on Wednesday at noon?
- Or how many kW will be required tomorrow at 6am in London?
- Or how many customers will come tonight to my restaurant?

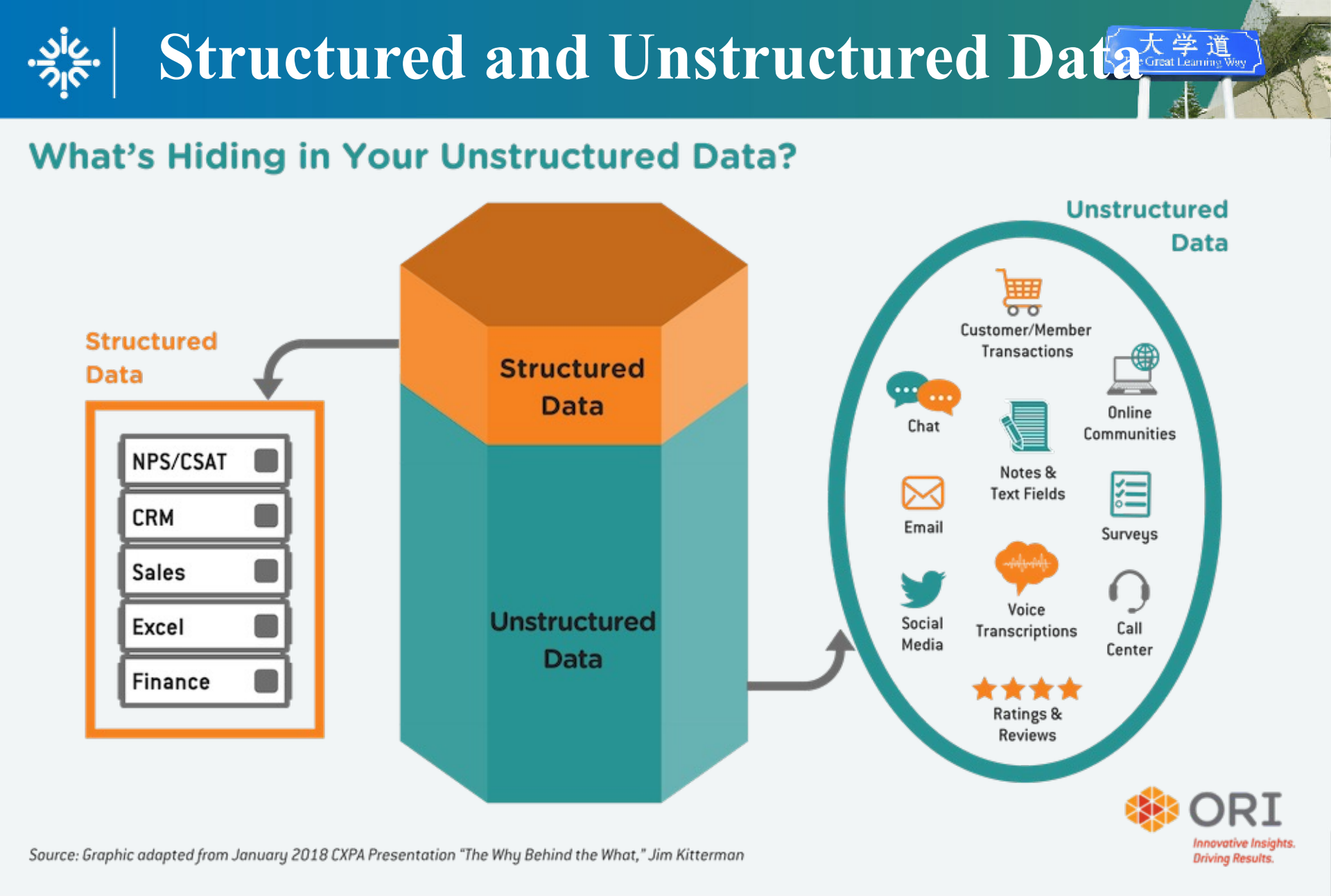


Model



Contents

Structured and Unstructured Data



Source: Graphic adapted from January 2018 CXPA Presentation "The Why Behind the What," Jim Kitterman

From Unstructured to Structured (Example 2)

- **“Thomas said: For our course, we’ll have our final exam on 9:30, 6/December. The exam will last 2 hours, place will be T7-501. Please revise carefully.”**
- {
  ”date”: 06/12/2021
  “time”: 09:30
  “duration”: 120
  “event”: “final exam”
  “type”: “closed book”
  “place”: “T7-501”
  “course code”: “GFQR1013”
  “course name”: “Hands on Data Analytics”
  “lecturer”: “Thomas Canhao XU”
  }

Common Structured Data Types

- Comma-separated values (CSV)
- Extensible Markup Language (XML)
- JavaScript Object Notation (JSON)
- Microsoft Excel (XLS)

Comma-separated Values

A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values.

- Each line of the file is a data record.
- Each record consists of one or more attributes. The attributes are separated by commas

- A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

Comma-separated Values

- Any field may be quoted (that is, enclosed within double-quote characters), while some fields must be quoted, as specified in the following rules and examples:
  - **`"1997","Ford","E350"`**
- Fields with embedded commas or double-quote characters must be quoted.
  - **`1997,Ford,E350,"Super, luxurious truck”`**
- Each of the embedded double-quote characters must be represented by a pair of double-quote characters.
  - **`1997,Ford,E350,"Super, ""luxurious"" truck"`**

Example of Quotes in CSV


```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended
Edition""","",4900.00
1999,Chevy,"Venture ""Extended Edition,
Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

Example of Quotes in CSV

Year	Make	Model	Description	Price
1997	Ford	E350	ac, abs, moon	3000.00
1999	Chevy	Venture "Extended Edition"		4900.00
1999	Chevy	Venture "Extended Edition, Very Large"		5000.00
1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded	4799.00

Contents

Data Table: Attribute and Instances

Attributes, features, variables…

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Instances, records, data objects, entries…

- Data can usually be described in terms of table or matrices
- Sometimes data are spread among different table that need to be **joined**

Data Types

Categorical | Ordinal | Numeric

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Numeric | Categorical

- Attributes differ for their **scale type**, according to the type of values that they can assume
- Three scale types:
  - Categorical / Nominal
  - Ordinal
  - Numeric

Data Processing

Chapter 2 – Hands on Data Analytics for Everyone

October 14, 2022

北京师范大学-香港浸会大学联合国际学院
United International College

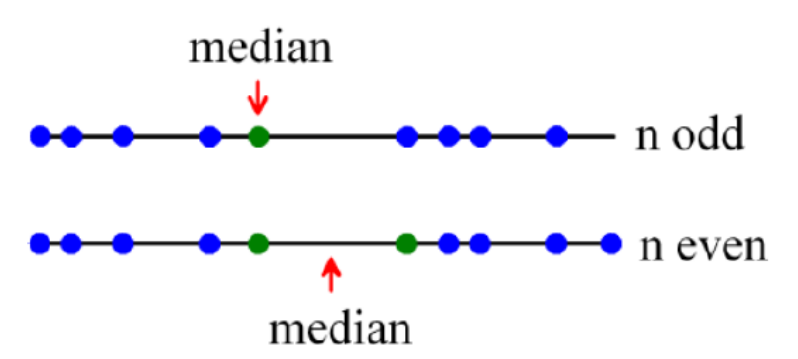
Contents

Basic Descriptive Statistics

Statistical measures can be used to describe a dataset:

- Range
- Min/max values
- Mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$
- Standard deviation $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}$
- Median (The middle number; found by ordering all data points and picking out the one in the middle - or if there are two middle numbers, taking the mean of those two numbers)
- Mode (Most frequently occurring value)
- Percentiles (Quartiles)
- Number of missing values
- ...

Median, Quantiles, Quartiles, Interquartile Range



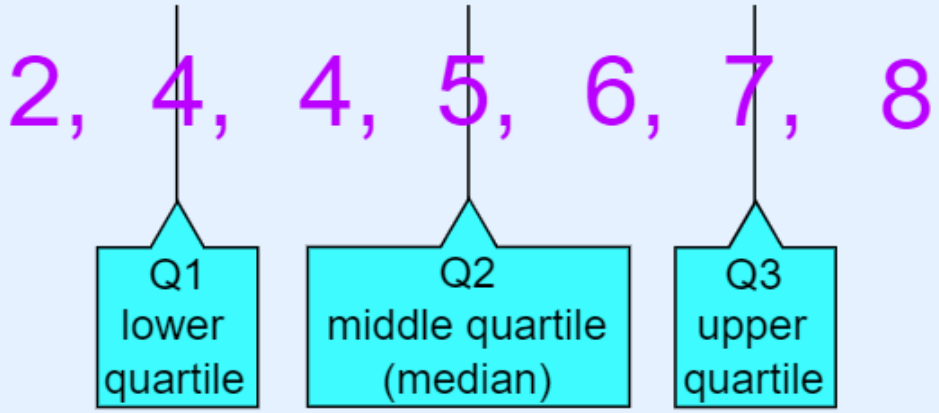
- **Median:** The value in the middle (for values sorted in increasing order)
- **q%-quantile** (0 < q < 100): The value for which q% of the values are smaller and 100-q% are larger. The median is the 50%-quantile
- **Quartiles**: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile)
- **Interquartile range (IQR):** 3rd quartile – 1st quartile

Median, Quantiles, Quartiles, Interquartile Range - Example

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

Cut the list into quarters:



And the result is:

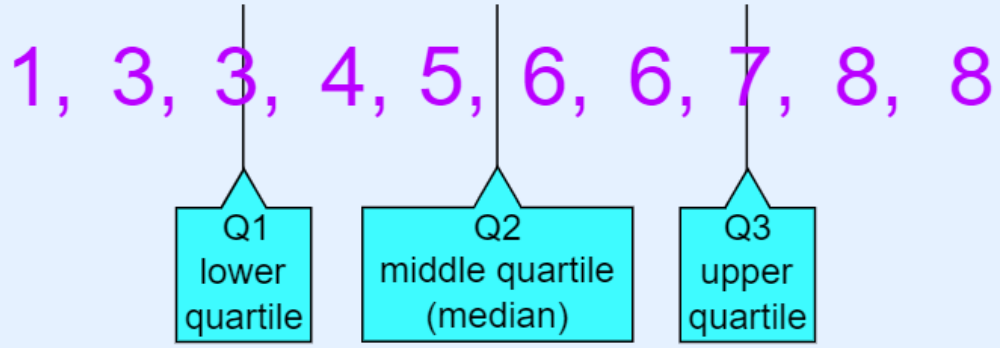
- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **7**

Median, Quantiles, Quartiles, Interquartile Range - Example

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:



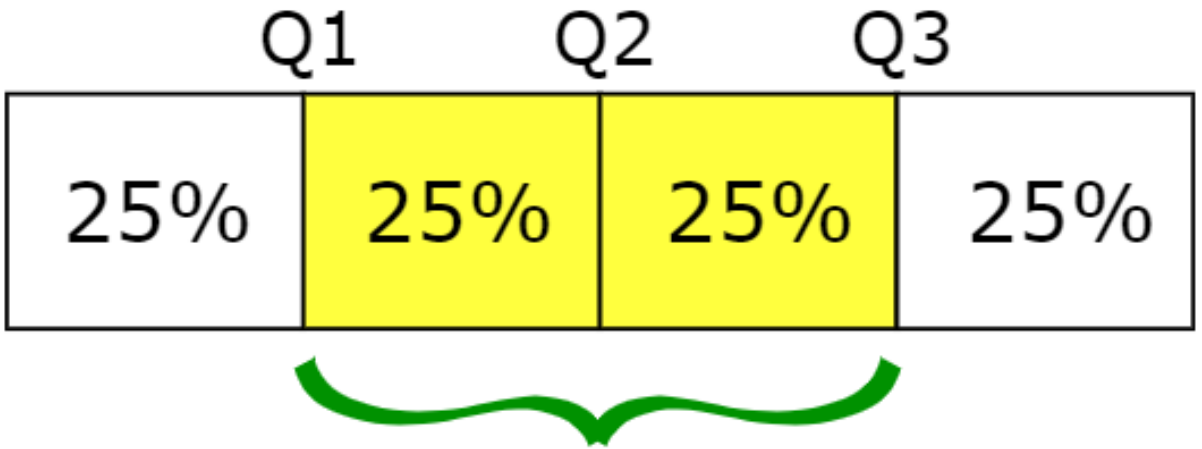
In this case Quartile 2 is half way between 5 and 6:

$$Q2 = (5+6)/2 = \mathbf{5.5}$$

And the result is:

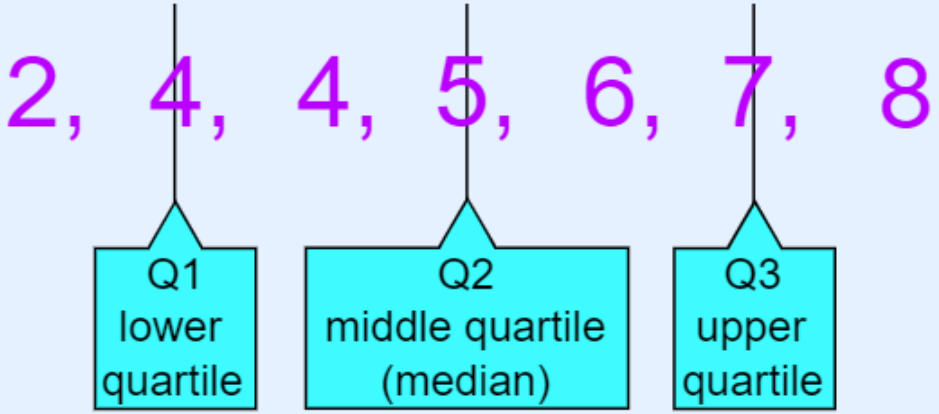
- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

Median, Quantiles, Quartiles, Interquartile Range - Example



Median, Quantiles, Quartiles, Interquartile Range - Example

Example:



The **Interquartile Range** is:

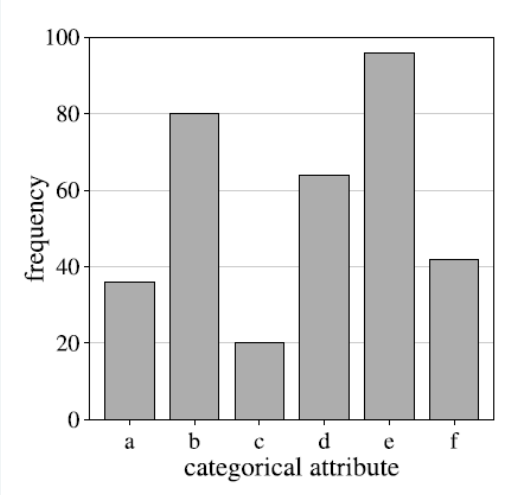
$$Q3 - Q1 = 7 - 4 = \mathbf{3}$$

https://www.mathsisfun.com/data/quartiles.html

Contents

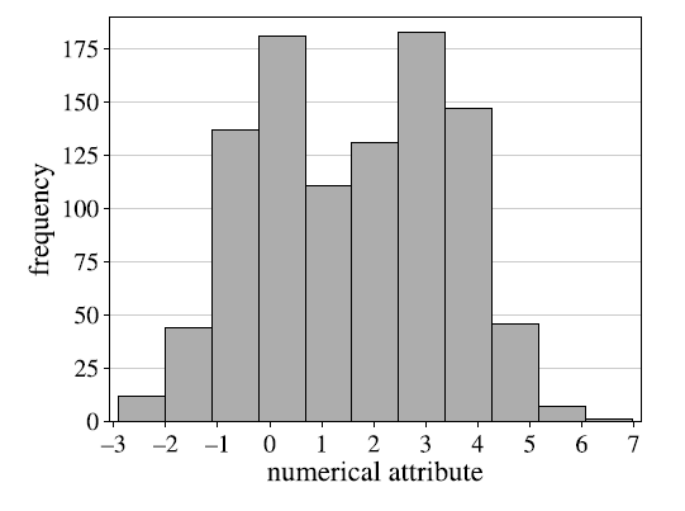
- **Data Summary and Visualization**
  - **Descriptive Statistics**
  - **Visualization for 1 or 2 Dimensions**
  - **Visualization for Higher Dimensions**
- **Feature Selection and Dimensionality Reduction**
- **Data Cleaning**
  - **Missing Values Imputation**
  - **Outliers**
  - **Data Type (Numerical and Categorical) Transformation**
  - **Data Normalization**
  - **String REGEX**
- **Feature Engineering**
- **Data Integration**
- **Lab (Demo): Data Import, Filtering and Visualization**
- **Assignment 3 (In-Class Lab) : Data Processing with KNIME**
  - **Lab3.1: Visualization**
  - **Lab 3.2: Data Cleaning**
- **Assignment 4: In-Class Quiz**

Bar Chart



A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

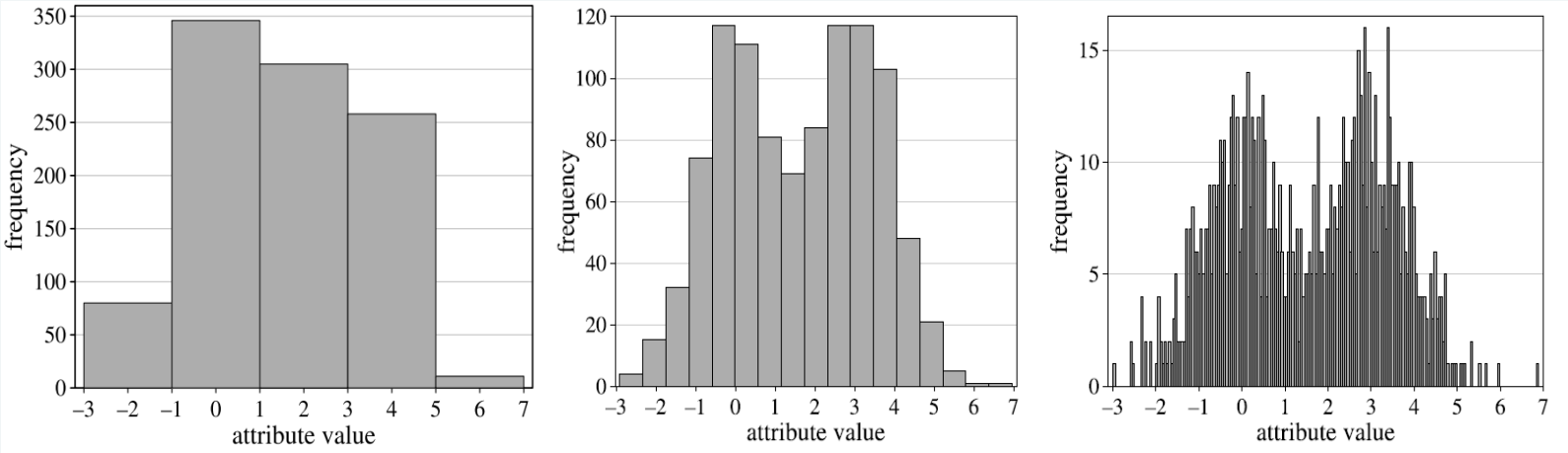
Histogram



- A histogram shows the frequency distribution for a numerical attribute.
- The range of the numerical attribute is discretized into a fixed number of intervals (bins), usually of equal length.
- For each interval, the (absolute) frequency of values falling into it is indicated by the height of a bar.

Choice of Number of Bins

The histogram in the figures resulted from a sample of size $n = 1000$



Choosing a low number of bins, the two peaks of the original distribution are no longer visible, and one gets the wrong impression that the distribution is unimodal

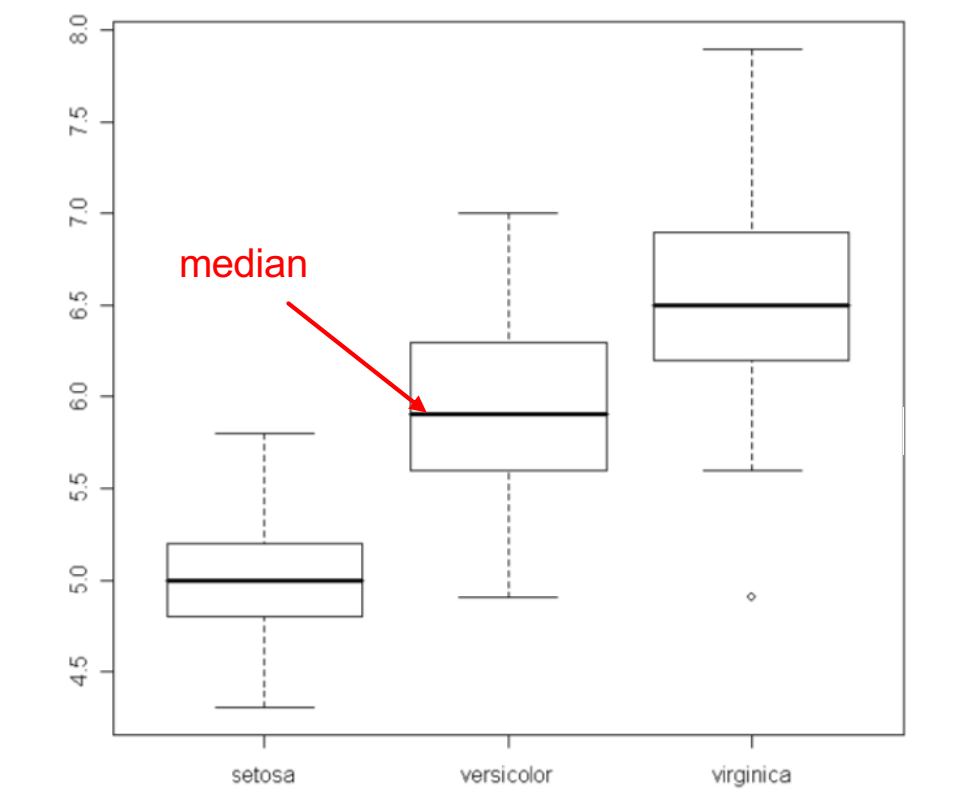
Choosing a high number of bins usually leads to a very scattered histogram in which it is difficult to distinguish true peaks from random peaks

Best choice for number k of bins in the histogram?

- Sturge ‘s Rule $k = \lceil log_2(n) + 1 \rceil$

Boxplots

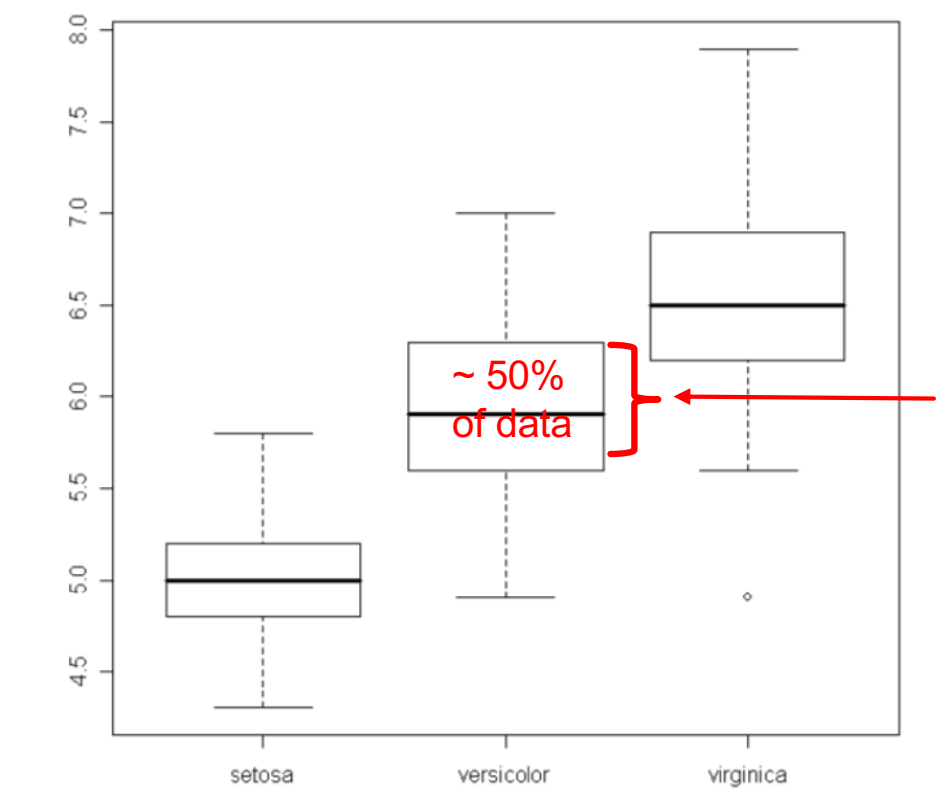
Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the ***median***, the IQR, and possible outliers



https://www.khanacademy.org/math/statistics-robability/summarizing-quantitative-data

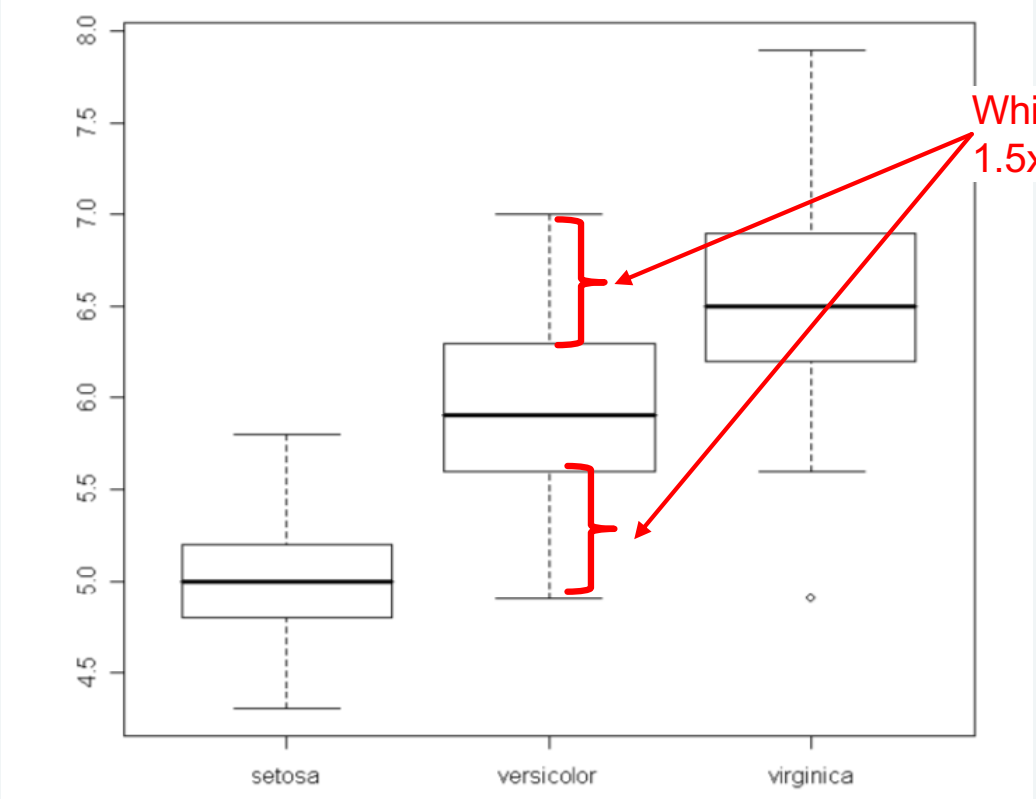
Boxplots

Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the ***IQR***, and possible outliers



Boxplots

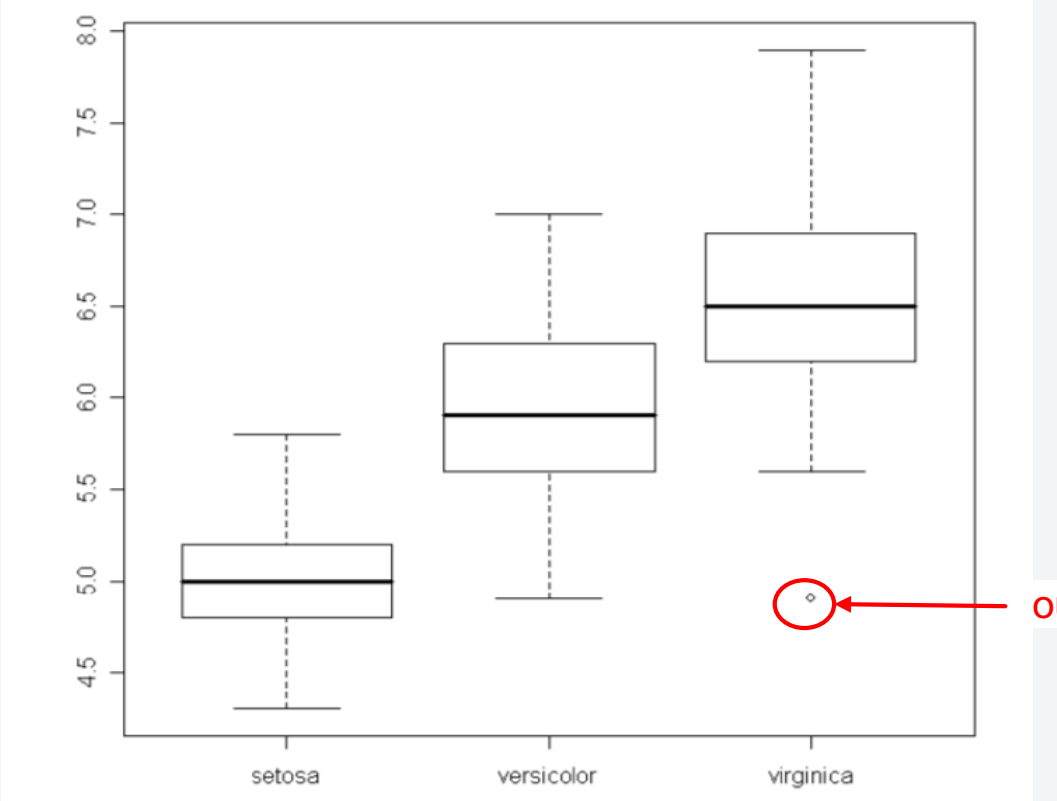
Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the ***IQR***, and possible outliers



The maximum length of each whisker is 1.5 times the length of the interquartile range. But if there is no data point at the maximum length of a whisker, the corresponding whisker is shortened until it reaches the next data point.

Boxplots

Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the *IQR*, and possible **outliers**



Data points lying outside the whiskers are considered as outliers and are indicated in the form of small circles.

Contents

- Data Summary and Visualization
  - Descriptive Statistics
  - Visualization for 1 or 2 Dimensions
  - Visualization for Higher Dimensions
- **Feature Selection and Dimensionality Reduction**
- Data Cleaning
  - Missing Values Imputation
  - Outliers
  - Data Type (Numerical and Categorical) Transformation
  - Data Normalization
  - String REGEX
- Feature Engineering
- Data Integration
- Lab (Demo): Data Import, Filtering and Visualization
- Assignment 3 (In-Class Lab) : Data Processing with KNIME
  - Lab3.1: Visualization
  - Lab 3.2: Data Cleaning
- Assignment 4: In-Class Quiz

Dimensionality Reduction Techniques

- Measure based
  - Ratio of missing values
  - Low variance
  - High Correlation

Dimensionality Reduction Based on Missing Values Ratio

First partition (as defined in dialog) - 0:337:0:276 - Partitioning (80% vs. 20%)

File Hilite Navigation View

Table "default" - Rows: 40000 | Spec - Columns: 231 | Properties | Flow Variables

Row ID	D Var16	I Var17	I Var18	I Var19	S Var20	I Var21	I Var22	I Var23	I Var24	I Var25	I Var26	I Var27	D Var28
Row0	?	?	?	?	?	464	580	?	14	128	?	?	166.56
Row1	?	?	?	?	?	168	210	?	2	24	?	?	353.52
Row2	?	?	?	?	?	1212	1515	?	26	816	?	?	220.08
Row4	?	?	?	?	?	64	80	?	4	64	?	?	200
Row7	?	?	?	?	?	32	40	?	2	16	?	?	230.56
Row8	?	?	?	?	?	200	250	?	2	64	?	?	300.32
Row10	?	?	?	?	?	92	115	?	6	112	?	?	133.12
Row11	?	?	?	?	?	236	295	?	8	40	?	?	133.12
Row12	?	?	?	?	?	0	0	?	?	0	?	?	240.56
Row13	?	?	?	?	?	480	600	?	10	216	?	?	176.56
Row14	?	?	?	?	?	148	185	?	0	8	?	?	236.08
Row16	?	?	?	?	?	584	730	?	6	320	?	?	220.08
Row17	?	?	?	?	?	168	210	?	2	32	?	?	166.56
Row18	?	?	?	?	?	12	15	?	2	0	?	?	253.52
Row20	?	?	?	?	?	168	210	?	2	56	?	?	272.08
Row21	?	?	?	?	?	20	25	?	2	0	?	?	86.96
Row22	?	?	?	?	?	[illegible]	[illegible]	?	[illegible]	[illegible]	?	?	166.56
Row23	?	?	?	?	?	[illegible]	[illegible]	?	[illegible]	[illegible]	?	?	198.88
Row24	?	?	?	?	?	216	270	?	8	128	?	?	200
Row25	?	?	?	?	?	152	190	?	4	16	?	?	20.08
Row26	?	0	0	0	?	?	?	?	?	?	?	?	?
Row28	?	?	?	?	?	0	0	?	?	0	?	?	257.28
Row29	?	?	?	?	?	312	390	?	0	120	?	?	200
Row30	?	?	?	?	?	112	140	?	4	56	?	?	166.56
Row31	?	?	?	?	?	28	35	?	0	16	?	?	285.2
Row33	?	?	?	?	?	160	200	?	4	40	?	?	[illegible]
Row36	?	?	?	?	?	612	765	?	14	360	?	?	200
Row37	?	?	?	?	?	380	475	?	4	208	?	?	336.56
Row38	?	?	?	?	?	76	95	?	0	16	?	?	213.36
Row40	?	?	?	?	?	228	285	?	22	56	?	?	200
Row41	?	?	?	?	?	120	150	?	10	80	?	?	133.12
Row42	?	5	0	0	?	?	?	?	?	?	?	?	?
Row43	?	?	?	?	?	72	90	?	0	40	?	?	191.36
Row44	?	?	?	?	?	0	0	?	?	0	?	?	120.4
Row47	?	?	?	?	?	0	0	?	?	0	?	?	186.64
Row48	?	?	?	?	?	172	215	?	4	200	?	?	137.68
Row49	?	?	?	?	?	0	0	?	?	0	?	?	274.16

IF (% missing value > threshold) THEN remove column

Missing Value

Dimensionality Reduction Based on Low Variance

Output table - 0:347:0:337 - Missing Value (Numeric: 0)

File Hilite Navigation View

Table "default" - Rows: 40000 | Spec - Columns: 231 | Properties | Flow Variables

Row ID	20	Var21	Var22	Var23	Var24	Var25	Var26	Var27	Var28
Row51	h	336	420	0	8	72	0	0	133.12
Row52	h	120	150	0	0	16	0	0	286.96
Row54	h	124	155	0	0	0	0	0	234.72
Row55	h	184	230	0	4	64	0	0	642.64
Row56	h	268	335	0	4	88	0	0	133.12
Row57	h	128	160	0	0	96	0	0	198.88
Row59	h	132	165	0	0	112	0	0	253.52
Row60	h	44	55	0	0	24	0	0	186.64
Row61	h	104	130	0	4	72	0	0	166.56
Row62	h	212	265	0	6	136	0	0	379.6
Row63	h	20	25	0	0	0	0	0	166.56
Row65	h	492	615	0	18	256	0	0	133.12
Row66	h	148	185	0	2	8	0	0	186.64
Row68	h	140	175	0	2	40	0	0	176.56
Row69	h	0	0	0	0	0	0	0	166.56
Row71	h	0	0	0	0	0	0	0	392.08
Row72	h	124	155	0	6	88	0	0	153.2
Row73	h	152	190	0	0	32	0	0	253.52
Row74	h	324	405	0	8	104	0	0	186.64
Row75	h	0	0	0	0	0	0	0	0
Row76	h	60	75	0	6	0	0	0	200
Row77	h	180	225	0	4	88	0	0	166.56
Row78	h	232	290	0	4	144	0	0	200
Row79	h	16	20	0	0	16	0	0	313.68
Row81	h	152	190	0	0	48	0	0	220.08
Row82	h	108	135	0	4	88	0	0	166.56

Note: requires min-max-normalization, and only works for **numeric** columns

If column has **constant** value (variance = **0**), it contains no useful information

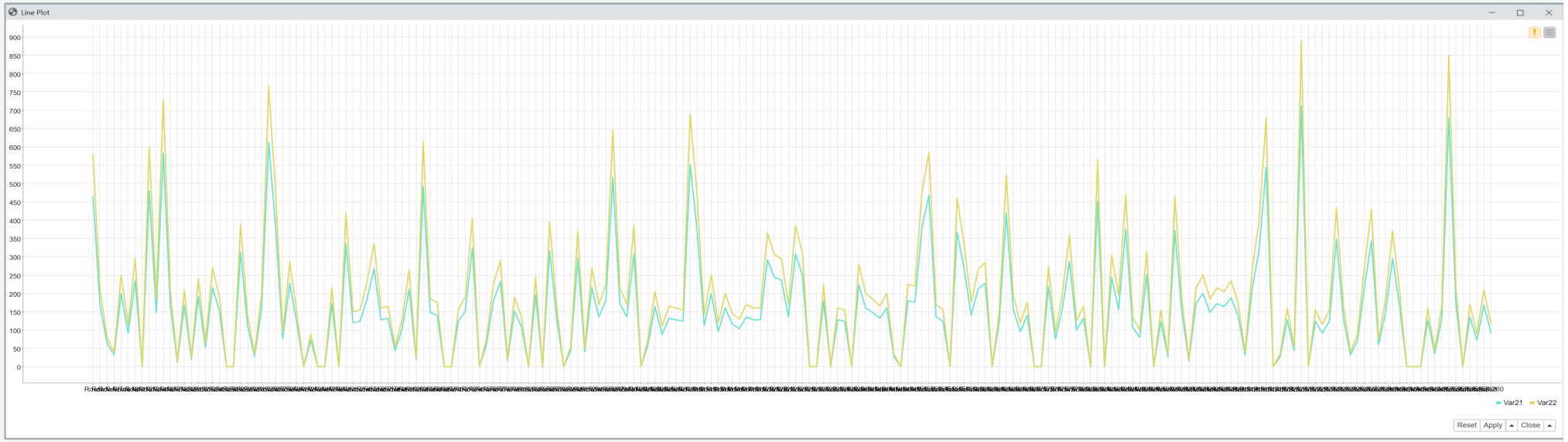
In general: **IF (variance < threshold) THEN remove column**

Dimensionality Reduction Based on High Correlation

Two **highly correlated** input variables probably carry similar information

$$IF\ (corr(var1, var2) > threshold\ \Rightarrow remove\ var1$$

Note: requires min-max-normalization of numeric columns



Contents

- Data Summary and Visualization
  - Descriptive Statistics
  - Visualization for 1 or 2 Dimensions
  - Visualization for Higher Dimensions
- Feature Selection and Dimensionality Reduction
- **Data Cleaning**
  - **Missing Values Imputation**
  - Outliers
  - Data Type (Numerical and Categorical) Transformation
  - Data Normalization
  - String REGEX
- Feature Engineering
- Data Integration
- Lab (Demo): Data Import, Filtering and Visualization
- Assignment 3 (In-Class Lab) : Data Processing with KNIME
  - Lab3.1: Visualization
  - Lab 3.2: Data Cleaning
- Assignment 4: In-Class Quiz

Missing Values: Types

Types of missing values:

Example: Suppose you are modeling weight Y as a function of sex X

- **Missing Completely At Random** (MCAR): the probability that a value for X is missing does neither depend on the value of X nor on other variables.
  There may be no particular reason why some people told you their weights and others didn't.

- **Missing At Random** (MAR): the probability that Y is missing depends only on the value of X.
  One sex X may be less likely to disclose its weight Y.

- **Not Missing At Random** (NMAR): the probability that Y is missing depends on the unobserved value of Y itself.
  Heavy (or light) people may be less likely to disclose their weight.

Missing Values Imputation

How to handle missing values?

- Ignore/delete the record
- Fill in (impute) missing value as:
- **Fixed value**: e.g., "unknown", -9999, -1 when only positive numbers in the domain, etc.
- Attribute **mean / median / mode**
- Attribute **most frequent value**
- **Next / previous /avg interpolation / moving avg value** (in time series)
- **A predicted value** based on the other attributes (inference-based such as Bayesian, Decision Tree, ...)

Contents

What are outliers?

- An **outlier** is a value or data object that is far away or very different from all or most of the other data.
- Errors in measurements or exceptional conditions that don't describe the common functioning of the underlying system
- Outliers are supposed to be rare

Outlier Detection

Causes for outliers:

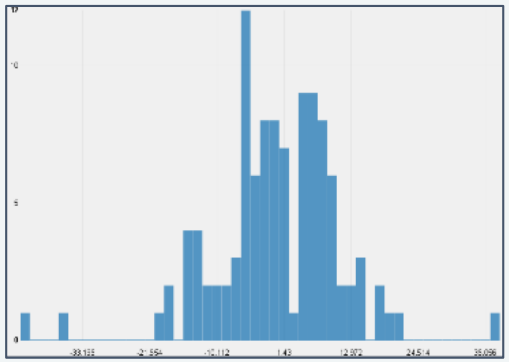
- Data **quality** problems (erroneous data coming from wrong measurements or typing mistakes)
- Exceptional or unusual **situations**/data objects.

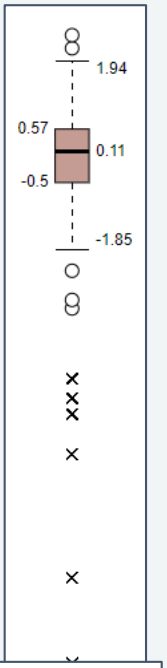
Outlier handling :

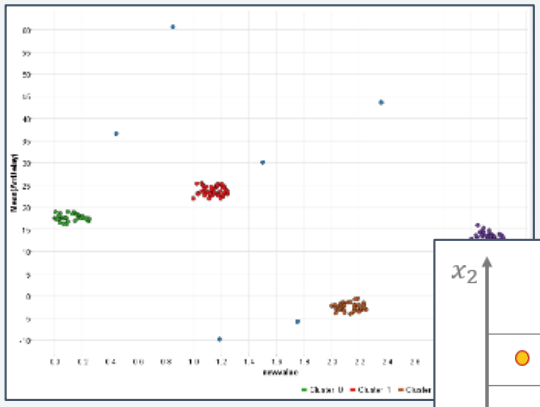
- Outliers coming from erroneous data should be **excluded** from the analysis.
- Even if the outliers are correct (exceptional data), it is sometimes useful to exclude them from the analysis.
- For example, a single extremely large outlier can lead to completely misleading values for the mean value.

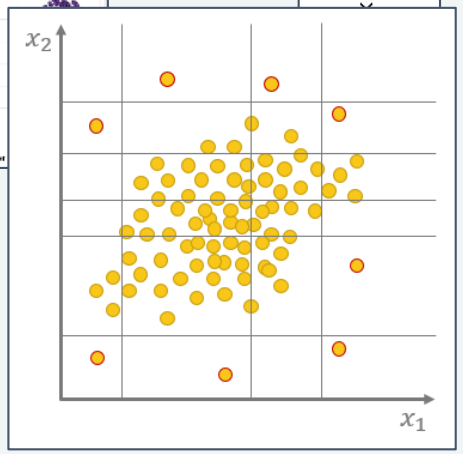
Outlier Detection Techniques

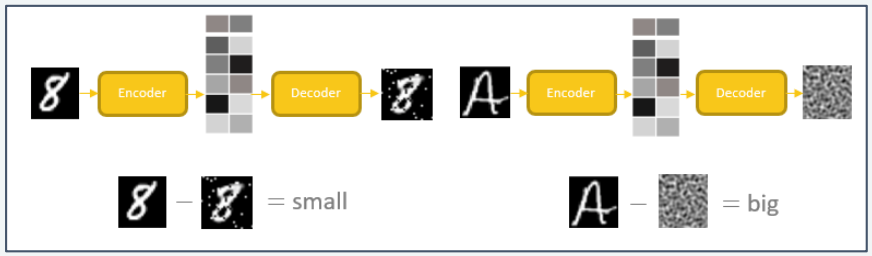
- Knowledge-based
  - We know that a 200 year old person must be a mistake
  - We know that "A" in a number corpus is an outlier
- Statistics-based
  - Distance from the median
  - Position in the distribution tails
  - Distance to the closest cluster center
  - Error produced by an autoencoder
  - Number of random splits to isolate a data point from other data











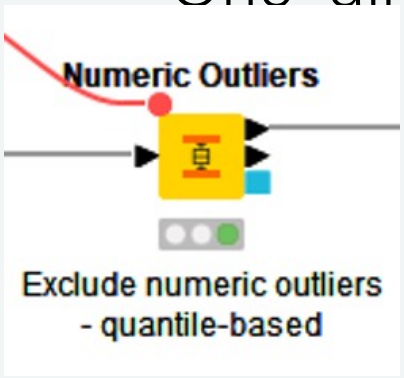
Statistical Methods for Outlier Detection

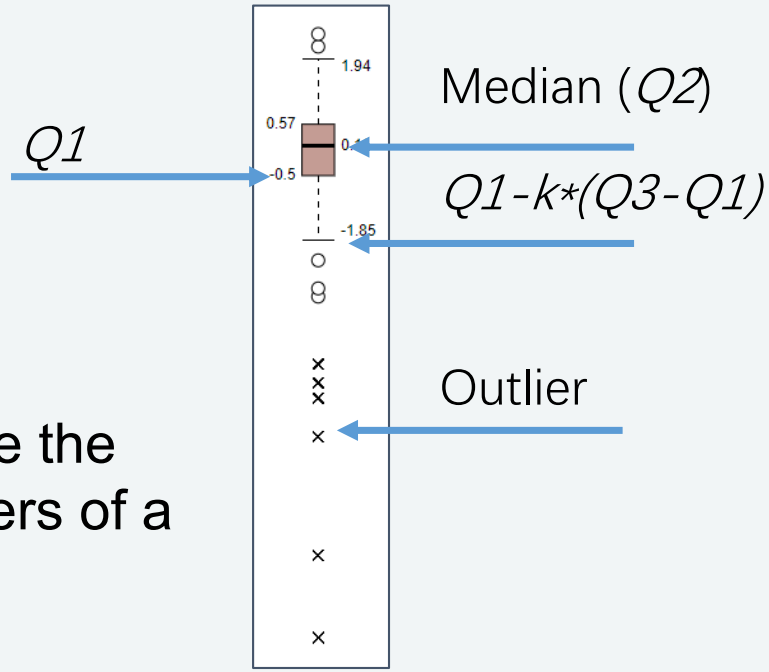
- Quantile-based: **Box plot**
- Distribution-based: **Z-Score**
- Cluster-based: **DBSCAN**
- Neural Autoencoder
- Isolation Forest
- …

Quartile based (Box Plots)

Challenges:

- Outliers in the data expand the quantiles
- Skewed data might require different *k* to detect upper and lower outliers
- One-dimensional





Flag data points outside the upper and lower whiskers of a box plot as outliers

Contents

- Data Summary and Visualization
  - Descriptive Statistics
  - Visualization for 1 or 2 Dimensions
  - Visualization for Higher Dimensions
- Feature Selection and Dimensionality Reduction
- **Data Cleaning**
  - Missing Values Imputation
  - Outliers
  - Data Type (Numerical and Categorical) Transformation
  - **Data Normalization**
  - String REGEX
- Feature Engineering
- Data Integration
- Lab (Demo): Data Import, Filtering and Visualization
- Assignment 3 (In-Class Lab) : Data Processing with KNIME
  - Lab3.1: Visualization
  - Lab 3.2: Data Cleaning
- Assignment 4: In-Class Quiz

Construct Data: Assure Impartiality

- In absence of domain knowledge, different techniques can be applied

- **min–max normalization** $n : dom\ X \to [0,1],\ x \longmapsto \frac{x - minX}{max_X - minX}$

Both sensitive to outliers!

- **z-score standardization** $s : dom\ X \to \mathbb{R},\ x \longmapsto \frac{x - \hat{\mu}_X}{\hat{\sigma}_X}$

- **robust z-score standardization** $s : dom\ X \to \mathbb{R},\ x \longmapsto \frac{x - \tilde{x}}{IQR_X}$

- **decimal scaling** $d : dom\ X \to [0,1],\ x \longmapsto \frac{x}{10^s}$

Contents

Example: Feature Engineering by Scale Conversion

Scale Conversion

- Categorical → Numerical: map categorical and ordinal values to a set of binary values
- Numerical → Categorical: **Discretization** (equal-width, equal-depth, V-optimal)

Machine Learning

Chapter 3 – Hands on Data Analytics for Everyone

November 3, 2022

北京师范大学-香港浸会大学联合国际学院
United International College

Contents

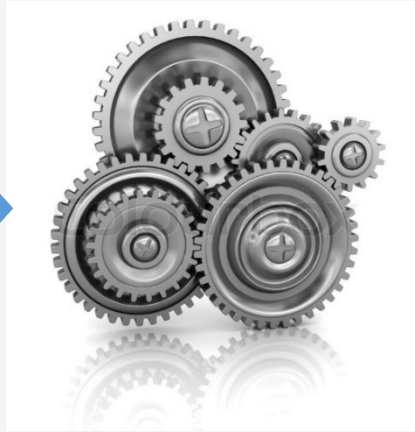
What is a Model (Learning Algorithm)?

A model or learning algorithm is simply a specification of a mathematical (or probabilistic) **relationship** that exists between different variables.

- Input features
- Input attributes
- Independent variables

$\boldsymbol{X} = (x_1, x_2, \ldots, xn)$

Model



y

- Class
- Label
- Target
- Output feature/attribute
- Dependent variable

$$y = f(\boldsymbol{\beta}, \boldsymbol{X}) \quad \text{with } \boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_m]$$

A learning algorithm adjusts (learns) the model **parameters $\boldsymbol{\beta}$** throughout a number of iterations to maximize/minimize a likelihood/error function on y.

What Is Machine Learning?

Learning = Improving with **experience** at some **task**

Arthur Samuel (1959)

- Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998)

- Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What Is Machine Learning?

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam.

What is the task T, the experience E and the measure P in this setting?

1. Classifying emails as spam or not spam.
2. Watching you label emails as spam or not spam.
3. The number (or fraction) of emails correctly classified as spam/not spam.

Contents

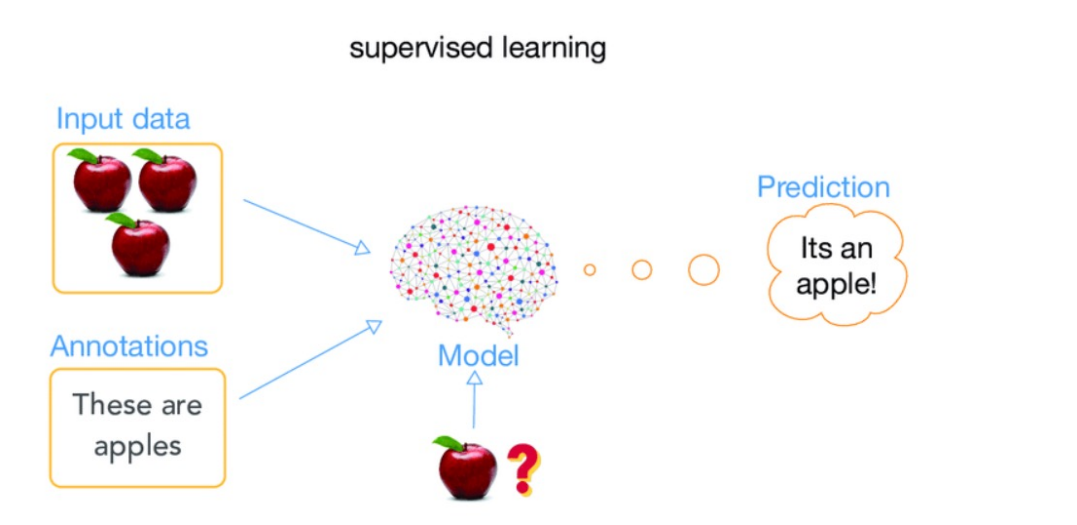
Supervised Learning

The learner is provided with a set of data inputs together with the corresponding desired outputs

- Data act as a "teacher"

Example:

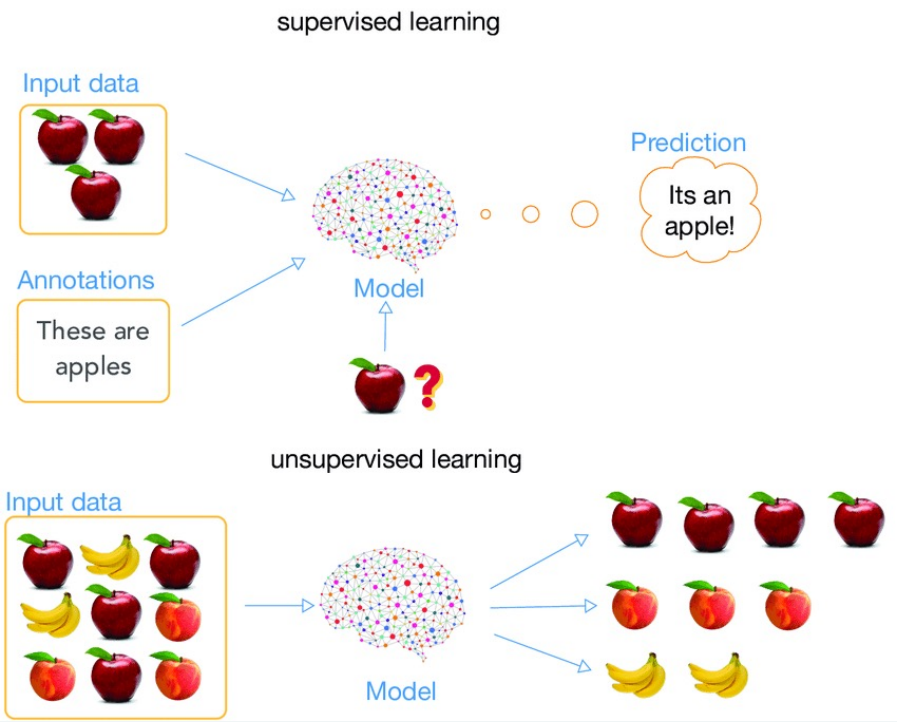
- teach kids to recognize different animals
- grade examinations with correct answer provided



Unsupervised Learning

Training examples as input patterns, with no associated output

- no “teacher”
- similarity measure exists to detect groupings/ clusterings



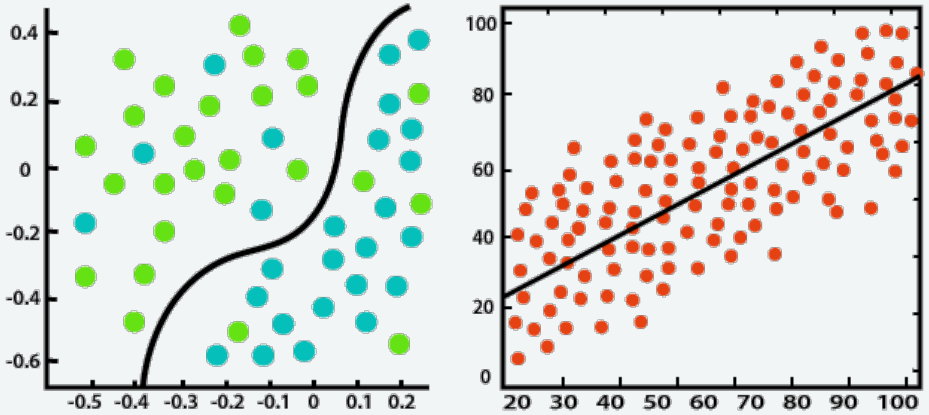
Contents

- Introduction to Machine Learning
  - Supervised and Unsupervised Learning
  - **Classification and Regression**
- Linear Regression (Supervised Learning)
  - Model
  - Performance Evaluation
- Classification (Supervised Learning)
  - How to Perform a Classification
  - Classification Tree Model
- Clustering Method (Unsupervised Learning)
  - Objective
  - Similarity Measures
  - (Optional) Method 1: Hierarchical Clustering
  - (Optional) Method 2: K-Means Method (Clustering by Partitioning)
- Lab (Demo): Unsupervised Learning
- Assignment 5: Supervised Learning
- Assignment 6: In-Class Quiz

Regression vs Classification (Supervised Learning)

Supervised learning: Given the “right answer” for each example in the data.

- When the target variable that we’re trying to predict is continuous, we call the learning problem a regression problem.
- When the target variable can take on only a small number of discrete values, we call it a classification problem.



Classification Regression

Classification Example

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
11	No	Married	80K	**?**
12	Yes	Single	100K	**?**

Regression Example

Suppose we have a dataset giving the **living areas** and **prices** of 47 houses from Portland, Oregon:

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Contents

Linear Regression – Problem Description

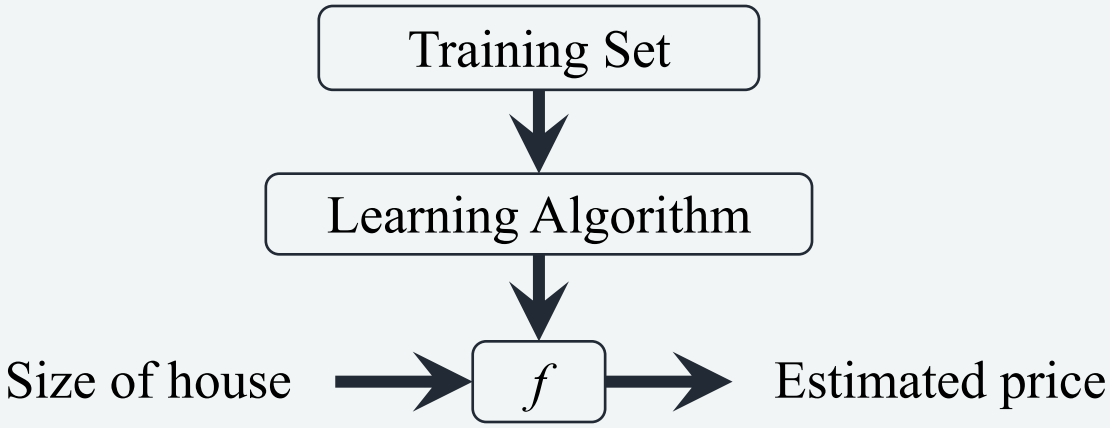
Notation:

- Input variable/feature: x
- Output/target variable: y
- Training example: $(x^{(i)}, y^{(i)})$
- Training set: $\{(x^{(i)}, y^{(i)}); i = 1,2, \dots, m\}$ (a list of m training examples)
- Space of input values: X ; space of output values: Y

To describe the problem slightly more formally, our goal is:

Given a training set, to learn a function (hypothesis/model) $f: X \longmapsto Y$, so that $f(x)$ is a "good" predictor for the corresponding value of y.

Linear Regression – Model Representation



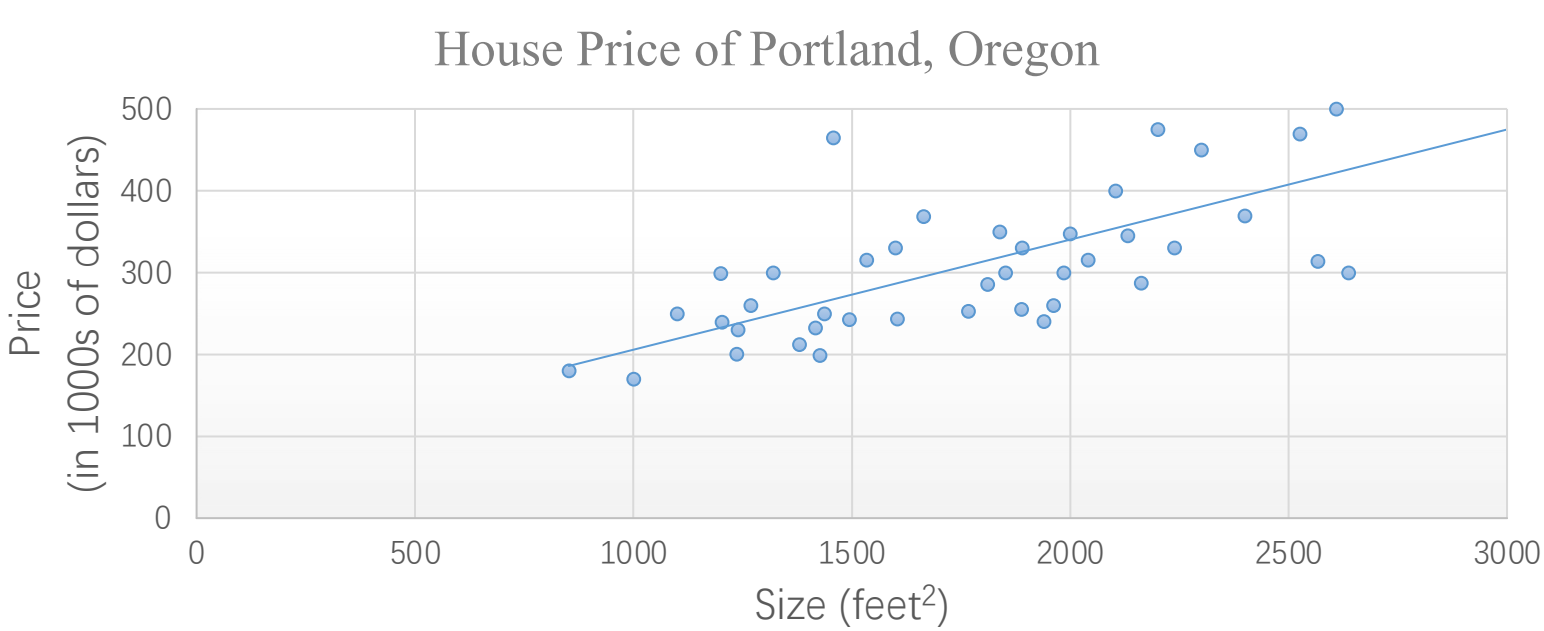
How do we represent *f*?

Linear Regression: $f(x) = \theta_0 + \theta_1 x$

- The model is in linear in terms of parameters θ_0 and θ_1
- Linear regression with one variable (univariate linear regression).

We will predict that y is a linear function of x (straight line)

Regression Example



We can predict that the house price is a linear function of house size

Regression Model Evaluation

Given dataset $D = \{(x^{(i)}, y^{(i)}); i = 1,2, \dots, m\}$ and a regression model f, evaluate the performance of the model using following metrics.

Error Metric	Formula	Notes
Mean absolute error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - f(x_i)\rvert$	Average of the absolute difference between the actual and predicted values.
Mean squared error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$	Average of the squared difference between the actual and predicted values.
Root mean squared error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$	Square root of Mean Squared error.
R-squared	$1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$	Proportion of the variance for a dependent variable that's explained by the regression model. Normally ranges from 0 to 1, the closer to 1 the better performance

Contents

Generally, the difference between the actual predicted output of the learner and the true output of the sample is called "error"

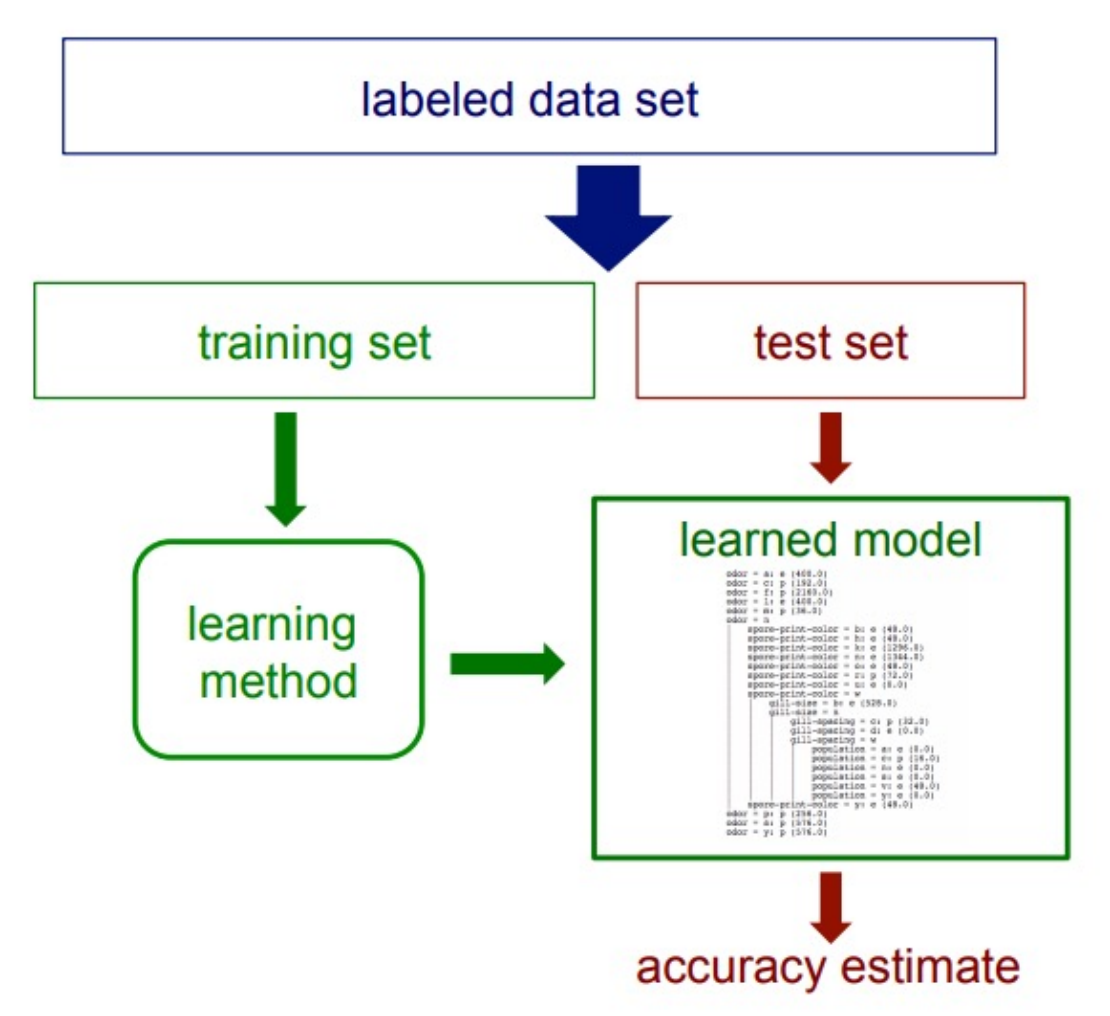
- Training error/ empirical error: the error of the learner/model on the training data
- Generalization error: the error on the new data

We want to get a learner with a small generalization error

However, we do not have the information for new data, instead we try to minimize the empirical error on the training data

- split data randomly into a training set and a test set (e.g., a 70%/30% split).
- train your model on the training set and see how it performs on the test set.
- use the "testing error" on the test set as an approximation of the generalization error

Model Evaluation and Selection



Contents

- Introduction to Machine Learning
  - Supervised and Unsupervised Learning
  - Classification and Regression
- Linear Regression (Supervised Learning)
  - Model
  - Performance Evaluation
- **Classification (Supervised Learning)**
  - **How to Perform a Classification**
  - Classification Tree Model
- Clustering Method (Unsupervised Learning)
  - Objective
  - Similarity Measures
  - (Optional) Method 1: Hierarchical Clustering
  - (Optional) Method 2: K-Means Method (Clustering by Partitioning)
- Lab (Demo): Unsupervised Learning
- Assignment 5: Supervised Learning
- Assignment 6: In-Class Quiz

How to measure the accuracy of a classifier?

- Suppose we have already selected the training and test sets, and we have created a classifier based on the training set

We will measure the accuracy based on the test set

- Try to predict the class value of every tuple in the test set, and compare it against their actual ones (which are already stored in the test set)

Accuracy Measures

Classification accuracy

- The percentage of test set tuples that are correctly classified by the classifier
- A useful tool for analyzing how well the classifier can recognize tuples of different classes is the confusion matrix

c_{ij}: number of tuples from class i that are classied as class j by the classifier

Predicted class

accuracy of classifying "yes" tuples

Classes	Buy="yes"	Buy="no"	Total	Accuracy
Buy="yes"	6,954	46	7000	99.34
Buy="no"	412	2,588	3000	86.27
Total	7,366	2,634	10,000	95.42

Actual class

total accuracy

Consider a two-class problem and the confusion matrix below

- The positives refers to the tuples of the main class of interest (C_1)

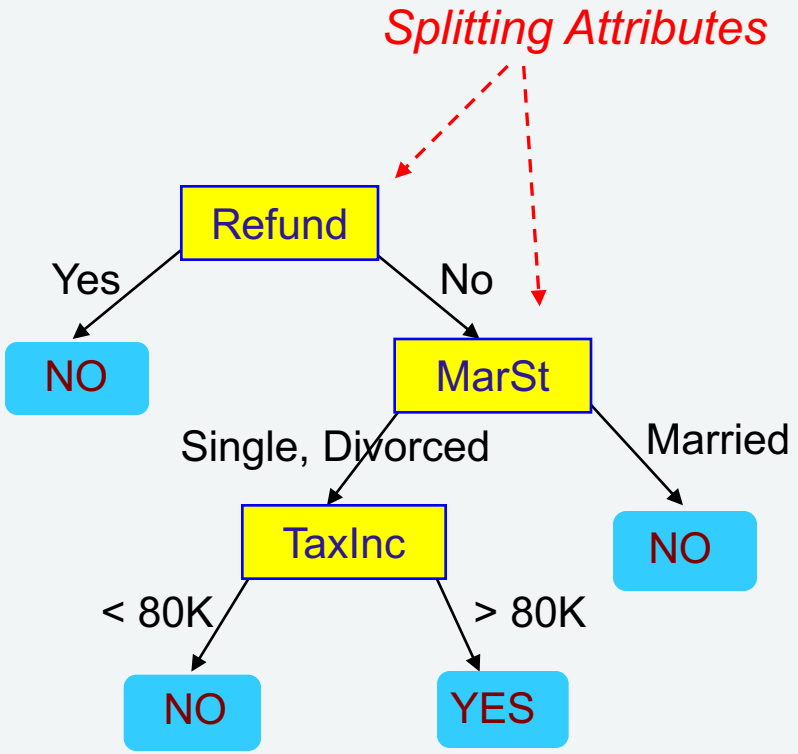
Predicted class

Actual class	Classes	C_1	C_2	Total
	C_1	true positives (TP)	false negatives (FN)	positives
	C_2	false positives (FP)	true negatives (TN)	negatives

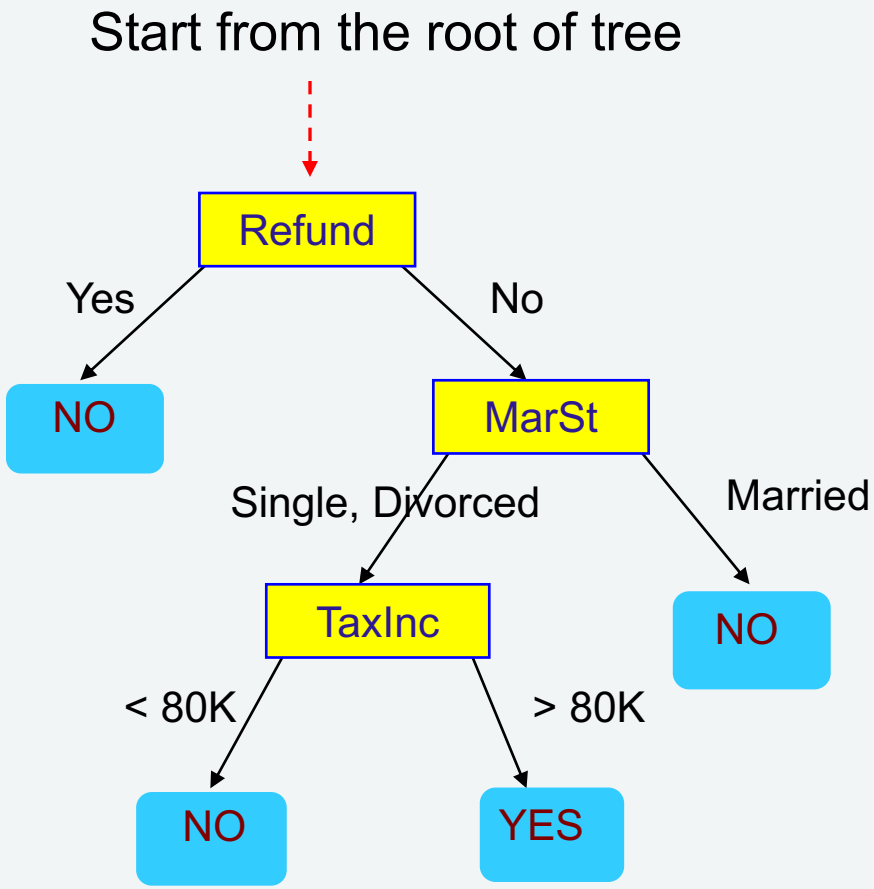
Contents

Model: Decision Tree

- Each internal node denotes a test on an attribute
- Each branch represents an outcome of the test
- Each leaf node holds a class label



Decision Tree: Prediction



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Contents

Goal of Cluster Analysis

Discover hidden structures in unlabeled data (unsupervised)

Clustering identifies a finite set of groups (*clusters*) $C_1, C_2 \cdots, C_k$ in the dataset such that:

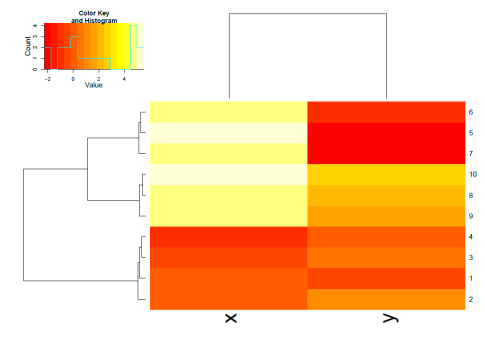
- Objects within the *same* cluster C_i shall be as similar as possible
- Objects of *different* clusters C_i, C_j $(i \neq j)$ shall be as dissimilar as possible

Application Examples

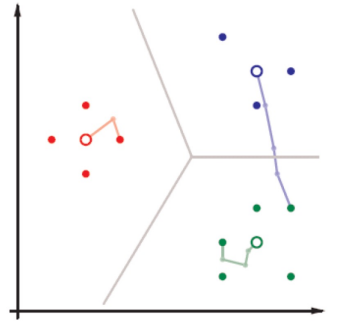
- **Customer segmentation**
  - Find groups of customers with similar behaviour; find customers with unusual behavior
- **Molecule search**
  - Find molecules with similar structure to already working ones
- **Anomaly detection**
  - Find unusual patterns in data from sensors monitoring mechanical engines
- **Determining user groups on the WWW**
  - *Clustering of activities in web-logs*
  - *Find groups of social media users with similar attitude.*
- **Structuring large sets of text documents**
  - *hierarchical clustering of the text documents*
- **Generating thematic maps from satellite images**
  - *clustering sets of raster images of the same area (feature vectors)*

Types of Clustering Approach

Linkage Based
e.g. Hierarchical Clustering



Clustering by Partitioning
e.g. k-Means



We will use those two approaches only

Contents

(Dis-)similarity Functions for Numeric Attributes

For two objects $\boldsymbol{x} = (x_1, x_2, \cdots, x_d)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_d)$:

- ***Minkowski-Distance*** ***(L_p-Metric)***

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \sqrt[p]{\sum_{i=1}^{d} |x_i - y_i|^p}$$

- ***Euclidean Distance*** ***($L_2 - \boldsymbol{p} = \mathbf{2}$)***

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

- ***Manhattan-Distance*** ***($L_1 - \boldsymbol{p} = \mathbf{1}$)***

$$d_M(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

- *Tschebyschew-Distance* ($L_\infty - p = \infty$)

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le i \le d} \{|x_i - y_i|\}$$

- *Cosine Distance*

$$d_C(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

- *Tanimoto Distance*

$$d_T(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}}$$

- *Pearson Distance*

Euclidean distance of z-score transformed $\boldsymbol{x}, \boldsymbol{y}$

Influence of Distance Function / Similarity

- Clustering vehicles:
  - red Ferrari
  - green Porsche
  - red Bobby car

A. Red Ferrari

B. Green Porche

C. Red Bobby car

- Distance function based on maximum speed (numeric distance function):
  - Cluster 1: Ferrari & Porsche
  - Cluster 2: Bobby car

The distance function affects the shape of the clusters

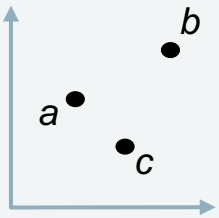
- Distance function based on color (nominal attributes):
  - Cluster 1: Ferrari and Bobby car
  - Cluster 2: Porsche

Similarities, Dissimilarities, and Distances

Given data points, how can we summarize how different they are? (Rather than summarizing the similarity)

- Dissimilarity metric – ***distance***
- Distance matrix – pairwise differences of all data points
- Distance $d_{i,j}$ (between i and j) calculated as the Euclidean distance

id	X	Y
a	1	2
b	3	3
c	2	1



$[d_{i,j}]$	a	b	c
a	0.00	2.23	1.41
b	2.23	0.00	2.23
c	1.41	2.23	0.00