December 2019

Data Preparation & Data Understanding

Avrahami Israeli

Advanced Analytics

Advanced Analytics group: Overview

Vision & Mission

Vision: Make Intel win the benefits of AI

#1: Optimize internal processes using AI

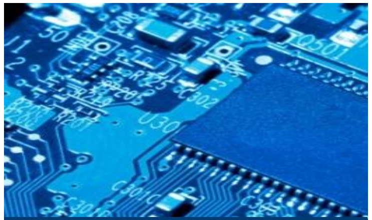
#2: Build competitive AI products

People

120

Machine Learning and Big Data Engineers

Portfolio: Breakthrough Technology that Scales

				
Design	Product Dev	Sales & Marketing	Healthcare	AI Products
Cut Time-to-Market, Find bugs early	Reduce Test Cost, Improve Quality, Higher Performance	Increase Sales, Focus Marketing	Wearables Analytics Platform, Parkinson's Monitoring	IoT Analytics, ML Bench, DL optimization

Self Intro

- Few words about myself
- [NASLAB]
- [Reddit]
- [r/place]
- More general – how do we represent communities


WHAT IS
DATA PREPROCESSING?!

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Imbalanced Data

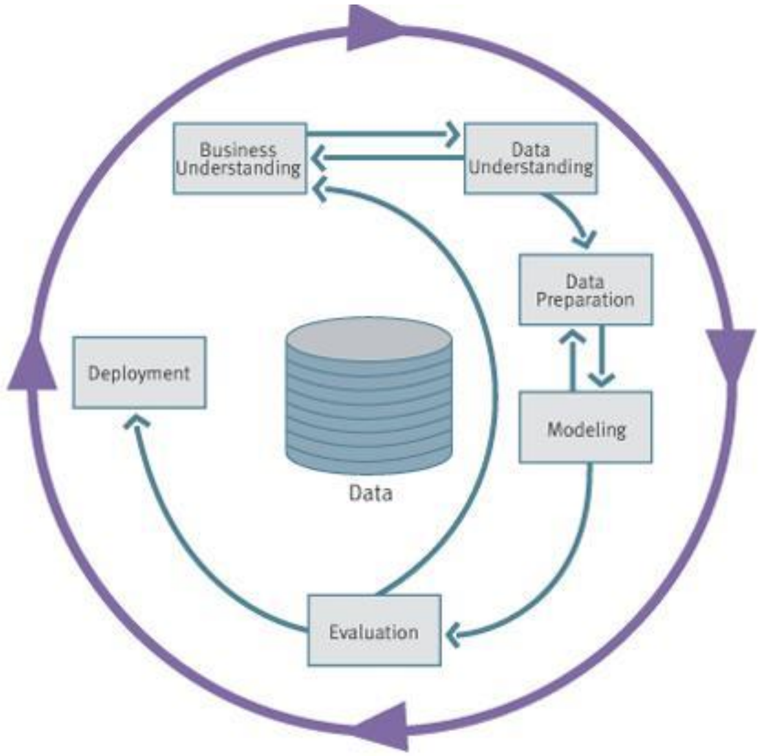
Agenda

1. **Introduction**
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Imbalanced data

Introduction (1) - CRISP-DM

CRISP-DM breaks the process of data mining into six major phases

1. Business Understanding
2. **Data Understanding**
3. **Data Preparation**
4. Modeling
5. Evaluation
6. Deployment



The sequence of the phases is not strict and moving

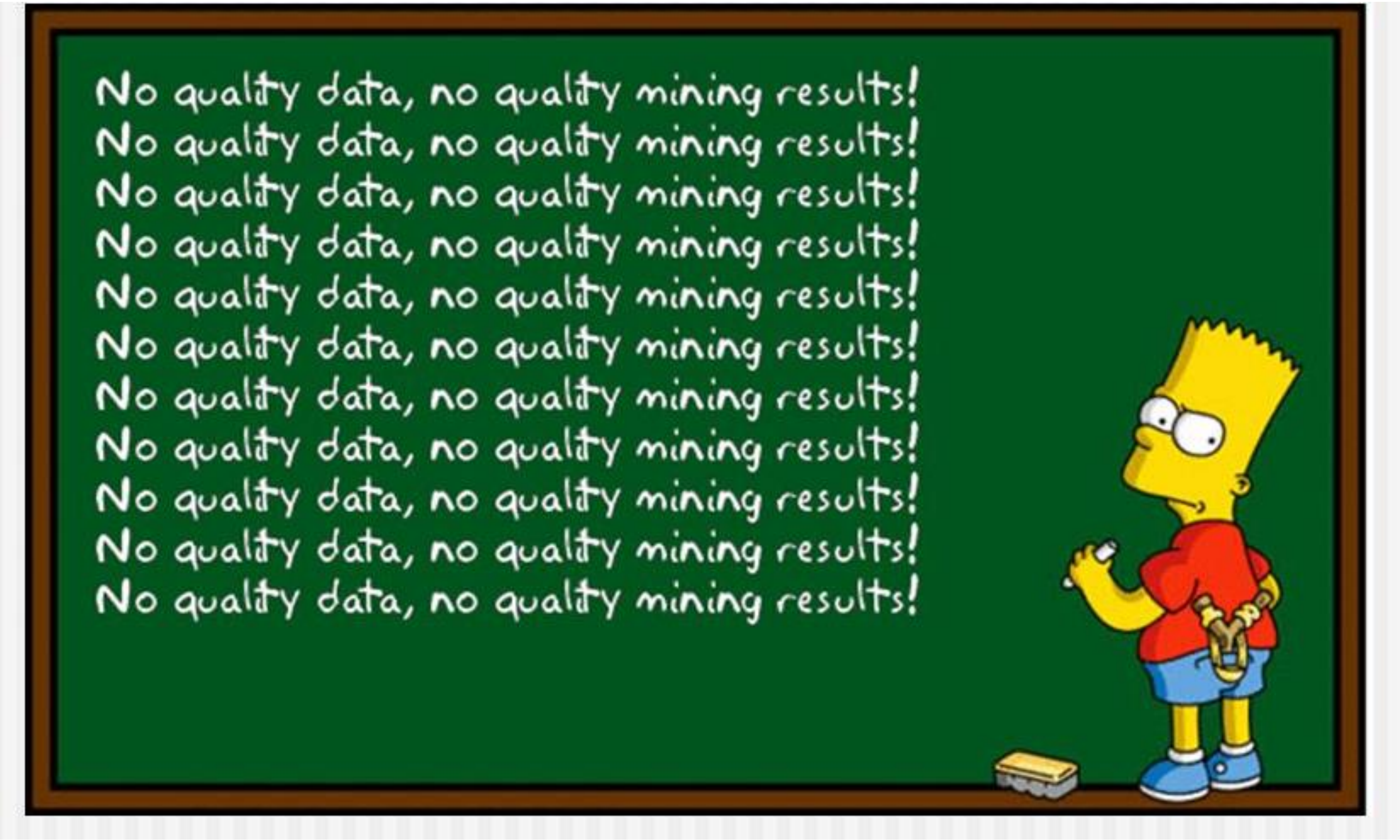
back and forth between different phases may be required

Introduction (2)

- Today - the first part of the CRISP-DM (and most important one!)
- What is NOT going to be covered here
- Statistical session VS ML 'hard core' session
- Not all topics involve heavy theoretical material

Introduction (3)

Why is data preprocessing important?

- 

Introduction (4)

Major tasks in data preparation

- Data cleaning (e.g. missing values ,outliers)
- Data transformation (e.g. normalization)
- Feature engineering
- Data discretization
- Data reduction

Agenda

1. Introduction
2. **Data types**
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Imbalanced data

Data types (1)

Type	Example
I. Numerical data (double)	Income (e.g. 650.34)
II. Numerical data (int)	# of children (e.g. 4)
III. Boolean	Gender (e.g. male)
IV. Categorical data	Colors (e.g. green)
V. Ordinal data	Satisfaction (e.g. 2/5)
VI. String	Description (e.g. “Bad”)
VII. Others	Comments

Data types (2)

Why is it so important ??

- A-normal input for modeling
- Distance measures
- Models results are based on this input

Ideal - Subset of HSW22 MIDAS Operation unit level data.txt

Console ~/

```
> head(car.test.frame)
                 Price   Country Reliability Mileage  Type Weight Disp.  HP
Eagle Summit 4    8895       USA           4      33 Small   2560    97 113
Ford Escort   4   7402       USA           2      33 Small   2345   114  90
Ford Festiva 4    6319     Korea           4      37 Small   1845    81  63
Honda Civic 4     6635 Japan/USA           5      32 Small   2260    91  92
Mazda Protege 4   6599     Japan           5      32 Small   2440   113 103
Mercury Tracer 4  8672    Mexico           4      26 Small   2285    97  82
> sapply(car.test.frame, class)
      Price     Country Reliability     Mileage        Type      Weight       Disp.          HP
  "integer"    "factor"   "integer"   "integer"    "factor"   "integer"   "integer"   "integer"
> |
```

For Help, press F1 | Table: Subset of HSW22 MIDAS C | Variables: 7 | Samples: 101085

Agenda

1. Introduction
2. Data types
3. **Distance measures**
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Imbalanced data

Distance Measures (Metric)

- A metric (or distance) is a function $d\colon X \times X \rightarrow [0, \infty)$

where for all $\mathrm{x}, \mathrm{y}, \mathrm{z} \in X$, the following conditions are satisfied:

1. $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$ — Non-negativity
2. $d(x, y) = d(y, x)$ — Symmetry
3. $d(x, z) \leq d(x, y) + d(y, z)$ — Triangle inequality

Distance Measures (Metric)

- Euclidean distance (L2):

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

- Manhattan distance (L1):

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

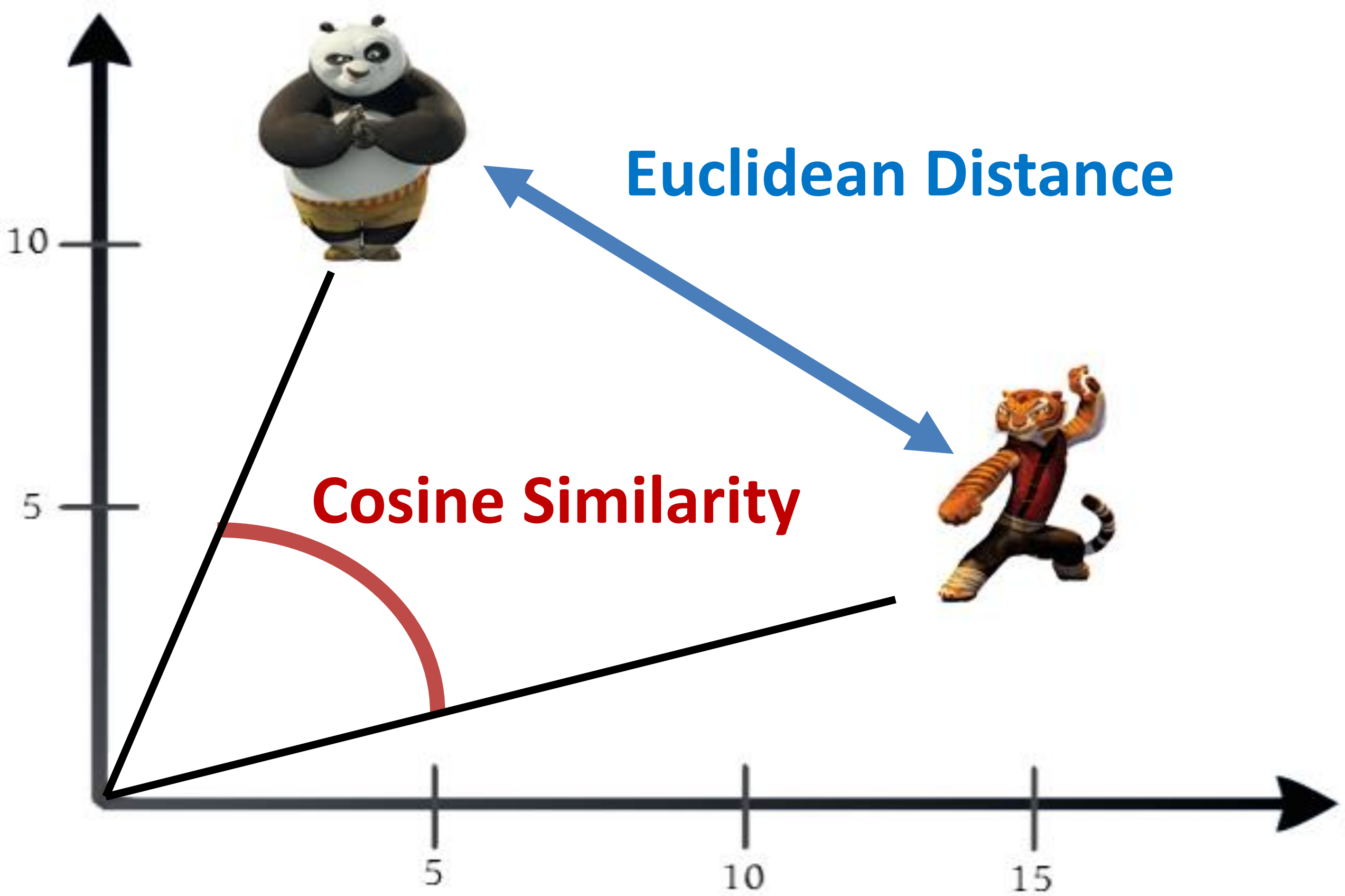
- Cosine similarity:

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i \, y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}} = \frac{\boldsymbol{X} \cdot \boldsymbol{Y}}{\|\boldsymbol{X}\|_2 \|\boldsymbol{Y}\|_2}$$

Intel Confidential

Euclidean vs Manhattan distance

Euclidean vs Cosine



Scikit-Learn

`metrics.pairwise.pairwise_distances` (X[, Y, …])	Compute the distance matrix from a vector array X and optional Y.
`metrics.pairwise.pairwise_kernels` (X[, Y, …])	Compute the kernel between arrays X and optional array Y.
`metrics.pairwise.polynomial_kernel` (X[, Y, …])	Compute the polynomial kernel between X and Y:
`metrics.pairwise.rbf_kernel` (X[, Y, gamma])	Compute the rbf (gaussian) kernel between X and Y:
`metrics.pairwise.sigmoid_kernel` (X[, Y, …])	Compute the sigmoid kernel between X and Y:
`metrics.pairwise.paired_euclidean_distances` (X, Y)	Computes the paired euclidean distances between X and Y
`metrics.pairwise.paired_manhattan_distances` (X, Y)	Compute the L1 distances between the vectors in X and Y.
`metrics.pairwise.paired_cosine_distances` (X, Y)	Computes the paired cosine distances between X and Y
`metrics.pairwise.paired_distances` (X, Y[, metric])	Computes the paired distances between X and Y.
`metrics.pairwise_distances` (X[, Y, metric, …])	Compute the distance matrix from a vector array X and optional Y.
`metrics.pairwise_distances_argmin` (X, Y[, …])	Compute minimum distances between one point and a set of points.
`metrics.pairwise_distances_argmin_min` (X, Y)	Compute minimum distances

Agenda

1. Introduction
2. Data types
3. Distance measures
4. **Correlation and Mutual information**
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Imbalanced data

Correlation

- Correlation refers to **any** of a broad class of statistical relationships involving dependence
- How is this related to our discussion ?
- Common correlations:

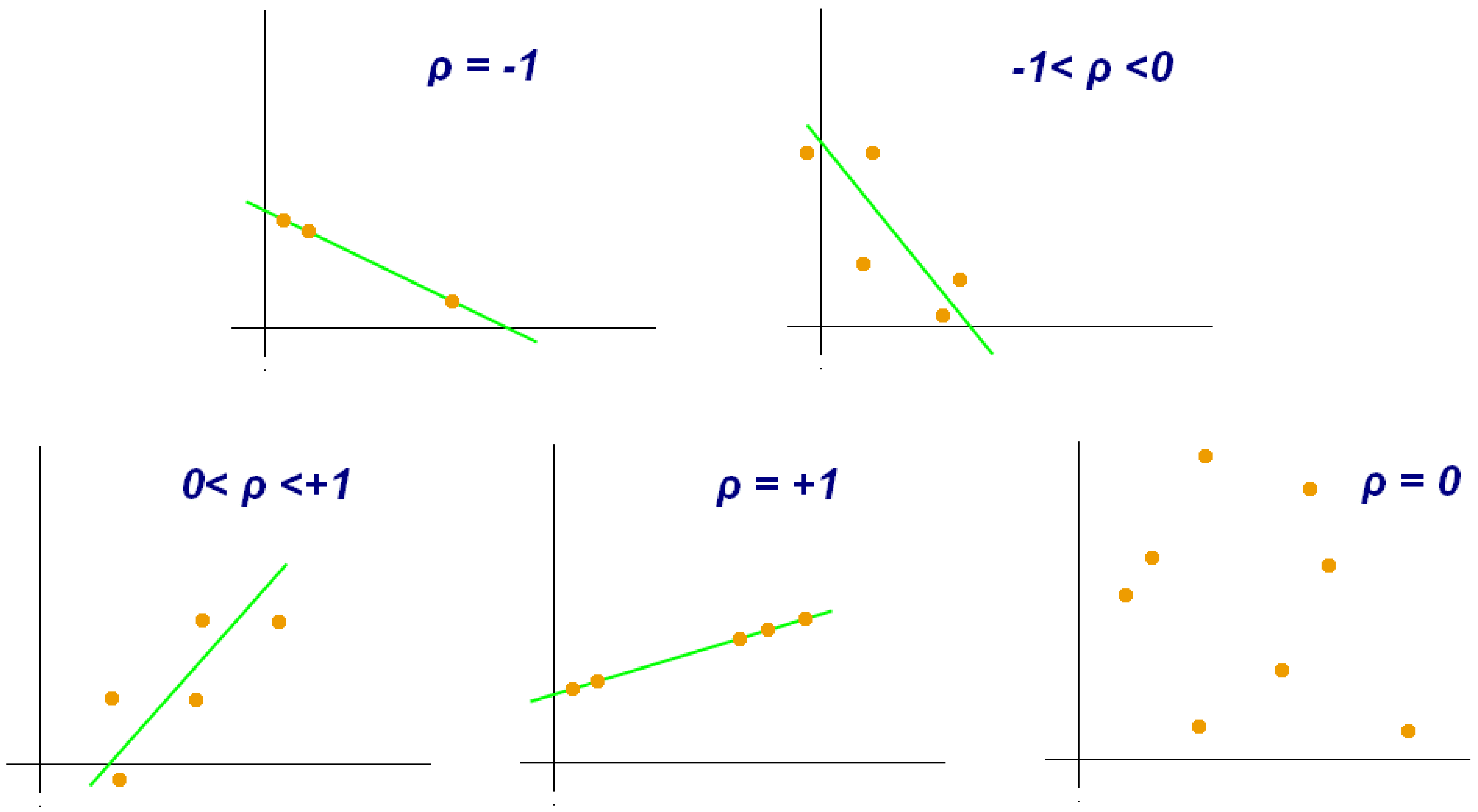
1. Pearson Correlation – measures the degree of **linear dependence** between two variables
2. Spearman correlation – measures how well the relationship between two variables can be described using a **monotonic function**
3. Kendall's tau correlation – measures the **“ordering” dependency** between two variables

Pearson Correlation

- A measure of the linear relation between two variables *X* and *Y*.
- It has a value between +1 and −1:
  - 1 is total positive linear correlation
  - 0 is no linear correlation
  - −1 is total negative linear correlation
- Definition:

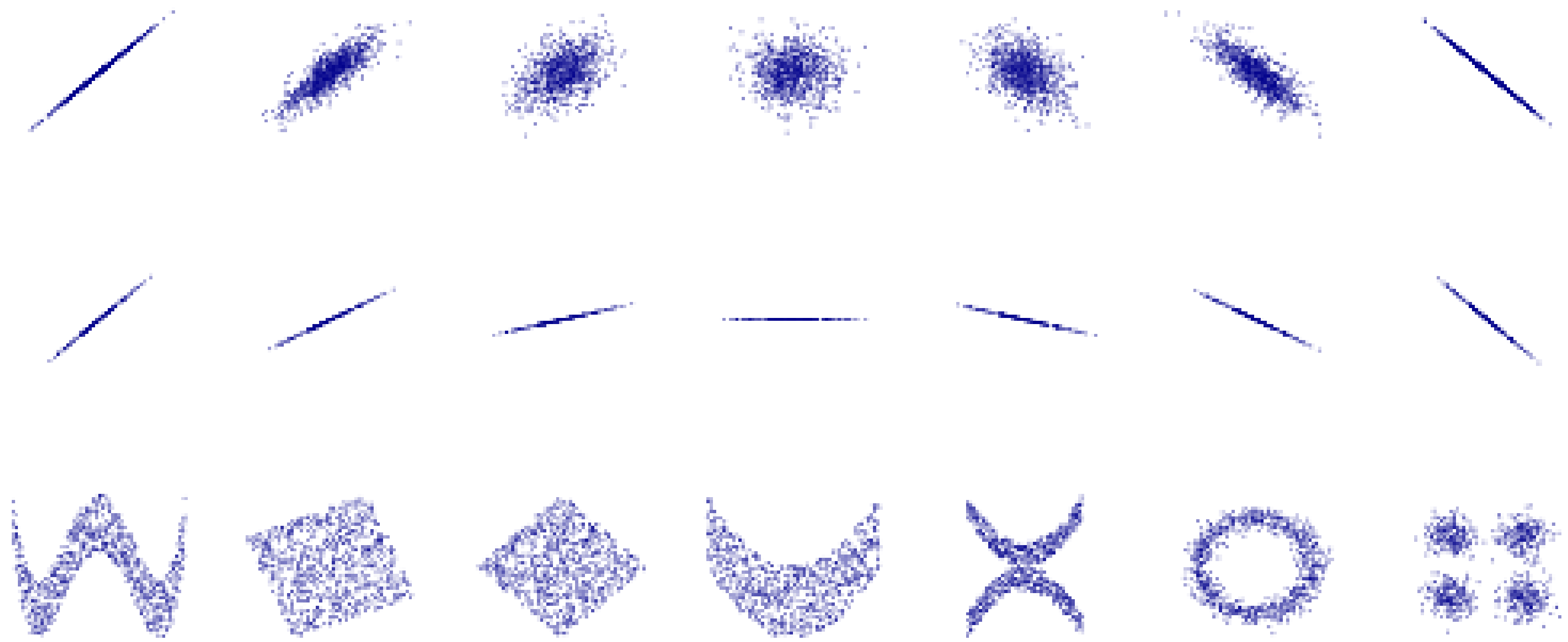
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

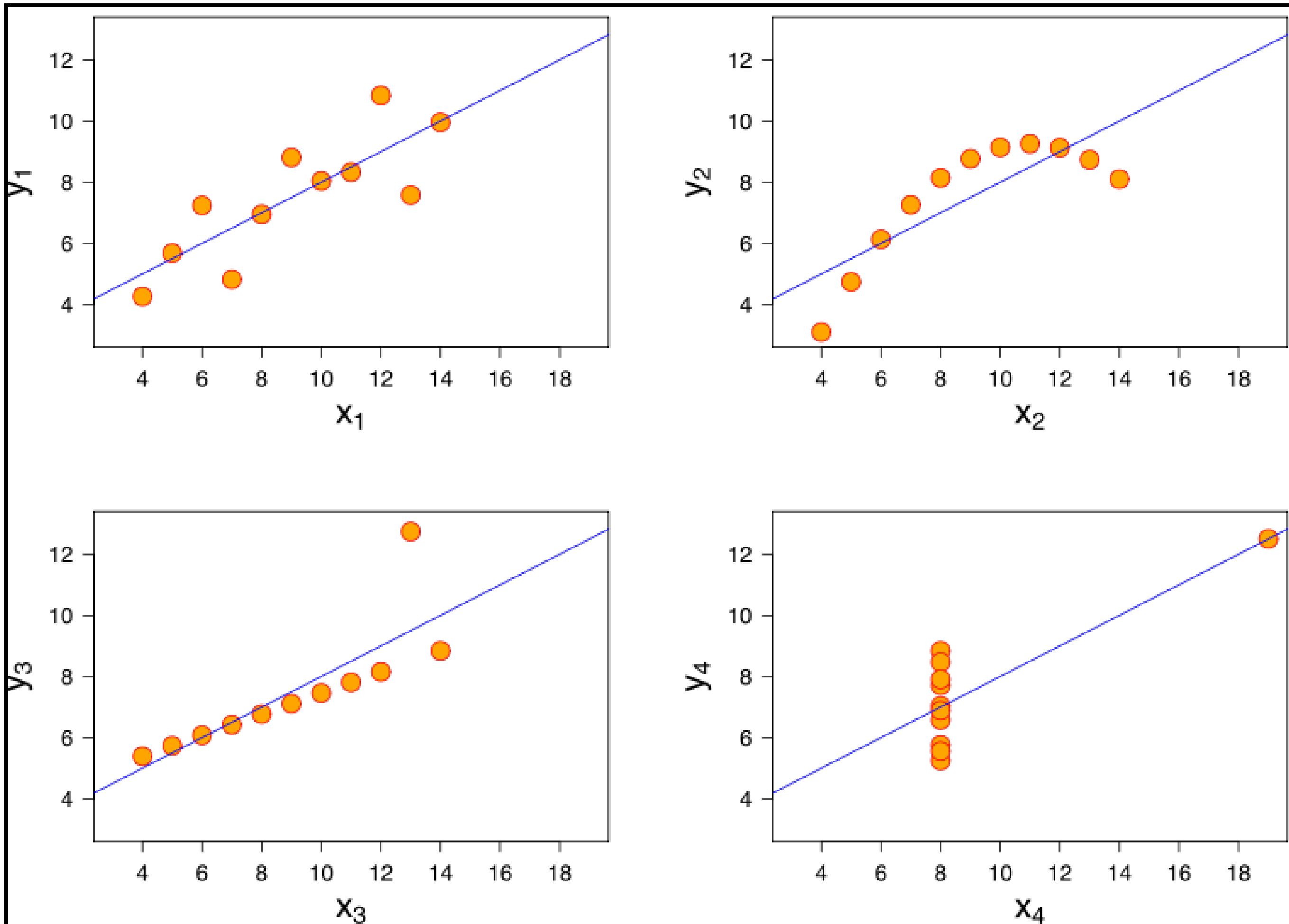
- $\rho_{X,Y} = 0 \overset{??}{\Leftrightarrow} Uncorrelated \overset{??}{\cancel{\Leftrightarrow}} Independent$



ρ = -1
-1< ρ <0
0< ρ <+1
ρ = +1
ρ = 0

Pearson Correlation - Examples





12
10
8
6
4
4 6 8 10 12 14 16 18
y_1
x_1
y_2
x_2
y_3
x_3
y_4
x_4

Spearman Correlation

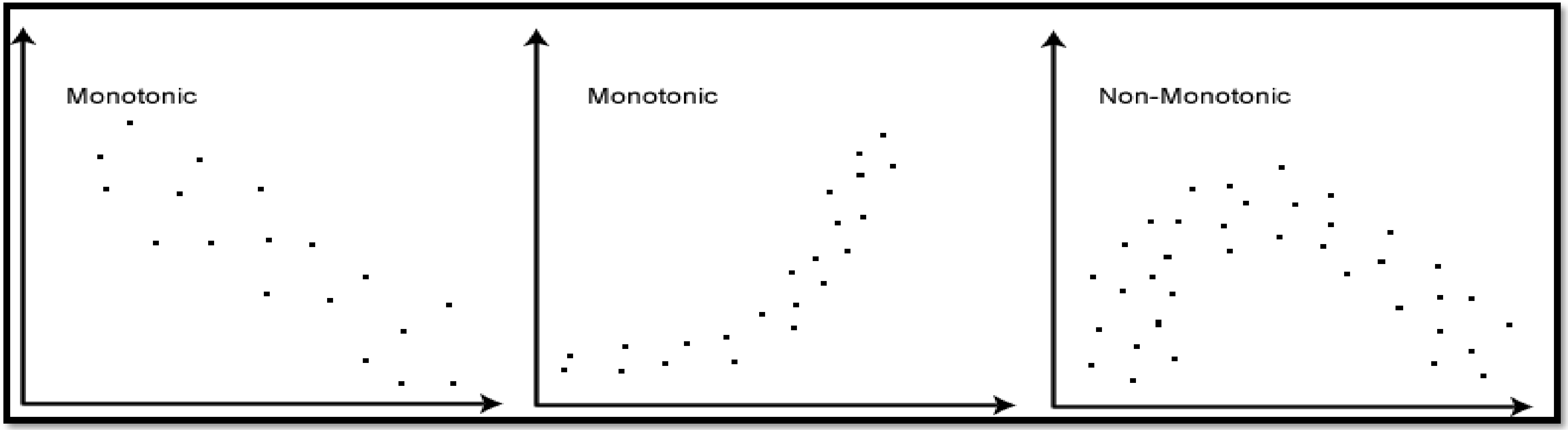
- Measures the **monotonic** behavior relationship between two features
- The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables
- Definition:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

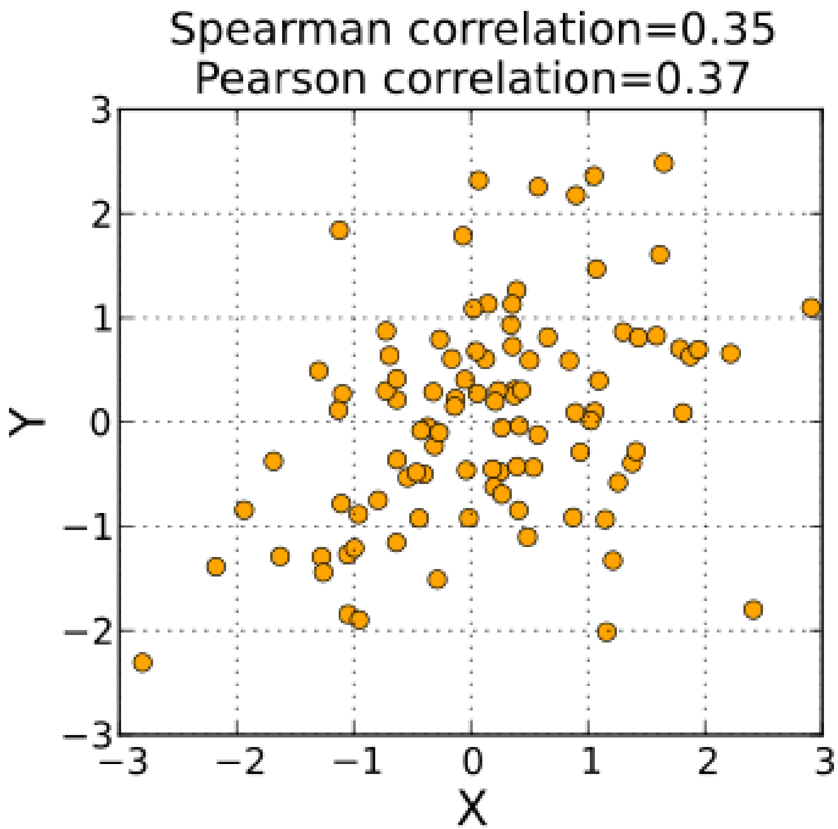
BUT – X, Y are the **ranked** features

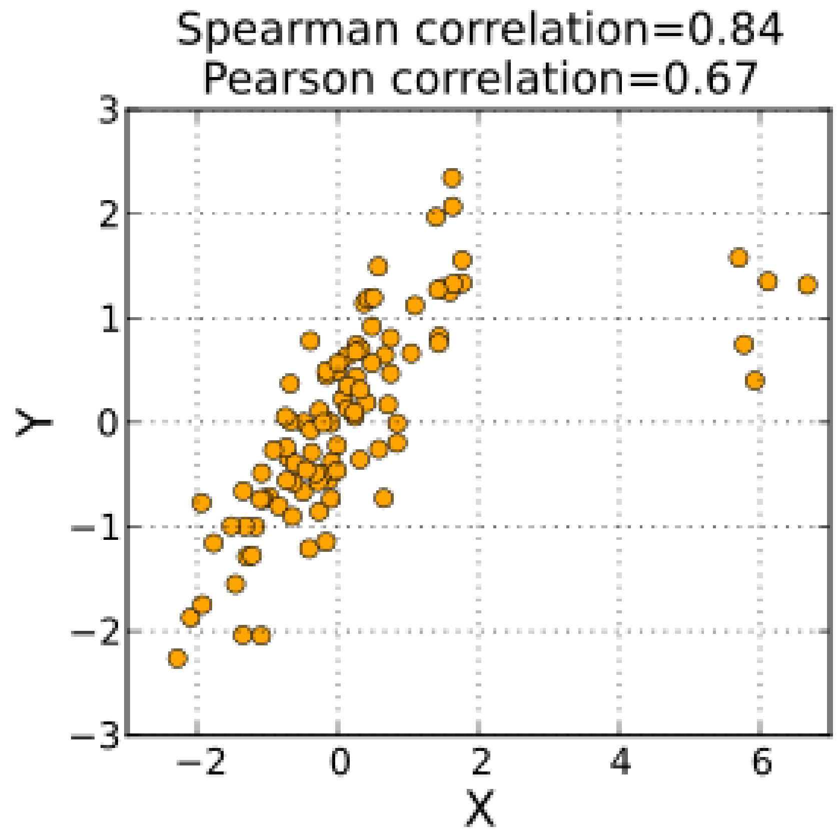
- Range: [-1,1] (what do the -1,0,1 values mean?)
- Advantages/disadvantages comparing to Pearson correlation

Montonic / Non-monotonic



Pearson vs Spearman Correlation

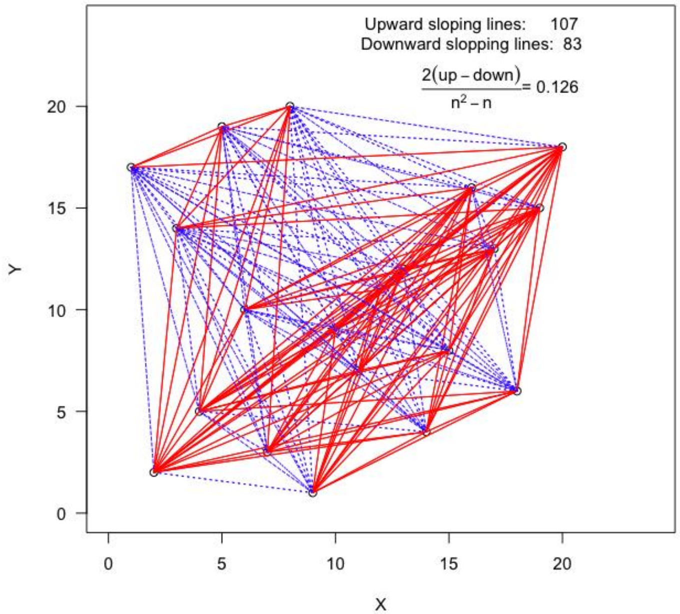




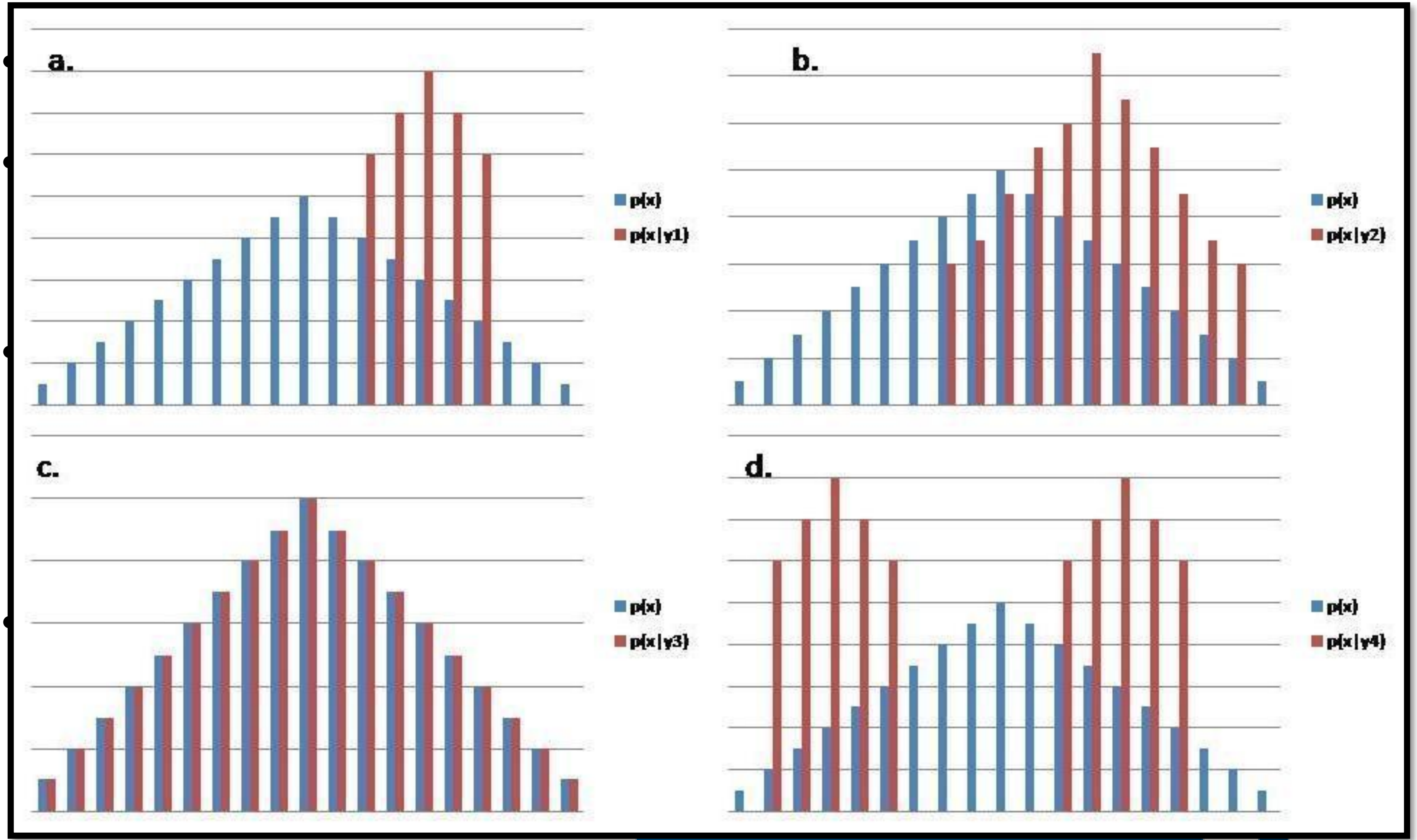
Kendall's Tau Correlation

- Measures the pair ordering correlation between two features
- Definition:

$$\tau_{X,Y} = \frac{(\#\ of\ concordant\ pairs) - (\#\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}$$



Mutual Information



Entropy

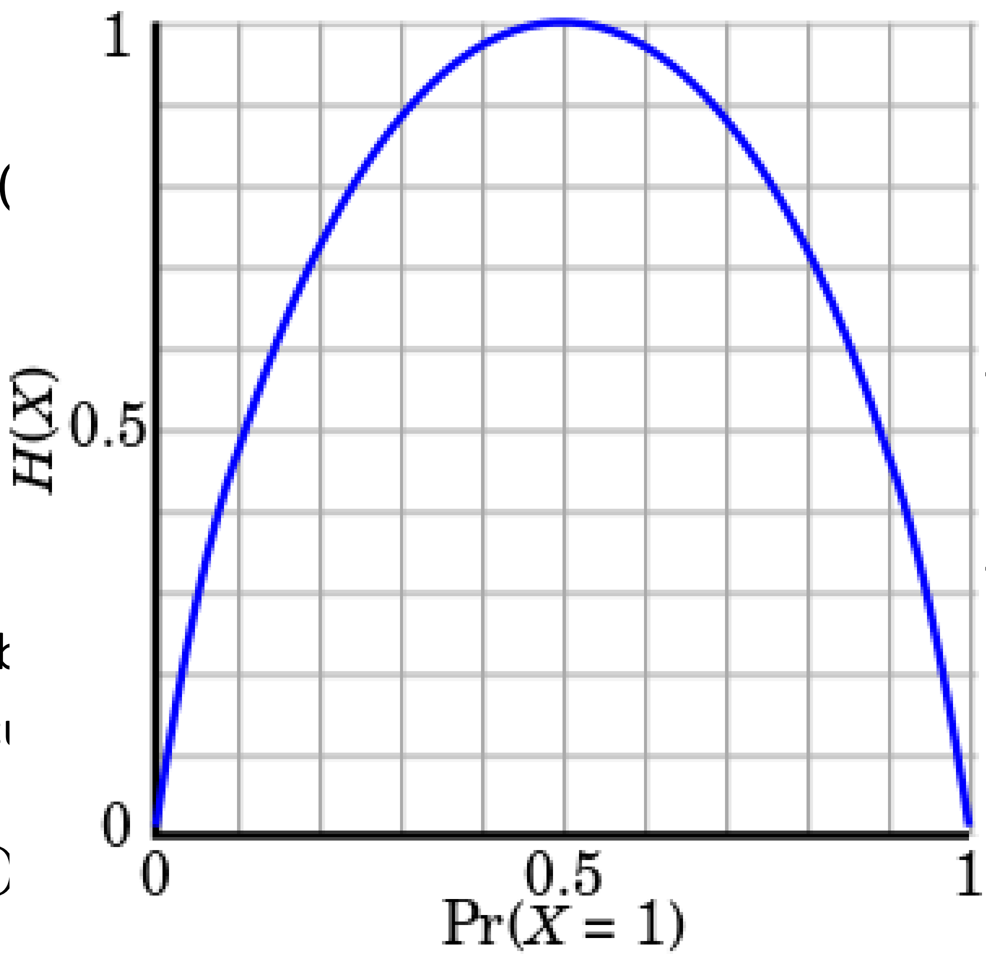
- Measures the uncertainty in a random variable
- Definition:

$H($ … $(P(x_i))$

)

- Example:
  - Consider … ning up heads or tails.
  - This can …
  - Let's calc…

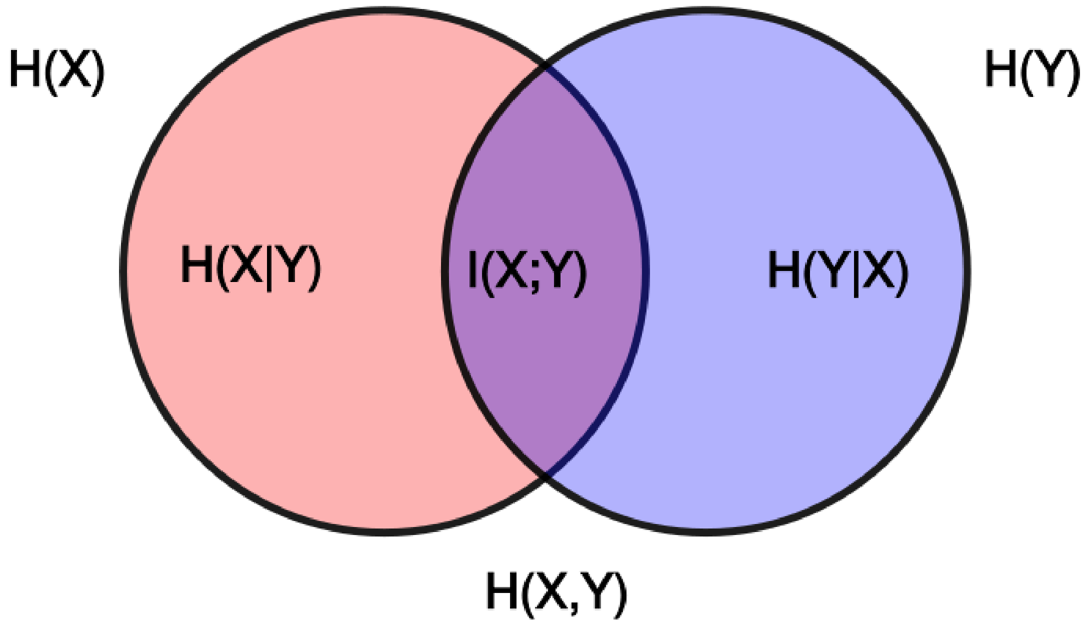
$H(X)$ … 1



Back to Mutual Information

- Reminder:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

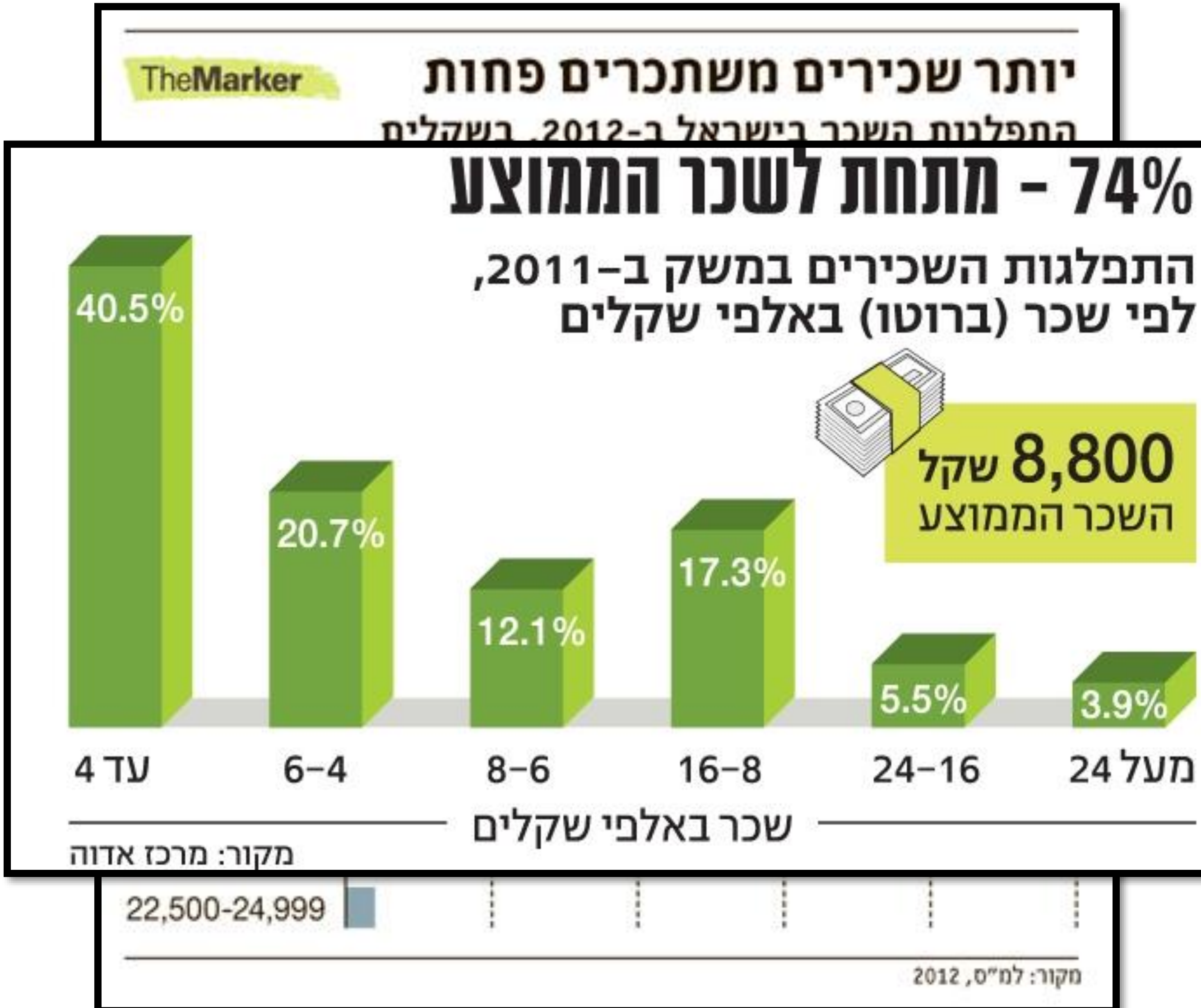


Correlation and Mutual Information - Summary

- Pearson correlation assumes a linear relationship, others don't
- Correlation can be calculated directly from a data sample, whereas mutual information requires knowledge of the distribution
  - How do we estimate the distribution?
- Various correlation measures:
  - Care about the actual values? If so – Pearson
  - Care only about the **rank** of value? If so – Spearman
  - Care about the **order** of the value? If so – Kendell's tau

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. **Data distribution**
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data



TheMarker
יותר שכירים משתכרים פחות
התפלגות השכר בישראל ב-2012, בשקלים
74% - מתחת לשכר הממוצע
התפלגות השכירים במשק ב–2011,
לפי שכר (ברוטו) באלפי שקלים
8,800 שקל
השכר הממוצע
40.5%
20.7%
12.1%
17.3%
5.5%
3.9%
עד 4
6–4
8–6
16–8
24–16
מעל 24
שכר באלפי שקלים
מקור: מרכז אדוה
22,500-24,999
מקור: למ"ס, 2012

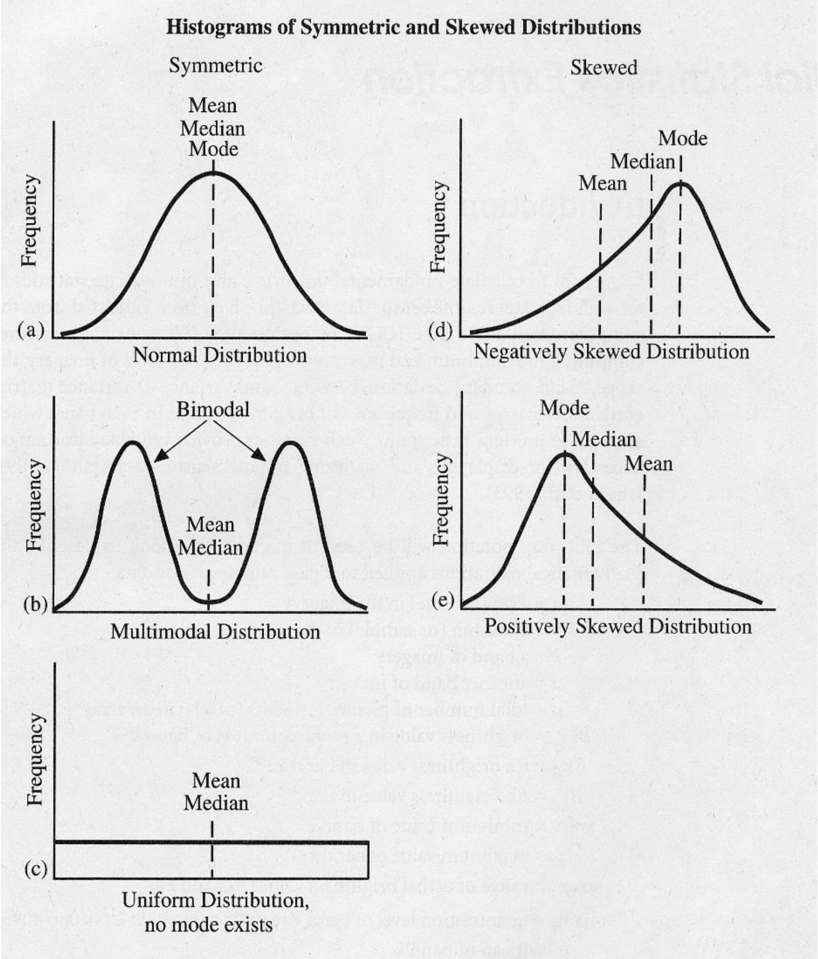
Basic measures (1)

Many statistical tests assume values are normally distributed, but this is not always the case

- Examine data prior to processing

Comparing Mean, Median & Mode

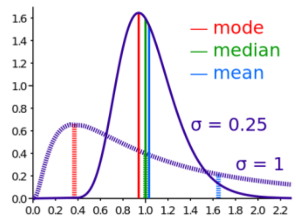
- Mode (שכיח)
  - Good for nominal variables
  - Quick and easy
- Median (חציון)
  - Robust central tendency statistics
    - Less sensitive to outliers and extreme values
  - Good for "bad" distributions
- Mean (ממוצע)
  - Most commonly used statistic for central tendency
    - Generally preferred except for "bad" distribution
  - Based on all data in the distribution
  - Used for inference as well as description
    - best estimator of the parameter

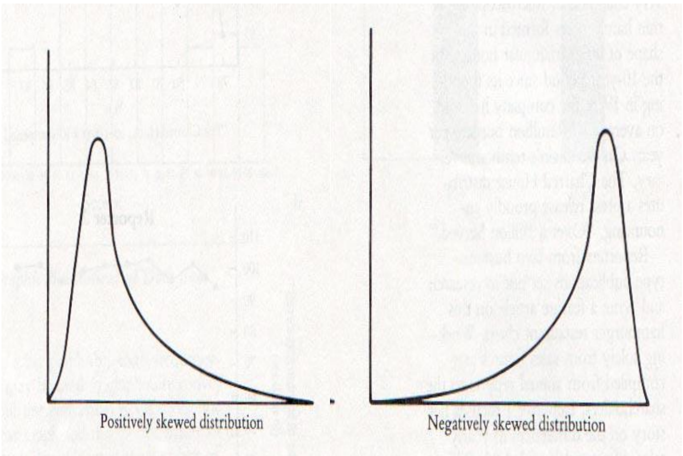


Basic measures (2)

Skewness *(tails)*

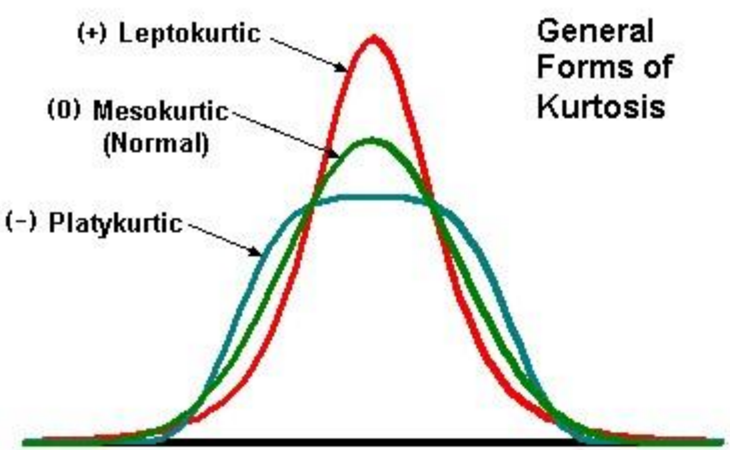
- Skewness is a measure of the asymmetry of the probability distribution
  - $\alpha_3 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3} = \frac{\mu_3}{\sigma_3}$
    - Right skew - $\alpha_3 > 0$
    - Left skew - $\alpha_3 < 0$
    - Symmetric - $\alpha_3 = 0$





Kurtosis *(shoulders, heavy tail)*

- Kurtosis is the degree of peakedness of a distribution relative to a normal distribution
  - $\alpha_4 = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} - 3 = \frac{\mu_4}{\sigma_4} - 3$
    - A normal distribution is a *mesokurtic* distribution
    - A pure *leptokurtic* distribution has a higher peak than the normal distribution and has heavier tails.
    - A pure *platykurtic* distribution has a lower peak than a normal distribution and lighter tails.



Intel Confidential

Data distribution (1)

Normal (Gaussian) Distribution

- $X \sim N(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- Z-score
  - $z = \frac{x-\mu}{\sigma}$
  - The distance of a value from the mean, measured in standard deviations

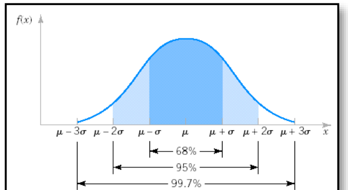
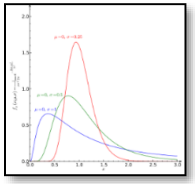


Figure 3-13 Probabilities associated with a normal distribution.

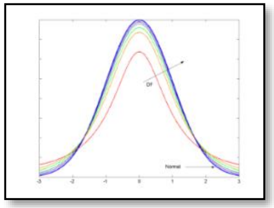
Log-normal Distribution

- $X \sim \ln N(\mu, \sigma^2)$, $x = e^z$, $z \sim N(\mu, \sigma^2)$
  - $f(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}$
- Used to model a variable which is a product of positive i.i.d vars,
  - A compound return from a sequence of many trades
  - Measures of size of living tissue



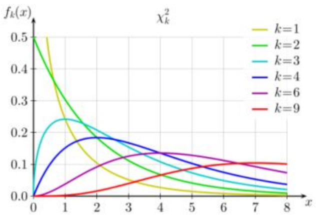
Student's t-Distribution (Gosset 1908)

- Sampling distrib. (i.i.d measures) of
  - $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$
- Approaches the Gaussian distrib. when
  - $n > 30$ or $s = \sigma$
- Used for
  - Test the diff. between two sample means
  - Inference when (μ, σ^2) are unknown



The χ^2 Distribution with k D.F

- $X \sim \chi_k^2$, $\chi_k^2 = \sum_{i=1}^{k} z_i^2$, $Z \sim N(0,1)$
  - $f(x; k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^k \Gamma\left(\frac{k}{2}\right)}$
- Heavily used in statistics
  - Estimating variance
  - Goodness-of-fit test



Data distribution (2)

- Bernoulli Distribution
  - Bernoulli trial
    - A trial with only two possible outcomes
  - Bernoulli Distribution
    - Represents success/failure (e.g. accuracy of prediction)
    - $X \in [0,1] \sim Bernoulli(p)$
  - $f(x;p) = p^x(1-p)^x$

  ($\Pr[X=1] = p$)

- Binomial distribution
  - Number of success in *n* independent trials
  - $K \sim B(p,n), \quad K = \sum_{i=1}^{n} z_i\ , \ Z \sim Bernoulli(p)$
    - $f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$

If *n* is large, then:
$Z \sim N(np, np(1-p))$
is a good approximation
for $K \sim B(p,n)$

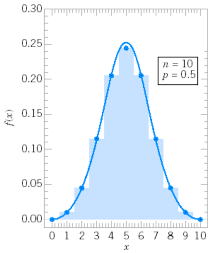


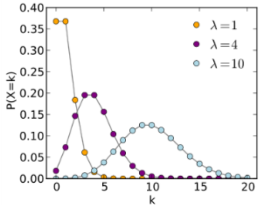
Figure 3-36 Normal approximation to the binomial distribution.

- Multinomial Distribution
  - Categorical Distribution
    - A trial with *k* possible outcomes
    - $f(x_1,\dots,x_k;p_1,\dots,p_k) = \prod_{i=1}^{k} p_i^{x_i}$

      where $x_i \in \{0,1\}$ and $\sum_{i=1}^{k} p_i = 1,\ \ p_i \in [0,1]$
  - Multinomial Distribution
    - Number of occurrences of *k* categories in *n* independent trials
    - $f(n_1,\dots,n_k;n,p_1,\dots,p_k) = \frac{n!}{n_1!\cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$

      where $n_i \in \mathbb{N}, \sum_{i=1}^{k} n_i = n$

- Poisson Distribution
  - Number of events occurring within a fixed time interval (or space)
    - λ , the shape param., indicates the average number of events in the given time interval
  - $K \sim Pois(\lambda),\ K \in \ \mathbb{N},\ \lambda > 0$
    - $f(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$



  - If λ is large, then $Z \sim N(\lambda,\lambda)$ is a good approximation for $K \sim Pois(\lambda)$

Testing the data distribution

Parametric Hypothesis and general test

- Statistical tests to check the mean/variance
- Q-Q plot

Testing a general distributions

- Shapiro's test for normality
- Kolmogorov–Smirnov test
- Cramér–von Mises criterion
- Anderson–Darling test

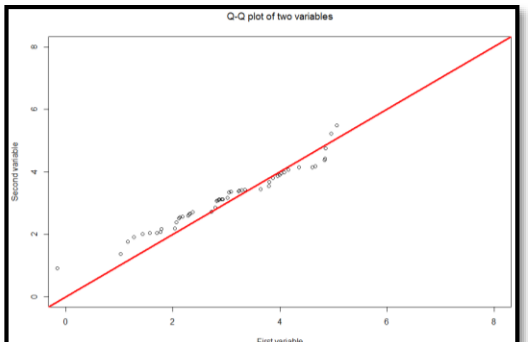
Testing the data distribution

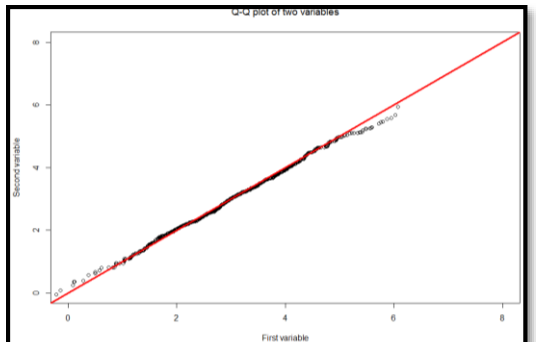
Data comparisons you are making	Data are normally distributed	Data are not normally-distributed, or are ranks or scores	Data are Binomial (Possess 2 possible values)
Compare one set of data to a hypothetical value	One-sample t-test	Wilcoxon test	χ^2 test
Compare two sets of independently-collected (unpaired) data	Unpaired t-test	Mann-Whitney test	χ^2 test or Fisher test
Compare two sets of data from the same subjects under different circumstances (paired)	Paired t-test	Wilcoxon test	McNemar's test
Compare three or more sets of data	One-way ANOVA	Kruskal-Wallis test	χ^2 test
Look for a relationship between two variables	Pearson Correlation coefficient	Spearman correlation coefficient	Contingency Correlation coefficients
Look for a linear relationship between two variables	Linear regression	Nonparametric linear regression	Simple logistic regression
Look for a non-linear relationship between two variables	Non-linear regression	Nonparametric non-linear regression	

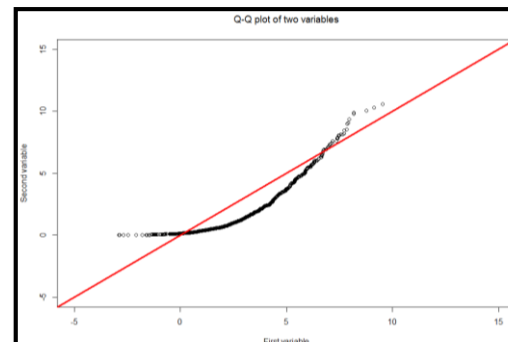
Let's see some examples how to run these tests

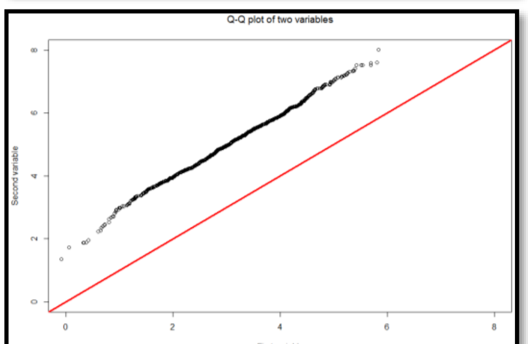
Q-Q plot

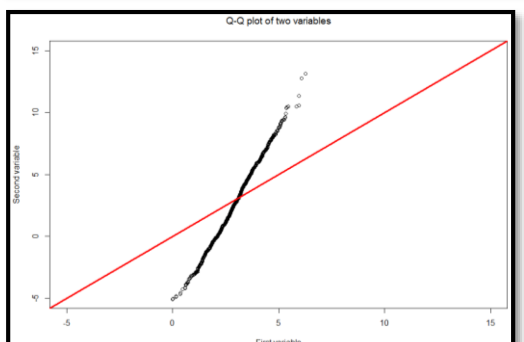
- A plot of the quantiles of the first data set against the quantiles of the second data set
- Data sets sizes don’t have to be equal
- The **greater** the departure from the 45 deg. reference line, the **greater** the evidence for the conclusion that the two data sets have come from populations with **different** distributions











Kolmogorov–Smirnov test

- A non-parametric test for the **equality** of continuous, one-dimensional probability distribution
- Can be applied to test a dataset distribution against a **known distribution** OR against **another dataset distribution**

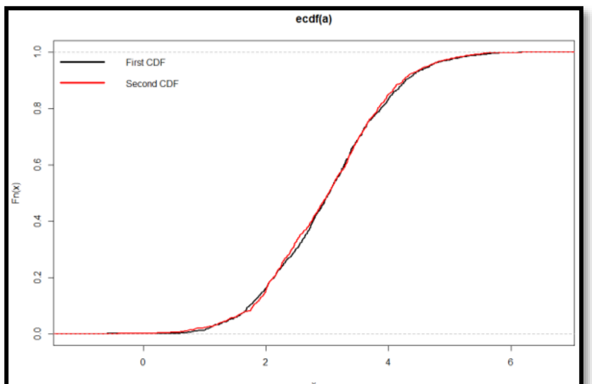
H_0: The data follow a specified distribution

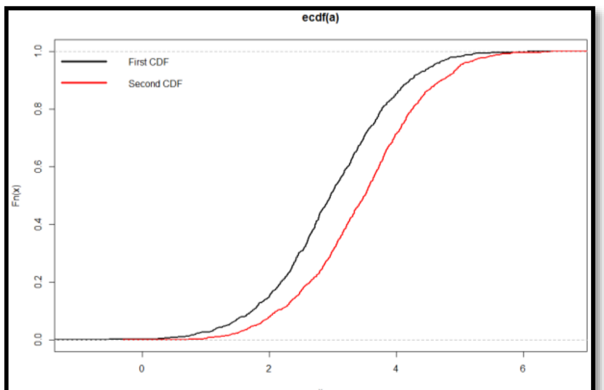
H_1: The data does not follow a specified distribution

- The K-S statistics is defined as:

$$\boxed{D_n = \sup_x |F_n(x) - F(x)|}$$

- Let's have an example in R





Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. **Missing values**
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Missing values handling (1)

- We don't always need to handle missing value

- But when we do…

- Any ideas?

Missing values handling (2)

- Ignore the entire tuple/feature

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
~~Hyundai Sonata 4~~	~~9999~~	~~Korea~~	~~NA~~	~~23~~	~~Medium~~	~~2885~~	~~143~~	~~110~~
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
~~Nissan Maxima V6~~	~~17899~~	~~Japan~~	~~5~~	~~22~~	~~NA~~	~~3200~~	~~180~~	~~160~~
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
~~Oldsmobile Cutlass Supreme V6~~	~~14495~~	~~NA~~	~~1~~	~~21~~	~~Medium~~	~~3220~~	~~189~~	~~135~~
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
~~Chevrolet Lumina APV V6~~	~~13995~~	~~USA~~	~~NA~~	~~18~~	~~Van~~	~~3195~~	~~151~~	~~110~~
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

  - Simple
  - Reduces statistical power, estimation might be biased if data is missing on purpose.

Missing values handling (3)

- Analyze only cases in which the relevant variables are present (Pairwise deletion)

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	~~NA~~	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	~~NA~~	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	~~NA~~	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	~~NA~~	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

  - Uses all possible information with each analysis

Missing values handling (4)

- Use attribute **mean**, **median** or **mode** to complete the missing data

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

Mean (Reliability): (5+5+2+1+3+3+1+3+3)/9 = **2.88**
Median (Reliability): 1 1 2 3 **3** 3 3 5 5
Mode (Country): **USA = 6**, Japan = 3, Korea = 1.

Intel Confidential

Missing values handling(5)

- Use attribute mean, median or mode to complete the missing data – **restricted to a class**

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP	Class
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110	A
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158	A
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160	A
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110	A
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135	B
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190	B
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165	B
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170	B
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150	C
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110	C
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150	C

Class A.**Mean** (Reliability): (5+5+2)/3 = **4**

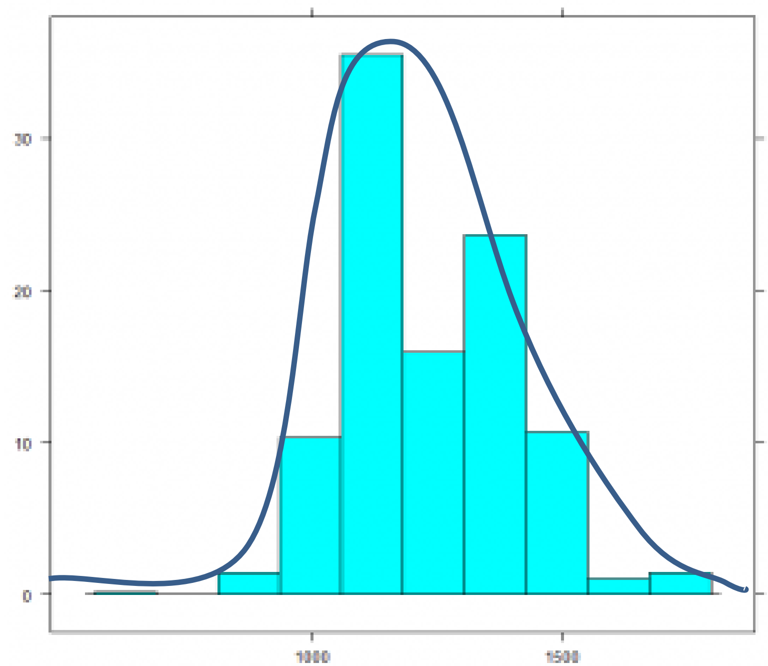
Class A.**Median** (Reliability): 2 **5** 5

Class B.**Mode** (Country): **USA = 2**, Japan = 1

Intel Confidential

Missing values handling (6)

- Sampling
  - If distribution is known, sample from it
  - Else, estimate distribution from data



Missing values handling (7)

- Use global closest fit to K nearest neighbors (take the value from the closest tuple.

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

- If K > 1, you can use either mean, median, mode or sampling to select the best fit.

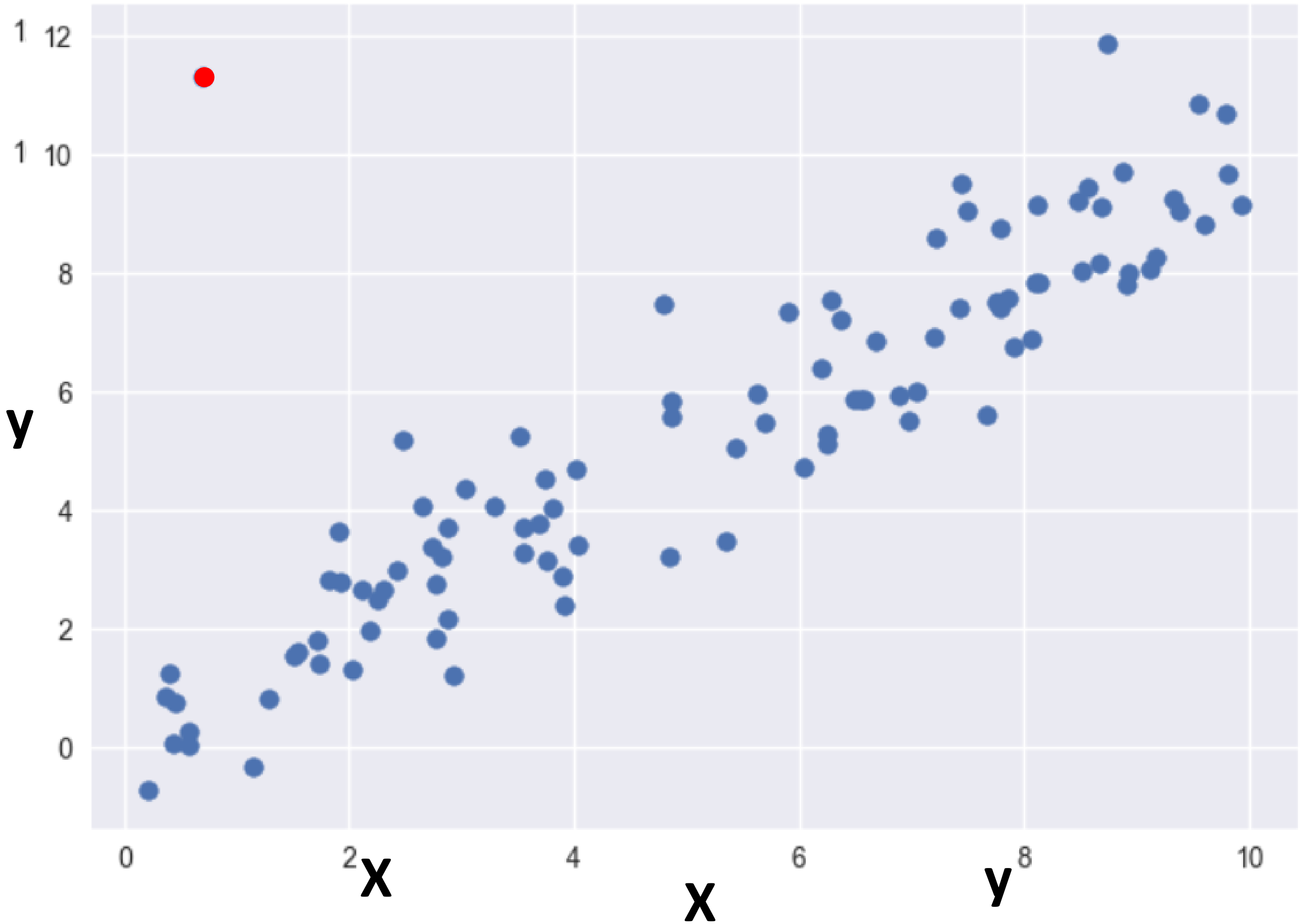
Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. **Outliers**
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Outliers (1)

- An observation point that is distant from other observations.
- Causes for outliers:
  - Variability in measurements
  - Experimental error
  - Can occur by chance (may indicate a heavy-tailed distribution)
- Why do we care?

Univariate vs Multivariate Outliers



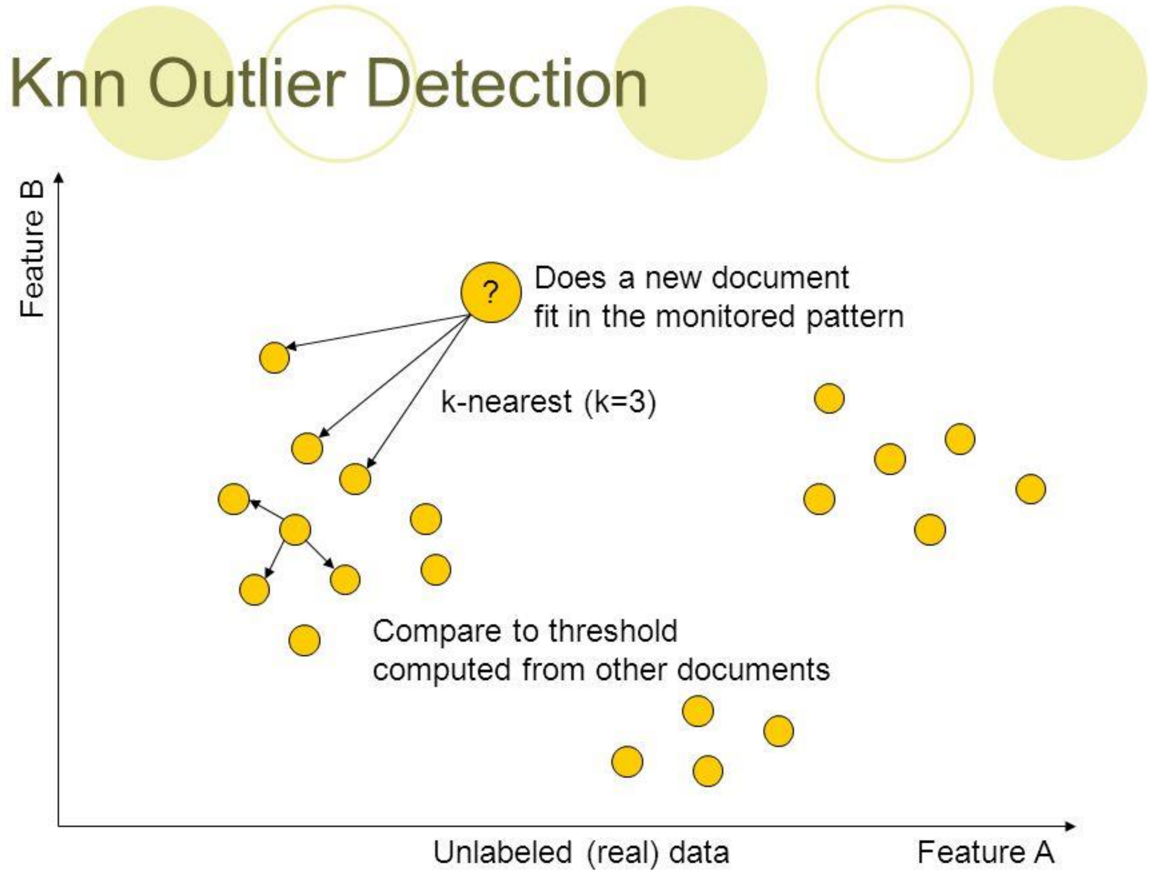
How do we detect outliers?

- Univariate methods:
  - Box Plot
  - 3 SD method
  - Grubbs' test:
    - Evaluates whether the maximal/minimal value is an outlier
    - $G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/2N,N-2)}}{N-2+t^2_{(\alpha/2N,N-2)}}}$
  - Rosner Test:
    - Sequentially apply Grubbs' test

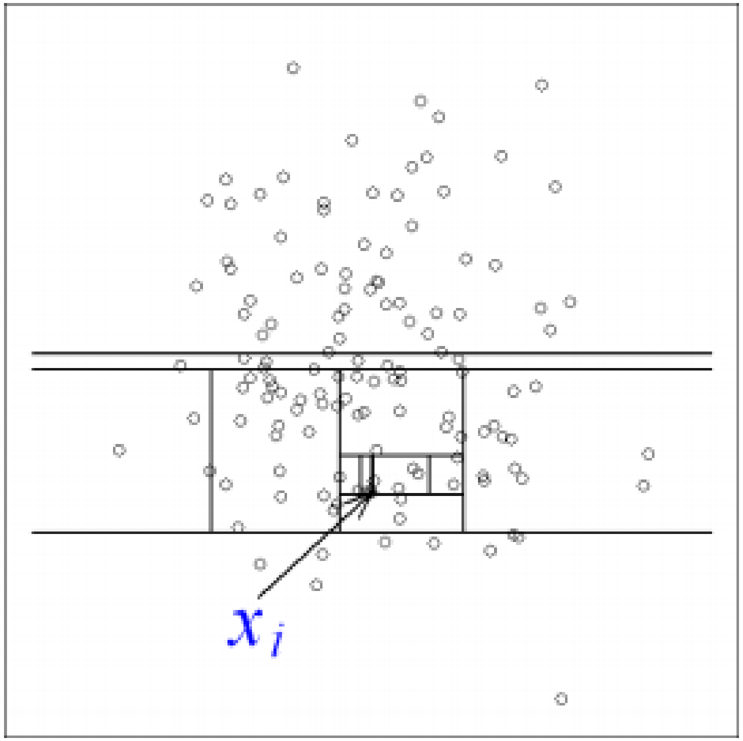
How do we detect outliers?

- Multivariate methods:
  - Nearest Neighbor based estimation:
    - KNN Outlier Detection
    - Local Outlier Factor (LOF)
  - Isolation Forest
  - Robust Covariance
  - One-Class SVM

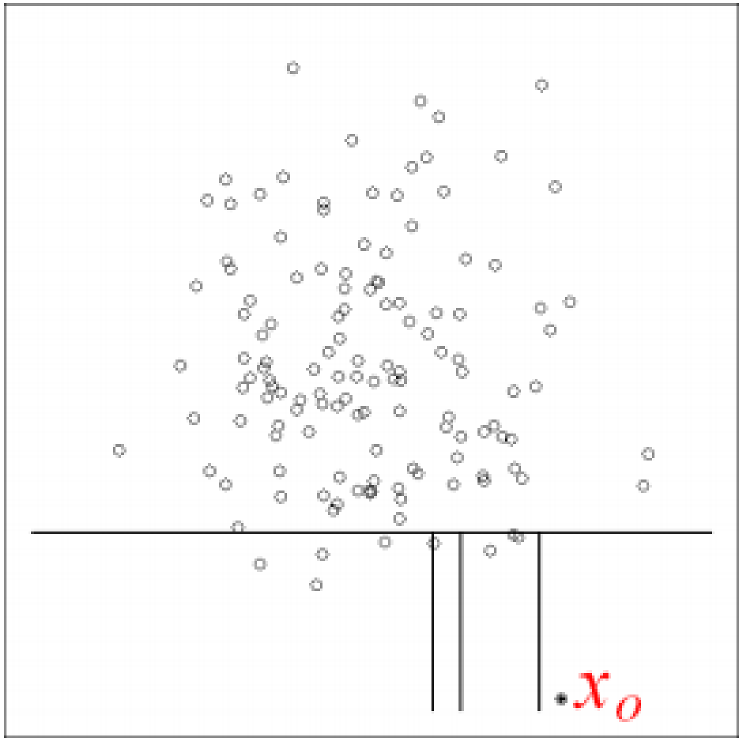
Nearest Neighbor Methods and LOF



Isolation Forest

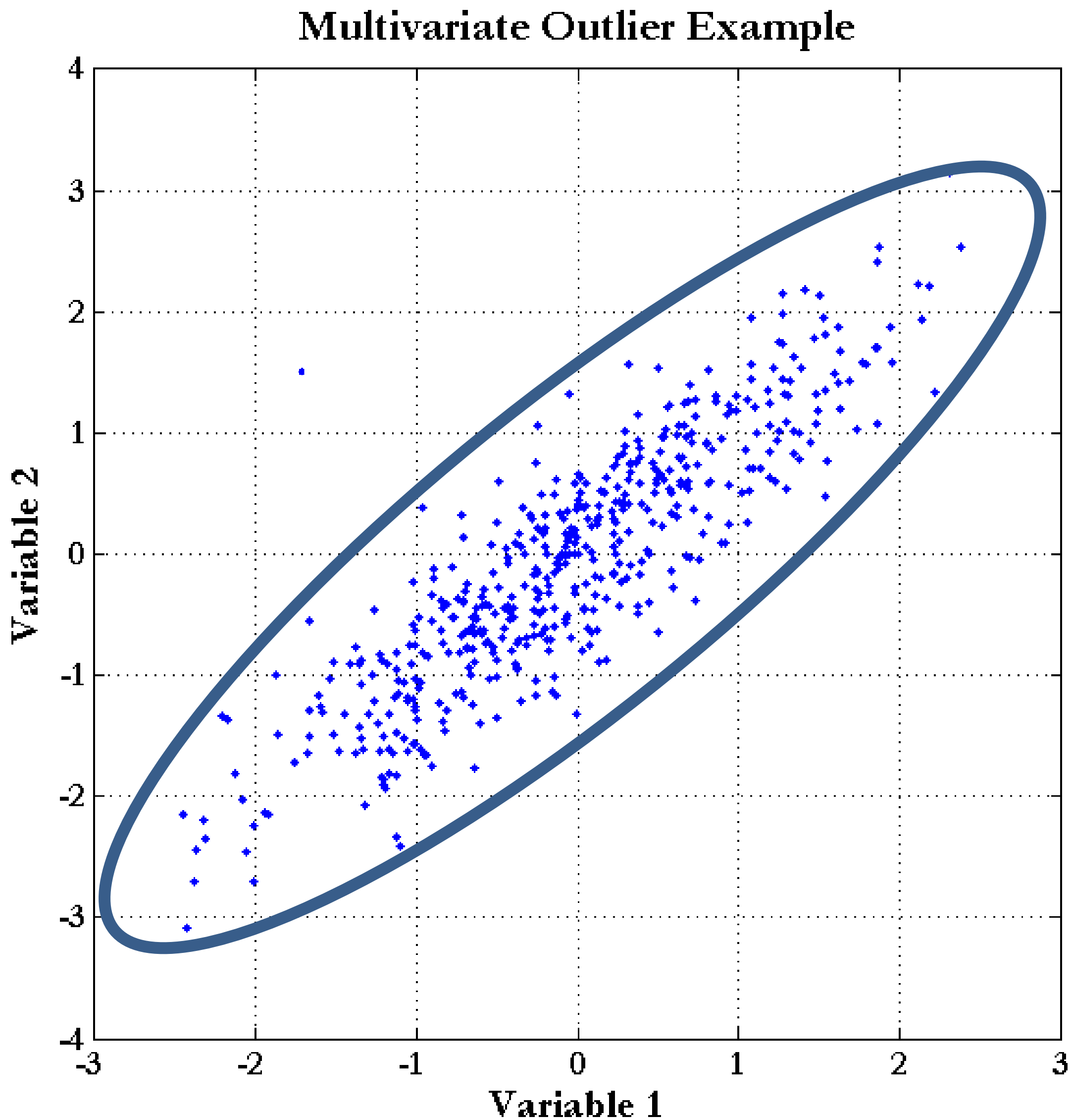


(a) Isolating x_i



(b) Isolating x_o

Robust Covariance Estimation



- Python code

How to deal with outliers?

- Remove them
- Give them unique value
- Use non-sensitive models

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. **Normalization & Transformation**
9. Discretization
10. Unbalanced data

Normalization (1)

- AKA Feature Scaling
- Why do we need to normalize the data?
  - Easy comparison of values
  - In some algorithms, objective functions will not work properly (or quick) without it
- Example:
  - Predict the cost of the house, giving it's size (squared meters) and the # of bedrooms

Rescaling

- The simplest method is rescaling the range of features to scale the range in [0, 1] or [−1, 1]:

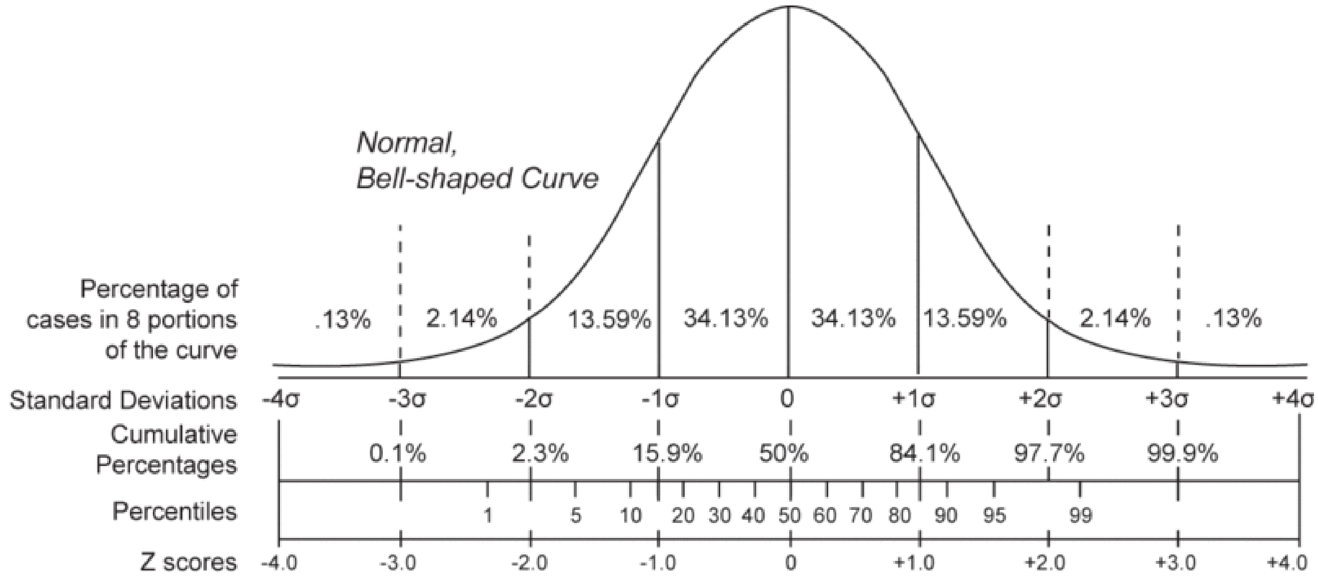
$$X_{i,\ 0\ \text{to}\ 1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

$$X_{i,-1\ \text{to}\ 1} = \frac{2X_i - X_{Min} - X_{Max}}{X_{Max} - X_{Min}}$$

Standardization (Z-normalization)

- Transforms the values of each feature in the data to have zero-mean and unit-variance.
- This method is widely used for normalization in many machine learning algorithms

$$x' = \frac{x - \bar{x}}{\sigma_x}$$



Intel Confidential

Robust Scaling

- Similar to Z-normalization but uses the median and quartiles:

$$x' = \frac{x - median(x)}{IQR_x}$$

IQR_x is defined as the interquartile range, i.e. the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile)

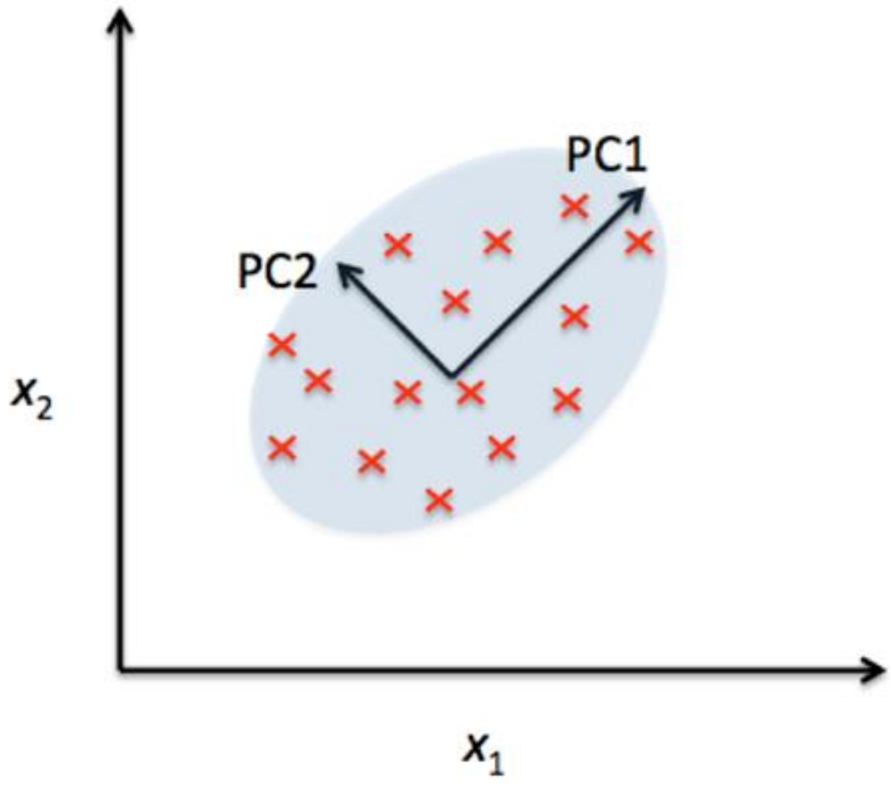
  – This method can cope with outliers better than z-normalization

Feature Transformation

- Various methods of feature transformations:
  - Univariate transformations:
    - log(x)
    - x^2
    - e^x
    - etc.
  - Multivariate transformations:
    - $x_1 \cdot x_2$
    - x_1 / x_2
    - etc.
  - Dimensionality reduction:
    - PCA

Intel Confidential

PCA



- Why use feature transformations?
  - Add nonlinearity to dataset
  - Add context and background experience to feature:

    Example: "123 Main Street, Seattle, WA 98101“
  - Reduce noise from features
  - Reduce number of features used

Polynomial Transformations

- Python example

Agenda

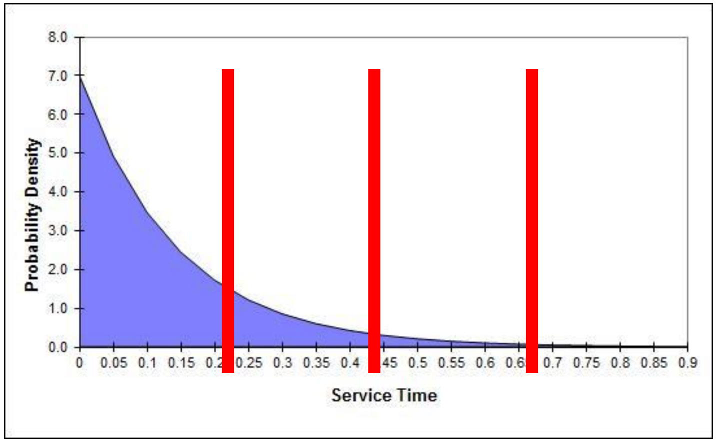
1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. **Discretization**
10. Imbalanced data

Discretization (1)

- Why do we need to change the data?
  - Some models/measures can't handle continuous values (i.e. Naïve Bayes, MI)
  - Some numeric values don't have a meaningful numeric insights (but when taking them as discrete ones – they do have)
  - The business might have useful information to give us.

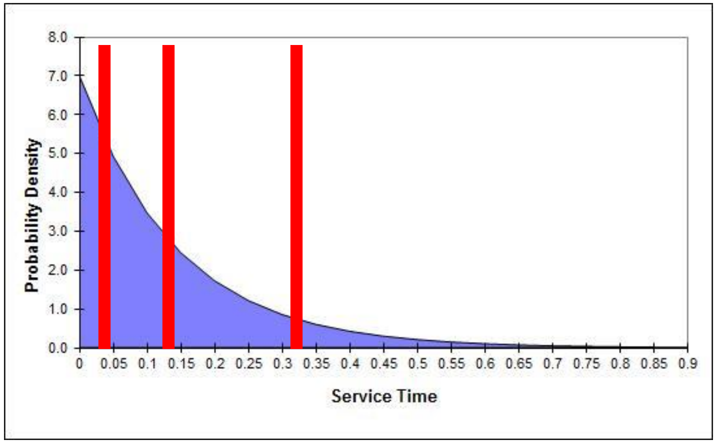
Discretization (2)

- **Equal-width** (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well



Discretization (3)

- **Equal-depth** (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky



Discretization (4)

- **Entropy based**
  - The entropy (or the information content) is calculated on the basis of the class label.
  - Intuitively, it finds the best split so that the bins are as pure as possible, i.e. the majority of the values in a bin correspond to having the same class label.
  - Formally, it is characterized by finding the split with the maximal information gain.

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. **Imbalanced data**

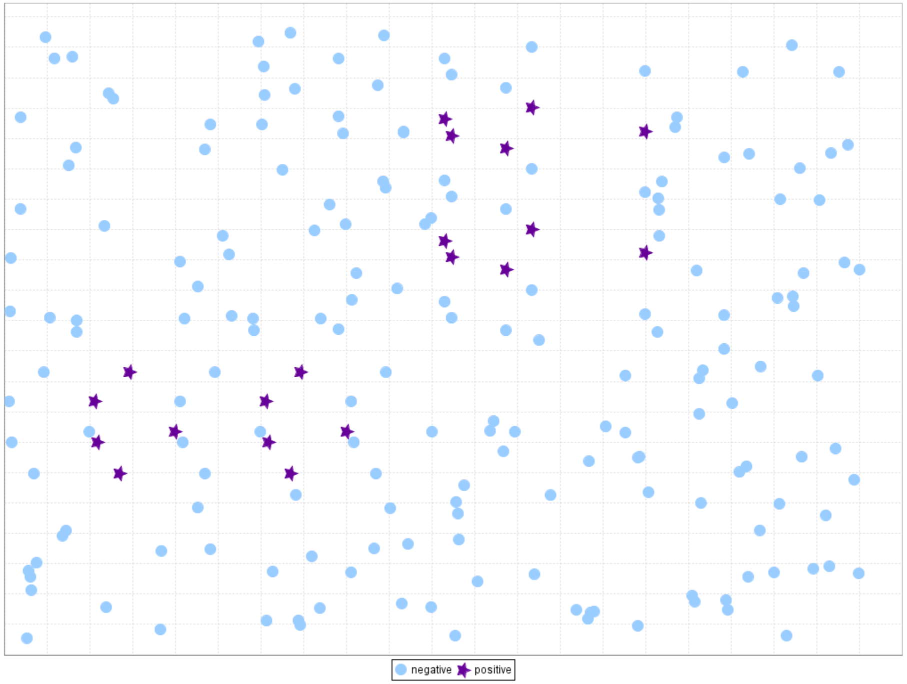
What is imbalanced data?

- Unequal distribution of classes
- Class of interest is often the minority
- Many real world situations
    - Fraud detection
    - Disease prediction
    - Faulty units in production

What’s the problem with imbalance?

- Negative (majority) samples “drown” positive (minority) samples
- Feature selection and modeling algorithms are thrown off course

Bad Models



What's the problem with accuracy?

- Accuracy = the proportion of true results over **all** classes
- A common performance measure

How would **YOU** maximize accuracy?

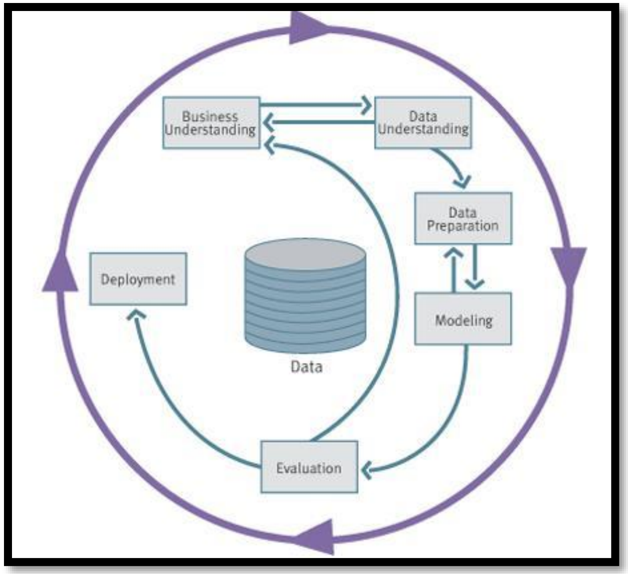
- Simplest solution –
  classify ALL samples as majority class

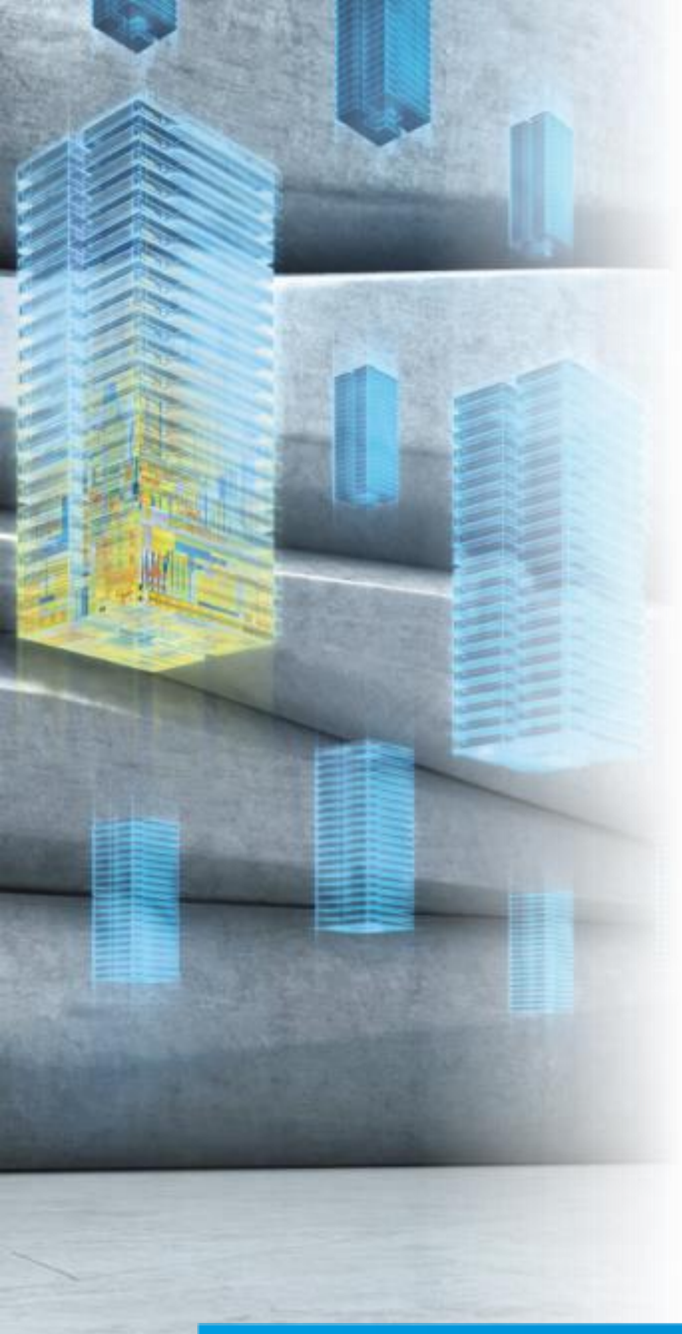
Solutions – high level approaches

- Stratified sampling
- Under-sampling
- Over-sampling
- Ensemble methods
- Cost sensitive methods

Summary

- Topics we have covered
- How CRISP-DM is related to the session
- In practice – what is being done in real life
- Anything else?



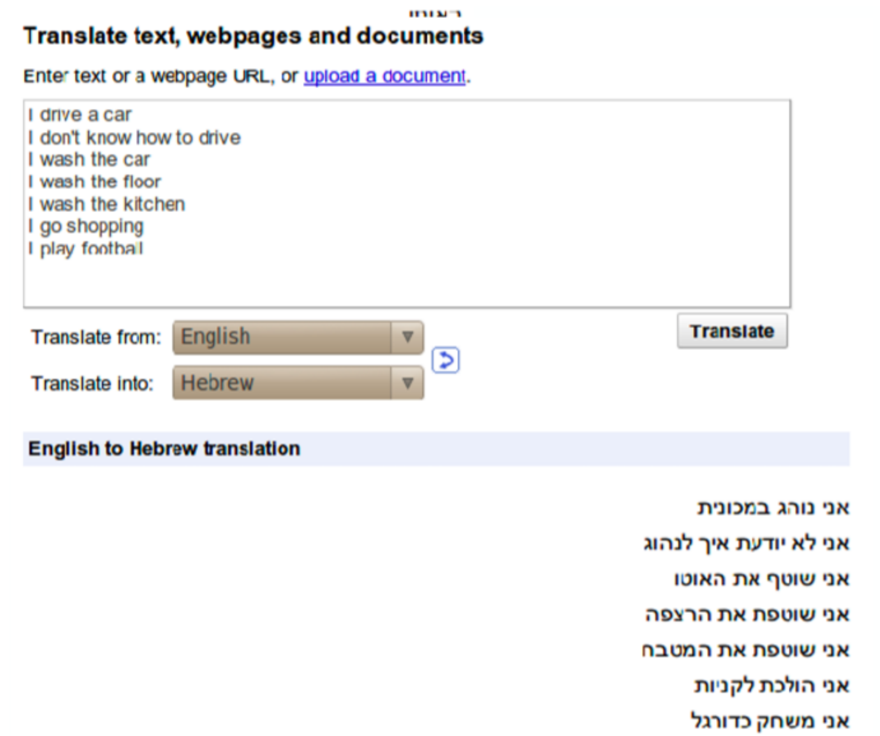


QUESTIONS?

Extra – NLP

- Why is it important?
- Why is it useful?
- Why is it hard?
- Why is it interesting?


15:40 77%
Google Translate
https://translate.google.com
Google Translate
ENGLISH
HEBREW
how about this
מה דעתך על כך
endorsements
רש


Translate text, webpages and documents
Enter text or a webpage URL, or upload a document.
I drive a car
I don't know how to drive
I wash the car
I wash the floor
I wash the kitchen
I go shopping
I play football
Translate from: English
Translate into: Hebrew
Translate
English to Hebrew translation
אני נוהג במכונית
אני לא יודעת איך לנהוג
אני שוטף את האוטו
אני שוטפת את הרצפה
אני שוטפת את המטבח
אני הולכת לקניות
אני משחק כדורגל

A Brief History of NLP

- 1907-11: de Saussure establishes modern linguistics (Structuralism)
- 1921: “All grammars leak” (Language: an intro to the study of speech, Edward Sapir)
- 193X: First patent for ‘translating machine’
- 1941: Turing @ Belchley Park: Breaking the (naval) Enigma
- 1950: ‘Computing Machinery and Intelligence’ (aka: The Turing Test)
- 1954: Hype and optimism – The Georgetown-IBM experiment (“MT is soon to be solved!”)
- 1957: Skinner publishes ‘Verbal Behavior’ (Behaviorism)
- 1957: Chomsky – Syntactic Structures (Universal Grammar, generative linguistics)
- 1960: A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation (Bar Hillel)
- 1964: ELIZA - a therapy chatbot (MIT AI Labs)
- 1988-2000: The rise of Machine Learning (and probabilistic models)
  - Machine Translation: IBM models 1,2,...6 (Bob Mercer, IBM)
  - Speech recognition: “Every time I fire a linguist, the performance of the speech recognizer goes up” (Frederick Jelinek, IBM)
- 200X: The rise of data (“the Internet”)
- 2011: IBM’s Watson wins Jeopardy
- 201X: The rise of “Deep Learning” methods (word2vec, Mikolov 2013)

Extra – NLP Tasks

Common NLP Tasks

Easy

- Chunking
- Part-of-Speech Tagging
- Named Entity Recognition
- Spam Detection
- Thesaurus

Medium

- Syntactic Parsing
- Word Sense Disambiguation
- Sentiment Analysis
- Topic Modeling
- Information Retrieval

Hard

- Machine Translation
- Text Generation
- Automatic Summarization
- Question Answering
- Conversational Interfaces

Extra – NLP Data Understanding

- Sentences level analysis
- Tokens level analysis
- Symbols level analysis (e.g. words, hashtags)
- Known linguistic behavior (e.g. RT in twitter)

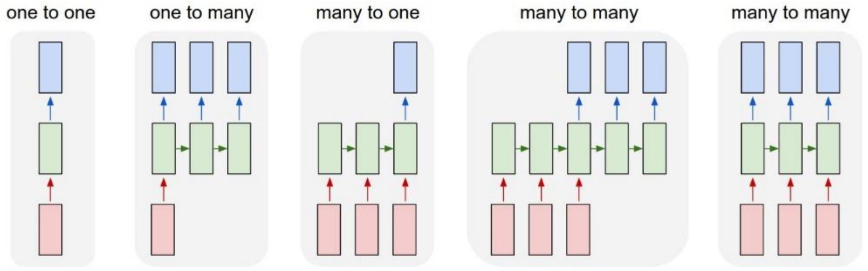
Extra – NLP Data Prep

- Tokenization
- Text cleaning (e.g. URL, hashtags)
- Data removal (e.g. stop-words, symbols)
- Negative wording
- Normalization (Stemming, Lemmatization)
- Part Of Speeach

Extra – NLP Modeling

- Classification
  - BOW and then any model you wish
  - NN (RNN, CNN, transformers)
- Other tasks
  - NN (RNN, CNN, transformers)

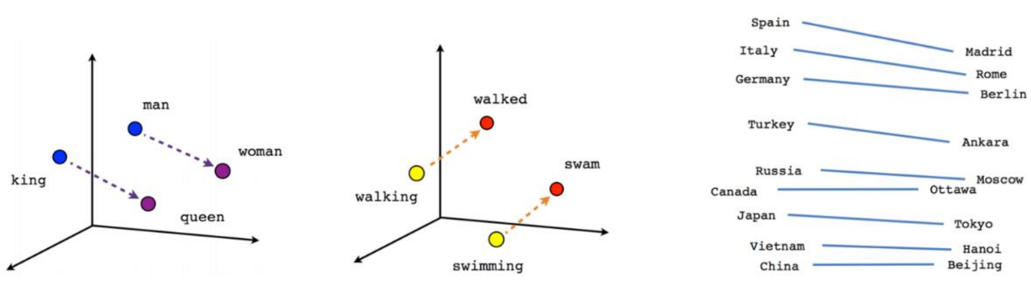
Today's trend: Transfer learning



Extra – NLP Embedding

- Mikolov 2013
  - New concept in words representation
  - Words into vectors – what can it allow us?

Today's trend: Dynamic embedding + all is [U_NAME_IT]2vec



Male-Female Verb tense Country-Capital

Extra – NLP tools

- Python main packages:
  - NLTK
  - Spacy
  - Gensim (+fasttext)
  - Pytorch/tf

- Other NLP tools:
  - Open source tools/repos