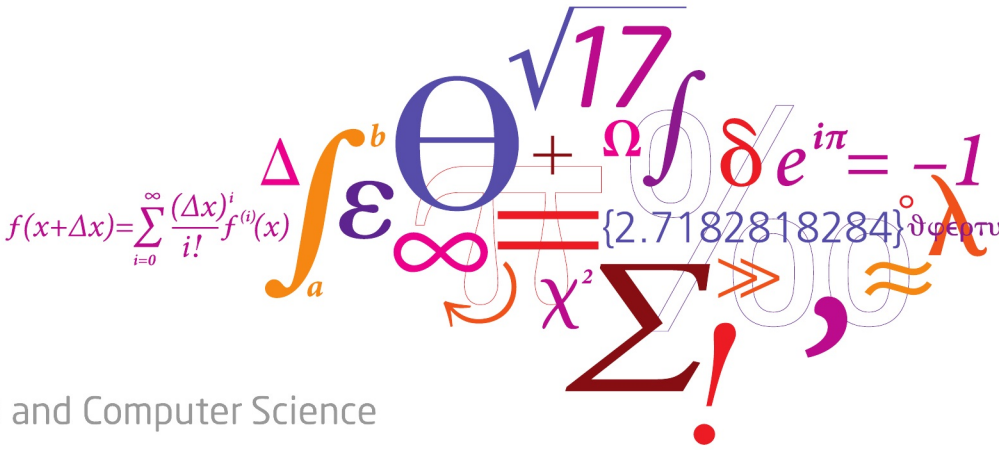
02450: Introduction to Machine Learning and Data Mining

Measures of similarity, summary statistics and probabilities

Tue Herlau

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

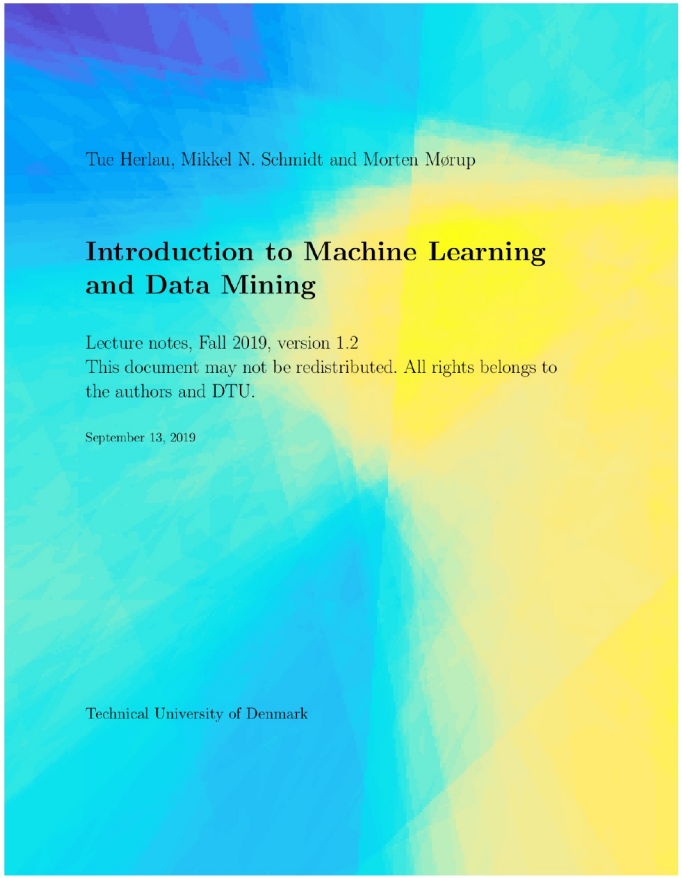
Today

Feedback Groups of the day:

Georgios Bekakos, Georgios Bekakos, Vasileios Boutlas, Cecilie Amalie Neijendam Thystrup, Clara Drachmann, Søren Steenstrup Zeeberg, Lisbeth Ditte Evald Sandvik, Agnete Marie Nørregaard, Robert Stig Grønlund Nielsen, Lukas Kofoed Gildhoff, Khushboo Nyman, Enrico Tolotto, Alexandros Spyropoulos, Louis Perot, Felix Gigler, David Onnen, Stefan Bîrs, Morten Bjerre, Mads Gramtorp Bjørn, Callum Blair, Kálmán Bogdán, Julian Böhm, Lisa Veibel Bonde, Julia Bonzanini, Annie Borch, Rishav Bose, Oskar Johannes Fred Bremberg, Mads Brink, Benjamin Bruun, Mikkel Bruus, Rasmus Stokholm Bryld, Razvan-Vlad Bucur, Jonathan Buhl, Philip Nielsen Butenko, Manuel Caballero Peña, Cemre Çadir, Yiyi Cao, Chiara Libera Carnevale, Dilara Eda Celepli, Emmanouil Chalvatzopoulos, Anshul Chauhan, Dominic Chen, Darius Chira, Seokhyun Choung, Emil Chrisander, Oskar Eiler Wiese Christensen, Line Wulff Christensen, Ann-Katrine Grodt Christiansen, Jonas Søbro Christophersen, Yeon Soo Chung, Sofus Rasmus Clausen, Joakim Nøddeskov Clifford

Reading material:

Chapter 4, Chapter 5



Lecture Schedule

1. Introduction
   3 September: C1

Data: Feature extraction, and visualization

2. Data, feature extraction and PCA
   10 September: C2, C3
3. **Measures of similarity, summary statistics and probabilities**
   17 September: C4, C5
4. Probability densities and data Visualization
   24 September: C6, C7

Supervised learning: Classification and regression

5. Decision trees and linear regression
   1 October: C8, C9 **(Project 1 due before 13:00)**
6. Overfitting, cross-validation and Nearest Neighbor (Note: Tentative)
   8 October: TBA
7. Bayes, Naive Bayes and performance evaluation (Note: Tentative)
   22 October: TBA
8. Artificial Neural Networks and Bias/Variance
   29 October: C14, C15
9. AUC and ensemble methods
   5 November: C16, C17

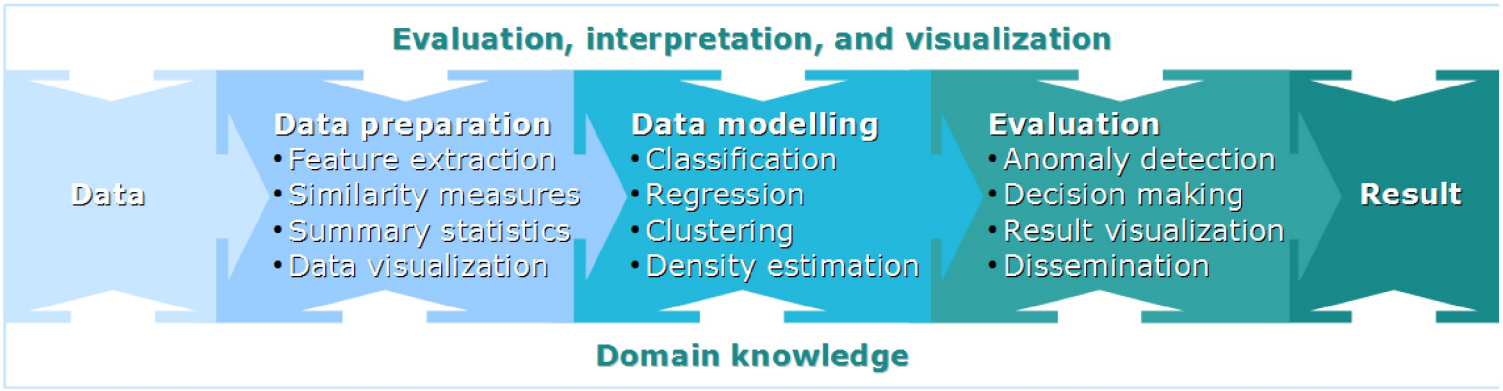
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    12 November: C18 **(Project 2 due before 13:00)**
11. Mixture models and density estimation
    19 November: C19, C20
12. Association mining
    26 November: C21

Recap

13. Recap and discussion of the exam
    3 December: C1-C21 **(Project 3 due before 13:00)**

Piazza online help: `https://piazza.com/dtu.dk/fall2019/02450`
Videos/streaming of lectures: `https://video.dtu.dk`



Learning Objectives

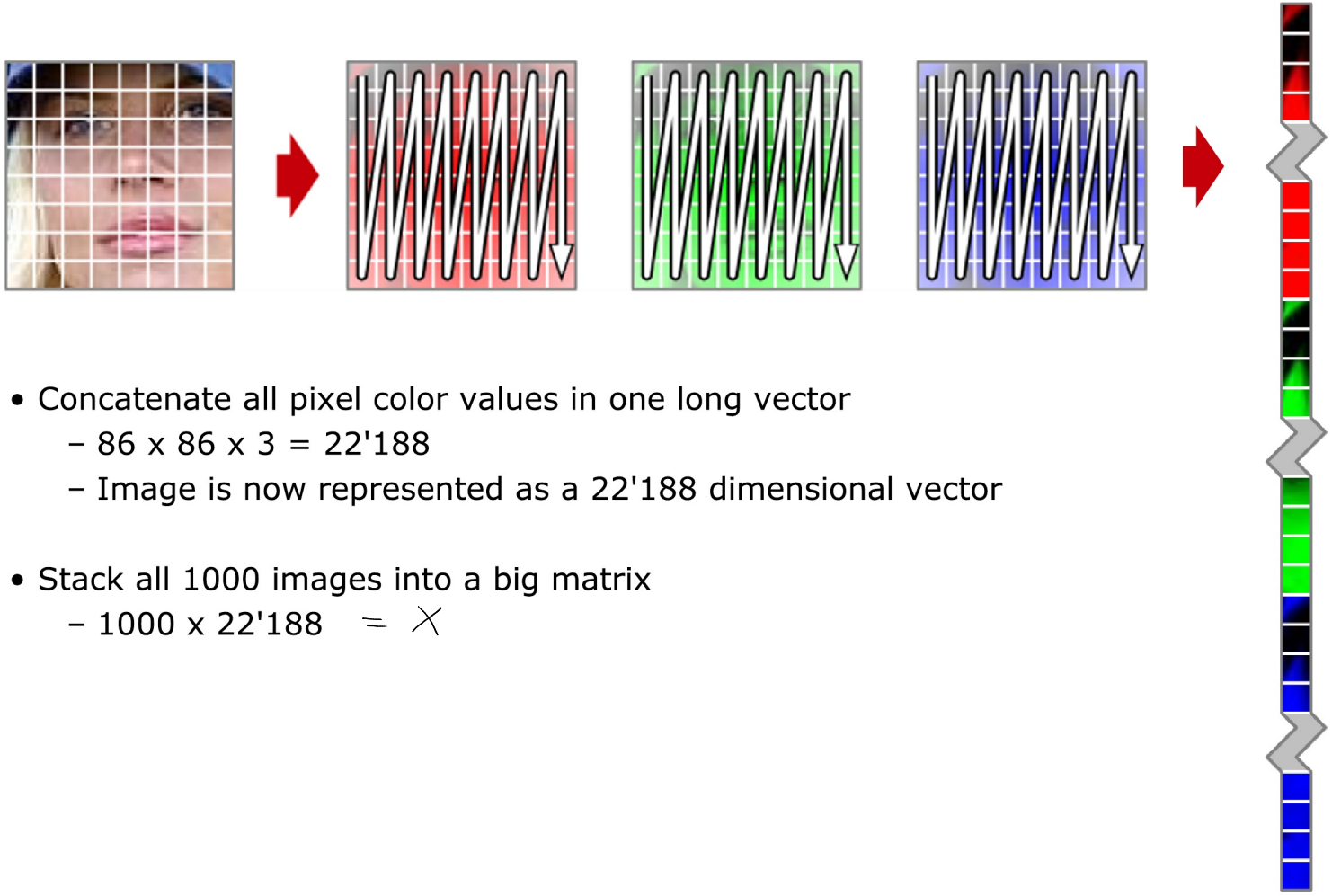
- Compute measures of similarity/dissimilarity (Lp distance, cosine distance, etc.)
- Understand basics of probability theory and stochastic variables in the discrete setting
- Understand probabilistic concepts such as expectations, independence and the Bernoulli distribution
- Understand the maximum likelihood principle for repeated binary events

PCA recap: Principal component analysis on images

- 1000 images, 86 x 86 pixels, 3 RGB intensities

Tamara Berg "Faces in the wild"

Pre-processing

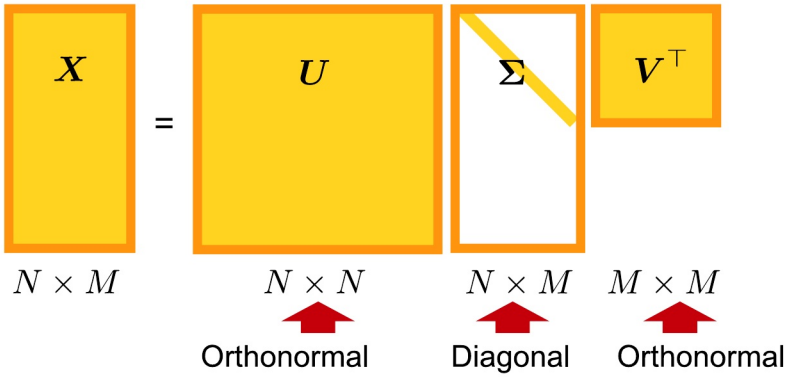


- Concatenate all pixel color values in one long vector
  - 86 x 86 x 3 = 22'188
  - Image is now represented as a 22'188 dimensional vector

- Stack all 1000 images into a big matrix
  - 1000 x 22'188 $\simeq$ $\times$

Principal component analysis (PCA)

1. **Subtract the mean**
   -Consider dividing with variance; use 1-out-of-K coding for nominal attributes
2. **Compute the singular value decomposition (SVD)**
   - Orthogonal linear transformation
   - Transforms data to a new coordinate system
     - Greatest variance along the first axis (first column of V)
     - Second greatest variance along the second axis



- **Plot data in the transformed coordinate system**
  - Corresponds to looking at data from an angle where it is most spread out

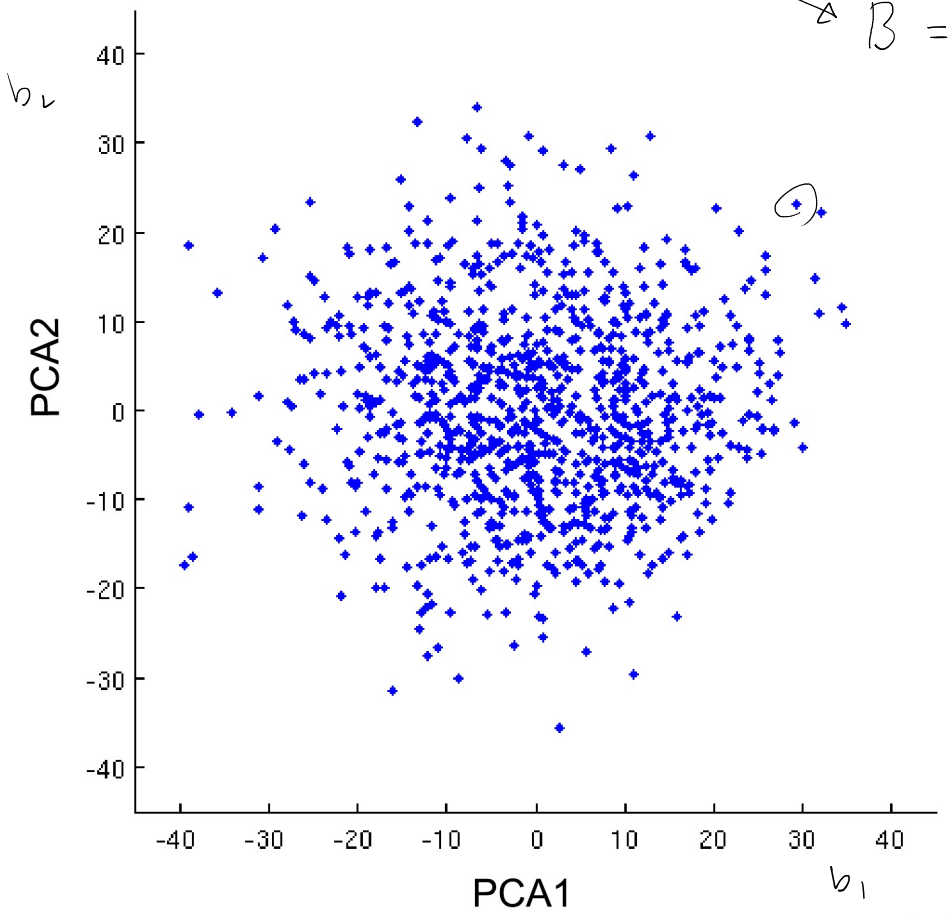
PCA on face images

$V_n = [V_1 \; V_2]$, $n = 2$

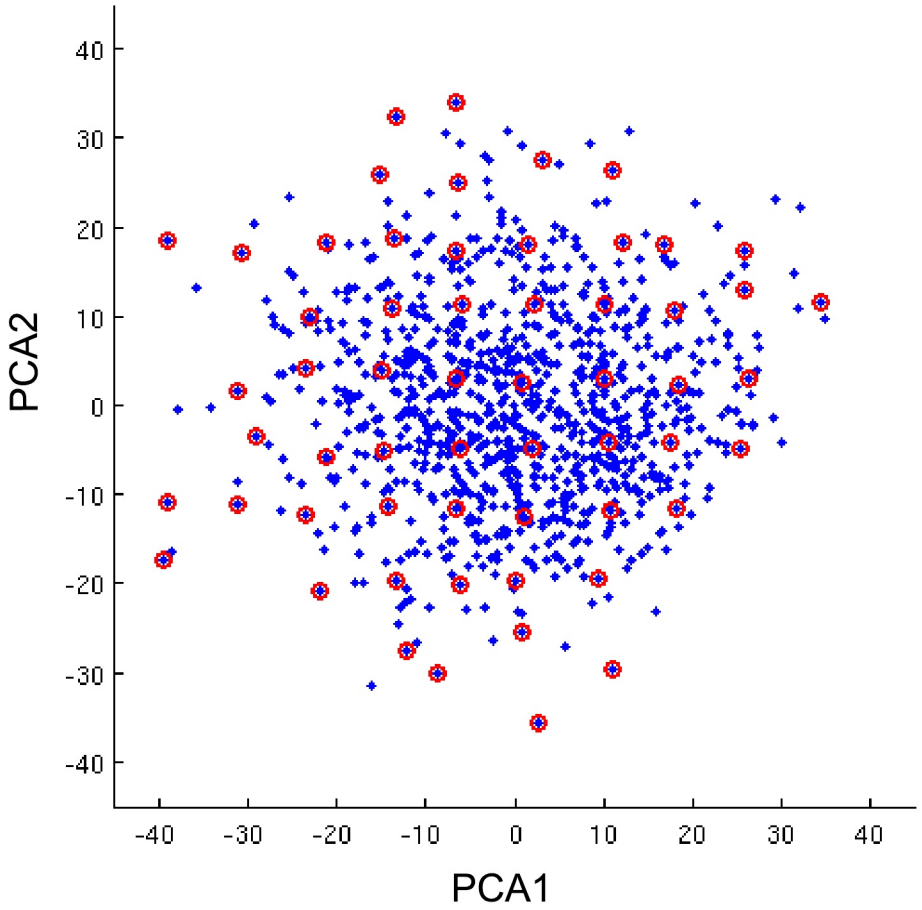
$[b_1 \; b_2] = [\tilde{X}_i V_1 \; \tilde{X}_i V_2]$

$= \tilde{X}_i^T V_n$

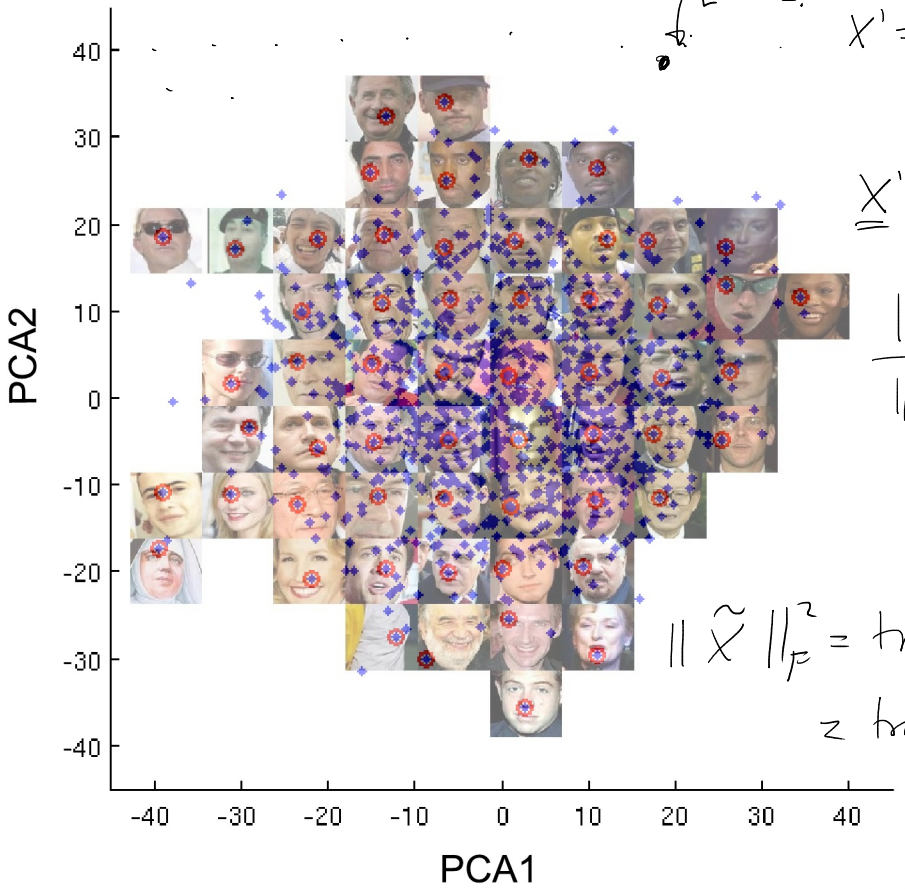
1000 × 2 ← $B = \tilde{X} V_n$ → 22188 × 2



PCA on face images



PCA on face images



$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = b$$

$$\tilde{X}' = b_1 V_1 + b_2 V_2$$

$$= V_n \cdot b$$

$$\underline{\tilde{X}}' = \left(V_n B^T\right)^T \qquad 1000 \times 22188$$

$$\frac{\|X'\|_F^2}{\|\tilde{X}\|_F^2} = \frac{\|B V_n^T\|_F^2}{\|\tilde{X}\|_F^2}$$

$$= \frac{\|\tilde{X} V_n V_n^T\|_F^2}{\|\tilde{X}\|_F^2} = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

$$\|\tilde{X}\|_F^2 = \text{trace}(\tilde{X}^T \tilde{X}) \overset{*}{=} \sum_{i=1}^{n} \sigma_i^2$$

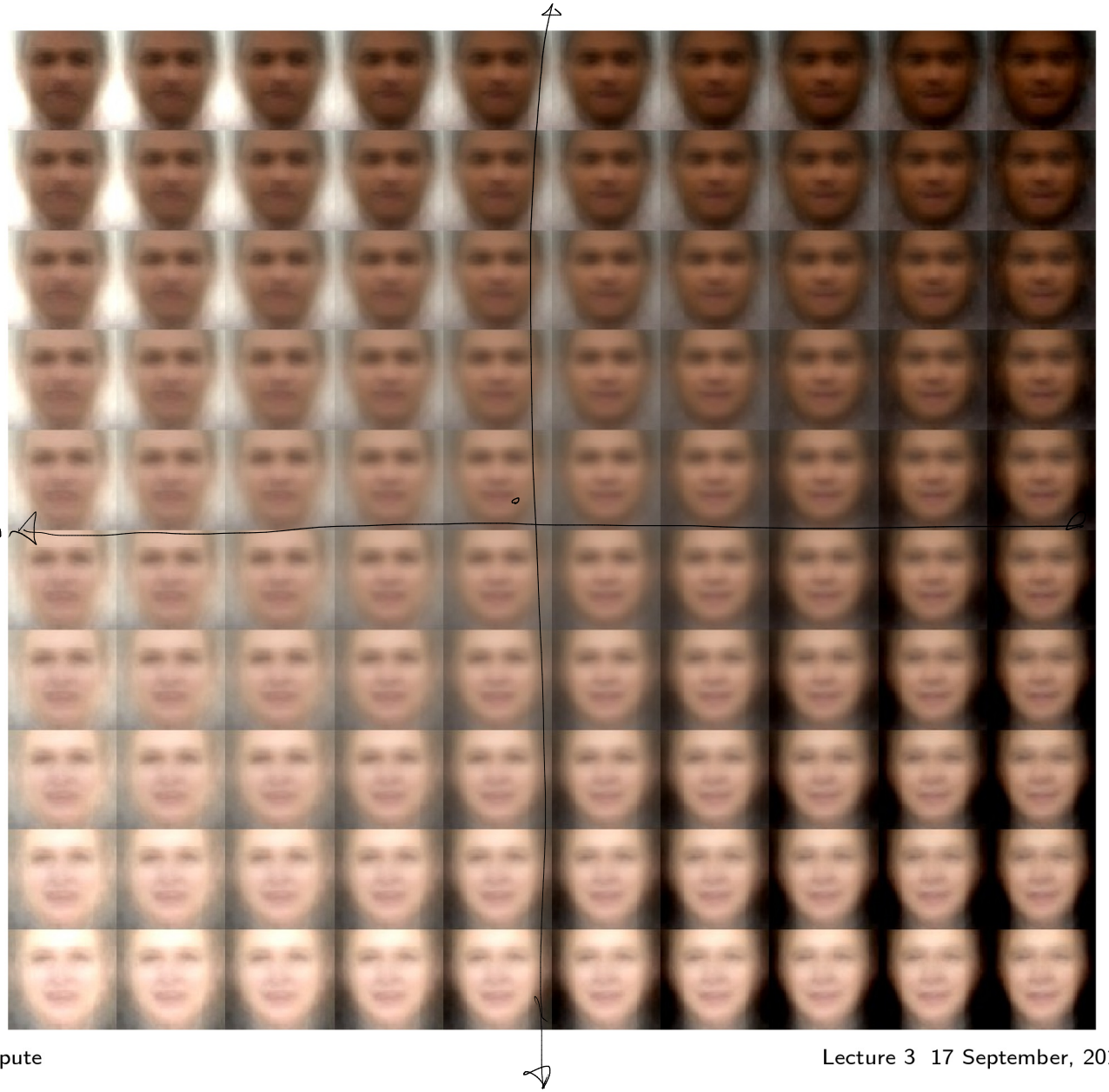
$$= \text{trace}\left((U \Sigma V^T)^T (U \Sigma V^T)\right)$$

$$= \text{tr}\left(V \Sigma^T U^T U \Sigma V^T\right)$$

$$= \text{trace}\left(V \Sigma^T \Sigma V^T\right)$$

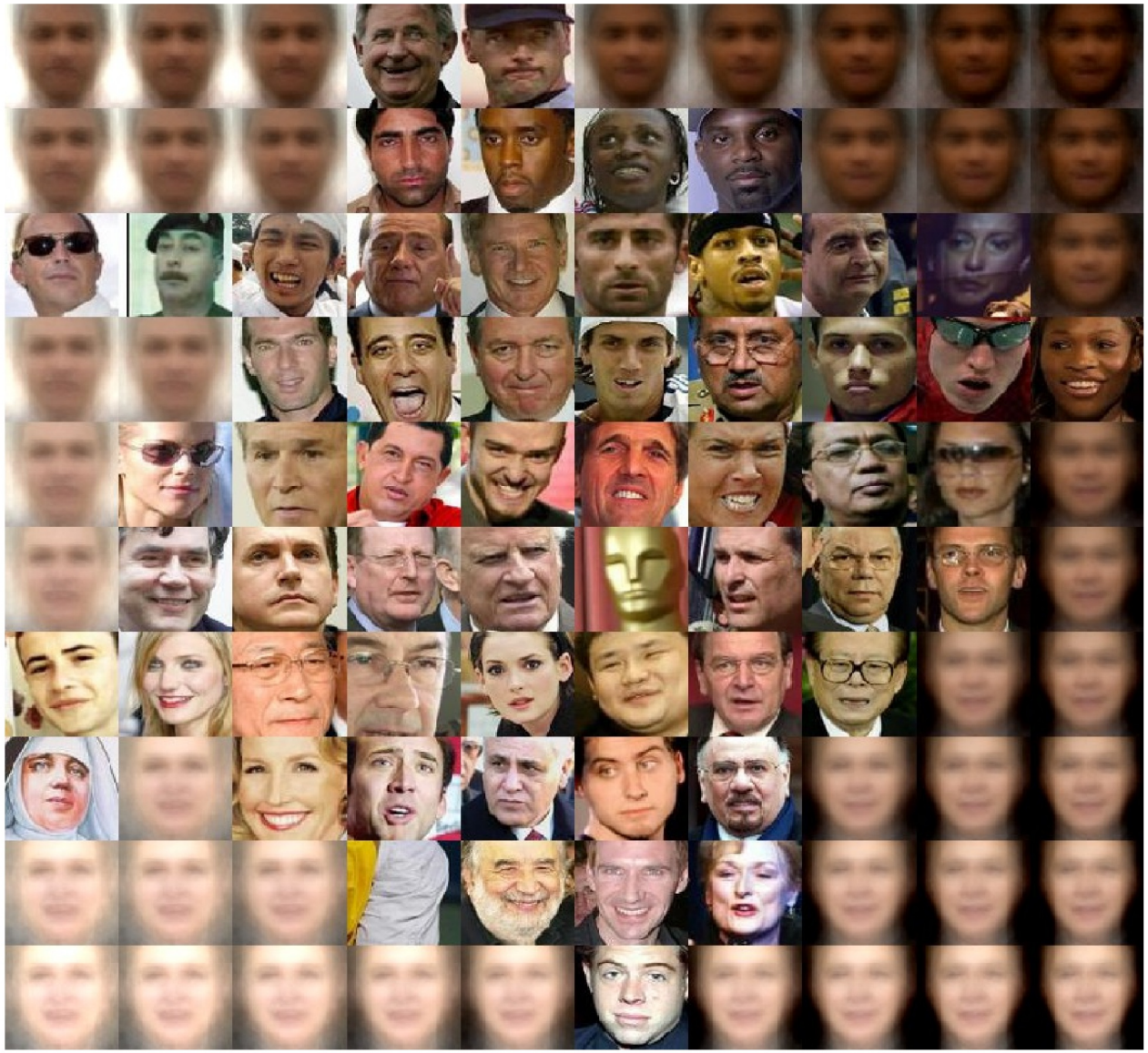
$$= \text{trace}\left(V^T V \Sigma^T \Sigma\right)$$

$$= \text{tr}\left(\Sigma^T \Sigma\right) \quad (*)$$

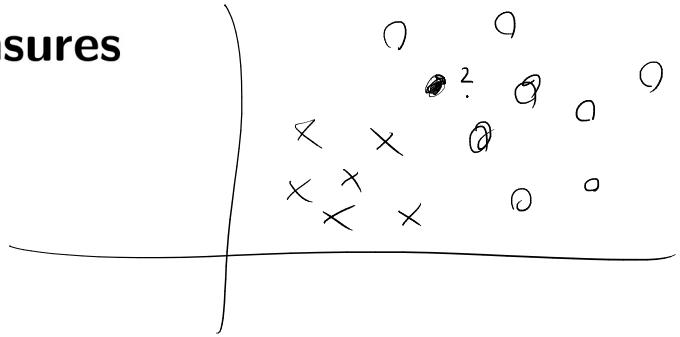


- What information do the two principal axes capture?

What information do the two principal axes capture?



Similarity / Dissimilarity measures

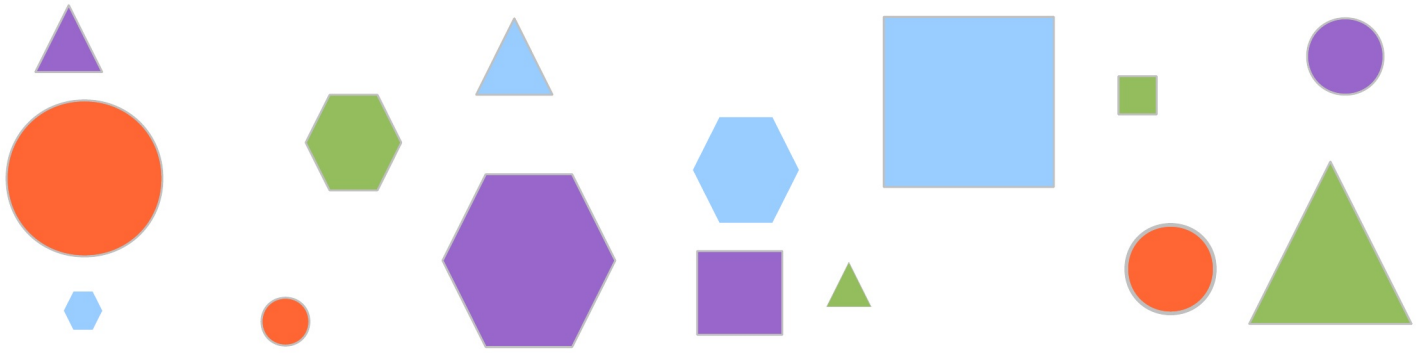


Similarity $s(\boldsymbol{x}, \boldsymbol{y})$ Often between 0 and 1. Higher means more similar

Dissimilarity $d(\boldsymbol{x}, \boldsymbol{y})$ Greater than 0. Lower means more similar.

Classification Classify a document as having the same topic y as the document is is **most similar**/**least dissimilar** to.

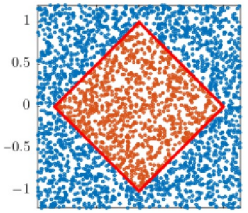
Outlier detection The observation most **dissimilar** to all other observations is an outlier



Dissimilarity measures

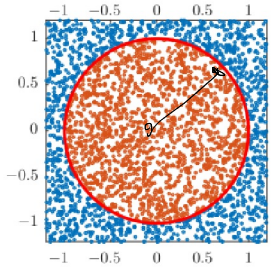
- General Minkowsky distance (p-distance) $d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}$
- One-norm ($p = 1$)

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{M} |x_j - y_j|$$



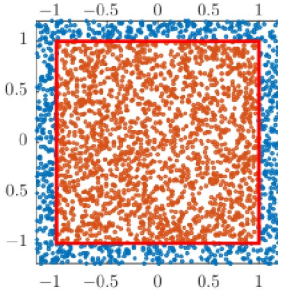
- Euclidean ($p = 2$)

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{M} (x_j - y_j)^2}$$



- Max-norm distance ($p = \infty$)

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_M - y_M|\}$$



Usage: Regularization and alternative optimization targets. For instance, $p = \infty$ is very affected by outliers, $p = 1$ much less so.

Similarity measures

$x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ $y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$ K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

DTU

Simple Matching Coefficient (SMC)

$$\mathrm{SMC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{00} + f_{11}}{K}$$

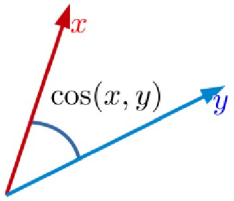
+ Symmetric

Jaccard Coefficient

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{11}}{K - f_{00}}$$

+ Positive matches

Cosine similarity



$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

+ Positive matches
+ Document length

Extended Jaccard coefficient

$$\mathrm{EJ}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^\top \boldsymbol{y}}$$

Also defined for continious data

Quiz 1, similarity measures

Calculate the simple matching coefficient, Jaccard, cosine and extended jaccard similarity between customer 1 and customer 2 in the market basket data below.

$f_{11} = 2$ $\quad f_{00} = 1$

	ID	Bread	Soda	Milk	Beer	Diaper
O_1	1	1	1	1	0	0
O_2	2	0	1	1	0	1

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Which of the following statements are true?

A. $\mathrm{SMC}(o_1, o_2) = \frac{3}{5}$ $J(o_1, o_2) = \frac{1}{2}$, $\cos(o_1, o_2) = \frac{2}{3}$,

B. $\mathrm{SMC}(o_1, o_2) = \frac{3}{5}$ $J(o_1, o_2) = \frac{3}{4}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,

C. ~~$\mathrm{SMC}(o_1, o_2) = \frac{2}{5}$~~ $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \frac{2}{3}$,

D. ~~$\mathrm{SMC}(o_1, o_2) = \frac{2}{5}$~~ $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,

E. Don't know.

$$\mathrm{SMC}(x, y) = \frac{f_{00} + f_{11}}{K} \quad \frac{2+5}{5}$$

$$\mathrm{J}(x, y) = \frac{f_{11}}{K - f_{00}} \quad \frac{2}{4}$$

$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}$$

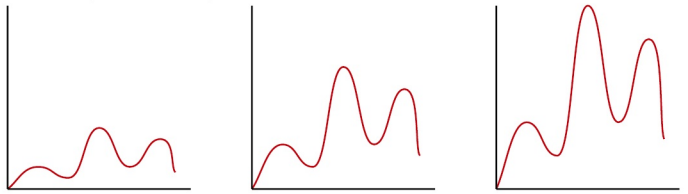
$$\mathrm{EJ}(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

The problem is easily solved by using the inserted formula. We obtain: $\text{SMC}(o_2, o_2) = \frac{3}{5}$ $J(o_1, o_2) = \frac{1}{2}$, $\cos(o_1, o_2) = \frac{2}{3}$ and therefore the A is true. Since the data is binary, the extended Jaccard and the jaccard coefficient agree.

Invariance

Scale invariance

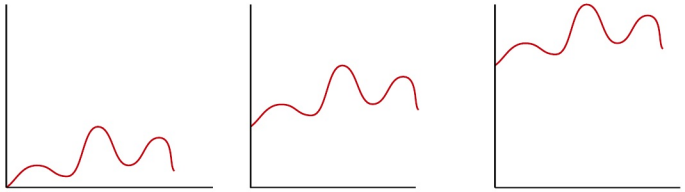
$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha \boldsymbol{x}, \boldsymbol{y})$$



Translation invariance

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha + \boldsymbol{x}, \boldsymbol{y})$$

$$d(x, y) = d_p(x - \bar{x}, y - \bar{y})$$



General invariances

$$d\left(\boxed{6}, \boxed{2}\right) = d\left(\boxed{16}, \boxed{2}\right)$$

Transformations

Standardization: Ensure a single attribute will not dominate:

$$\tilde{x}_{ik} = \frac{x_{ik} - \hat{\mu}_k}{\hat{\sigma}_k}$$

Example:
- **Number of children** ~ 0-5
- **Age** ~ 0-100 years
- **Annual income** ~ 0-50.000 €

Combining heterogeneous attributes Transform measures and combine

$$\begin{aligned} s_{\text{Edu.}} &= \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}}) \\ s_{\text{Age.}} &= a\left(a + d_1(x_{\text{Age.}}, y_{\text{Age.}})\right)^{-1}, \ a = 1 \\ s(x, y) &= \frac{1}{2}\left(s_{\text{Edu.}} + s_{\text{Age.}}\right) \end{aligned}$$

Example:
- **Age:** Continuous
- **Education:** Binary
  - Primary (yes/no)
  - Secondary (yes/no)
  - Tertiary (yes/no)

Weighting Attributes have different importance

$$s(x, y) = 0.99 s_{\text{Edu.}} + 0.01 s_{\text{Age.}}$$

Empirical statistics

Given two samples $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$:

- Empirical mean

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Empirical variance

$$\hat{s} = \hat{\mathrm{var}}[x] = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

- Empirical covariance

$$\hat{\mathrm{cov}}[x, y] = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

- Empirical standard deviation

$$\hat{\sigma} = \hat{\mathrm{std}}[x] = \sqrt{\hat{s}}$$

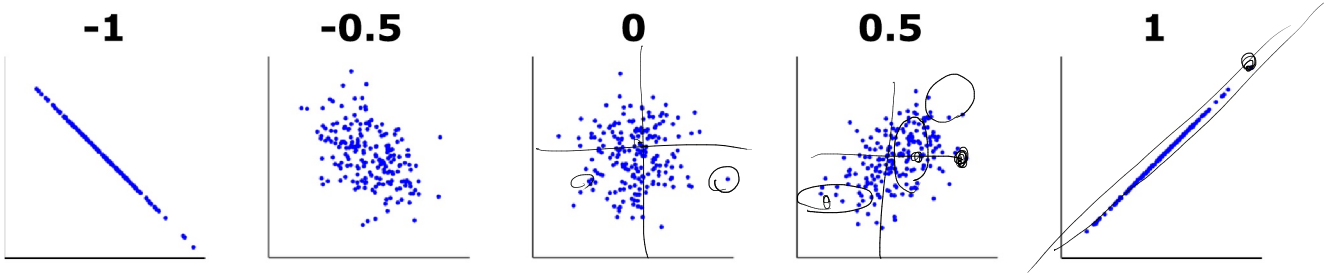
Correlation

- Measure of degree of linear relationship

$$\hat{\mathrm{cor}}[x,y] = \frac{\hat{\mathrm{cov}}[x,y]}{\hat{\sigma}_x \hat{\sigma}_y}$$

  - A correlation of **1** or **-1** means there is a perfect linear relation

$$x_k = a y_k + b$$



$$= \frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \hat{\mu}_y\right)\left(a y_i + b - \left(a\hat{\mu}_y + b\right)\right)}{\hat{\sigma}_x \hat{\sigma}_y}$$

$$= \frac{\frac{1}{N-1}\, a \sum_{i=1}^{N}\left(y_i - \hat{\mu}_y\right)\left(y_i - \hat{\mu}_y\right)}{\sigma_x \sigma_y}$$

$$= \frac{\frac{a}{N-1}\sum_{i=1}^{N}\left(y_i - \hat{\mu}_y\right)\left(y_i - \hat{\mu}_y\right)}{|a|\,\hat{\sigma}_y \hat{\sigma}_y}$$

$$= \frac{a}{|a|} = \mathrm{sign}(a)$$

Quantiles

Given N observations of an attribute $x_1, x_2, \ldots, x_N$. The **q'th quantile** is the value x_q of x such that a fraction q of the sample is **smaller** than x_q.

- Sort in ascending order $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$
- q'th quantile is then (approximately) $x_{(\lceil Nq \rceil)}$
- Percentile is the same except q is given in percent $q = \frac{p}{100}$.
- **Median** is the $q = \frac{1}{2}$ quantile:

$$\text{median}[x] = \begin{cases} x'_{\frac{(N+1)}{2}} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x'_{\frac{N}{2}} + x'_{\frac{N}{2}+1}\right) & \text{if } N \text{ is even.} \end{cases}$$

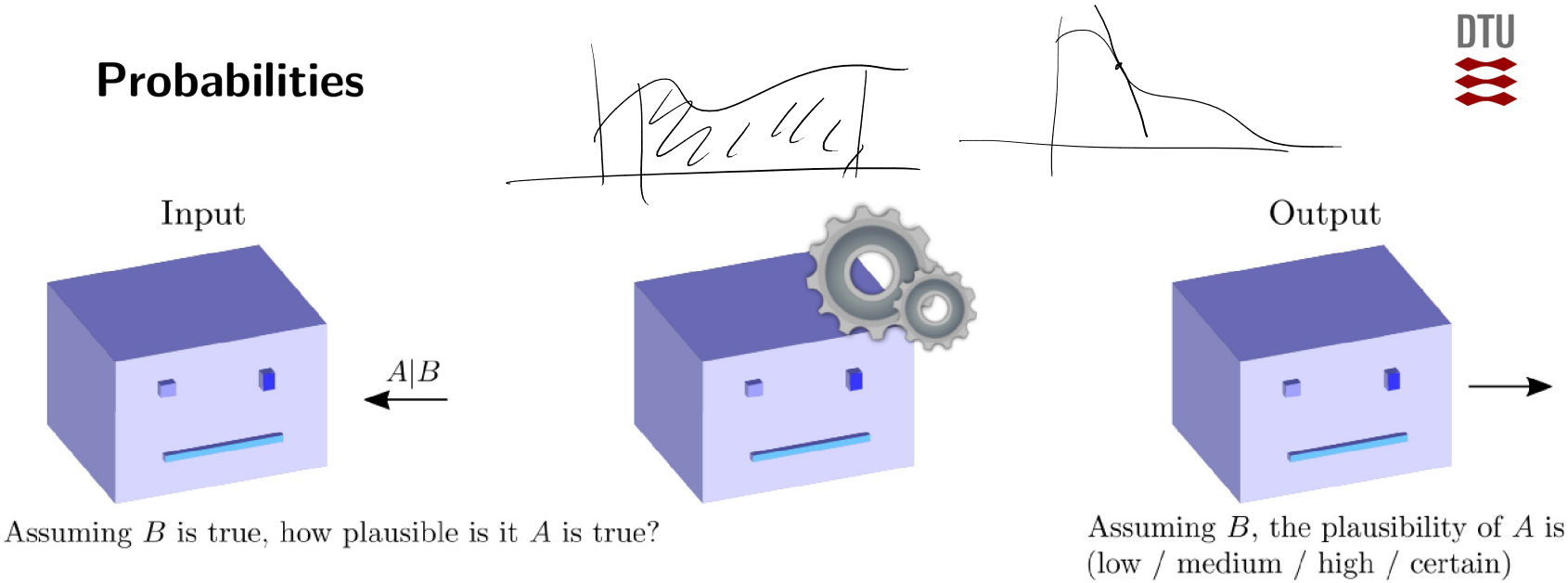
Probabilities

Pragmatically: A big part of AI is dealing with **uncertainty** and **incomplete information**. Probabilities is the formal framework for doing so

Algorithmically: **If** and image belongs to a particular category is a discrete event. The **probability** it belongs to a category is continuous. Algorithmically, easier to optimize continuous quantities

Convenience: There are boiler-plate ideas for transforming a **probabilistic** assumption into an algorithm (maximum likelihood)

Probabilities



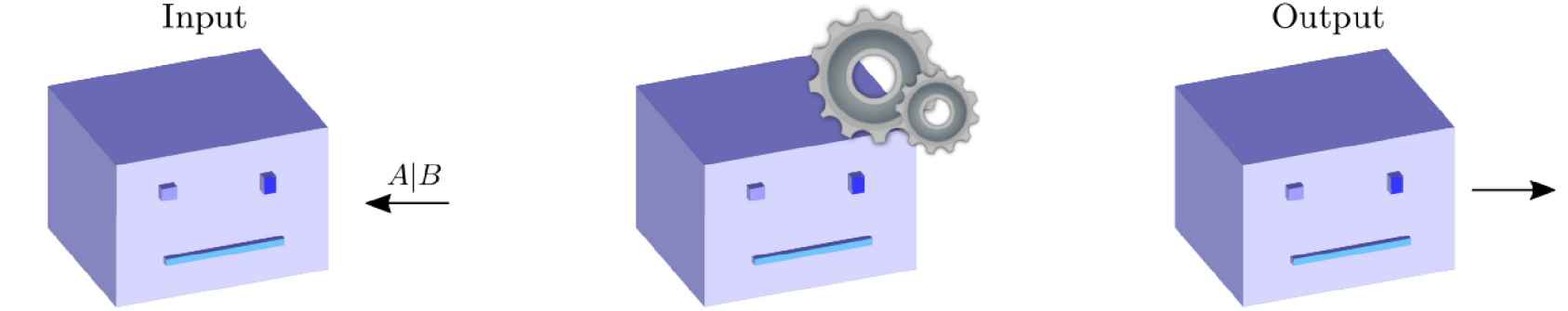
We reason about a proposition A in light of evidence B:

$$P(A|B) = x$$

The degree-of-belief that A is true given B is accepted as true is at a level x

- A number between 0 and 1
- A and B are always binary (true/false) propositions
- Represents a *state of knowledge*

Probabilities: Trial example



Assuming B is true, how plausible is it A is true?

Assuming B, the plausibility of A is
(low / medium / high / certain)

G : *The accused is guilty*

E_1 : *His mom says he was home on the night*

E_2 : *A large sum of money was found in his posession*

E_3 : *His fingerprints was found at the door of the bank.*

Probabilities express states-of-knowledge

$$E \equiv E_1 \text{ and } E_2 \text{ and } E_3$$

$$P(G|E) > P(G|E_2)$$

Binary propositions

A binary proposition is a statement which is either true or false (we might not know, but someone with complete knowledge would)

A : *In 49 BCE, Caesar crossed the Rubicon*

B : *Acceleration sensor 39 measures more than 0.85*

C : *Patient 901 has high cholesterol*

Propositions can be combined with **and**, **or** and **not**:

$AB \equiv$ True if A and B are both true

$A + B \equiv$ True if either A or B are true

$\overline{A} \equiv$ True if A is false

We define two special propositions which as always **true**/**false**:

1 : *A proposition which is always true*

0 : *A proposition which is always false*

...and the following identities: $A1 = A, \quad A + \overline{A} = 1, \quad \overline{\overline{A}} = A$ and

$$A(B_1 + B_2 + \cdots + B_n) = AB_1 + AB_2 + \cdots + AB_n$$

Quiz 2, Probabilities

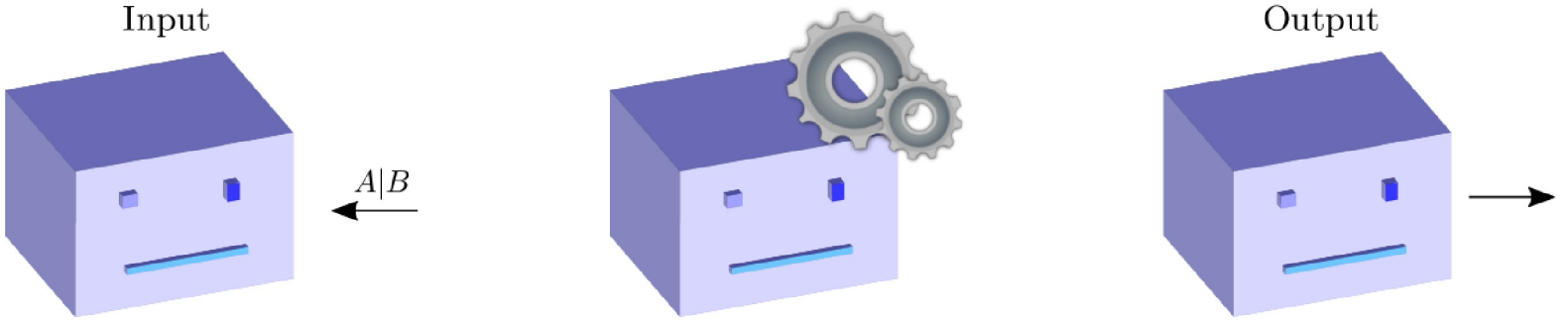
Assume we define the following 4 boolean variables.

R_1 : *Handed in report 1*
R_2 : *Handed in report 2*
R_3 : *Handed in report 3*
F : *Student failed 02450*

How would you express the probability of the statement:

If a student hand in report 1, 2 and 3, the chance of passing 02450 is greater than 90%?

A. $P(R_1R_2R_3|F) > 0.9$
B. $P(\overline{F}|R_1 + R_2 + R_3) > 0.9$
C. $P(\overline{F}|R_1R_2R_3) > 0.9$
D. $P(R_1 + R_2 + R_3|F) > 0.9$
E. Don't know.



Assuming B is true, how plausible is it A is true?

Assuming B, the plausibility of A is (low / medium / high / certain)

Passing can be defined as not failing. Therefore, express the statement as:

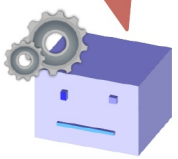
$$P(\overline{F}|R_1R_2R_3) > 0.9$$

namely, if it is true report 1, 2 and 3 are all handed in, what is the chance of passing?

Rules of probability

The sum rule: $P(A|C) + P(\overline{A}|C) = 1$

The product rule: $P(AB|C) = P(B|AC)P(A|C)$



Interpretation:

$P(A|B) = 0$ *(interpretation: given B is true, A is certainly false)*

$P(A|B) = 1$ *(interpretation: given B is true, A is certainly true)*

We also use the shorthand:

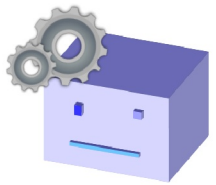
$P(A|1) = P(A)$

$P(A) + P(\overline{A}) = 1$

$P(AB) = P(A|B)P(B)$

Remarkably, this is the mathematical basis for this course

Marginalization and Bayes' theorem



DTU

Sum rule $P(A|C) + P(\overline{A}|C) = 1$

Product rule $P(AB|C) = P(B|AC)P(A|C)$

$$P(B|C) = P(B|C)\left[P(A|BC) + P(\overline{A}|BC)\right] = P(AB|C) + P(\overline{A}B|C)$$
$$= P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C).$$

Bayes' theorem: $$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|C)}$$

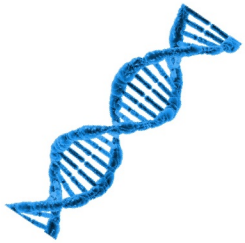
$$= \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C)}.$$

Marginalization and Bayes' theorem

$$P(B|C) = P(B|C)\left[P(A|BC) + P(\overline{A}|BC)\right] = P(AB|C) + P(\overline{A}B|C)$$
$$= P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C).$$

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|C)}$$
$$= \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C)}.$$

DNA



Bayes theorem

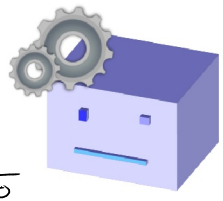
$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C)}$$

Crimes may be solved by matching crime-scene DNA to DNA in a database

- If the two samples are from the same person, a DNA test will always give a positive match
- If the DNA are from different persons, DNA will incorrectly give a positive match one time out of a million

A crime is committed in Racoon City by an unidentified male. Assume all 8000 possible perpetrators undergo a DNA test, and suppose the DNA test gives a positive result for George. What is the chance George is guilty?

G : *George is guilty*, D : *There was a positive DNA match*



$$P(G|D) = \frac{P(D|G)P(G)}{P(D|G)P(G) + P(D|\overline{G})P(\overline{G})} = \frac{1 \times \frac{1}{8000}}{1 \times \frac{1}{8000} + 10^{-6}\left(1 - \frac{1}{8000}\right)} \approx 99\%$$

Solution:

$$
\begin{aligned}
P(G|D) &= \frac{P(D|G)P(G)}{P(D|G)P(G) + P(D|\overline{G})P(\overline{G})} \\
&= \frac{1 \times \frac{1}{8000}}{1 \times \frac{1}{8000} + 10^{-6} \times \left(1 - \frac{1}{8000}\right)} \\
&= 1 - \frac{1}{126} \approx 99\%
\end{aligned}
$$

Exclusive and exchaustive events

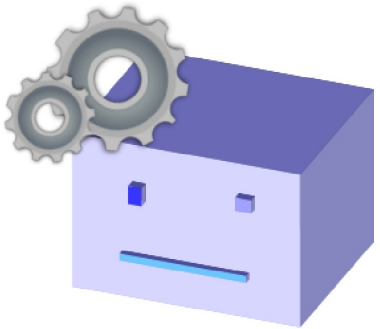
A_1 : *The side* ⚀ *face up.* A_2 : *The side* ⚁ *face up.*

A_3 : *The side* ⚂ *face up.* A_4 : *The side* ⚃ *face up.*

A_5 : *The side* ⚄ *face up.* A_6 : *The side* ⚅ *face up.*

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_i A_j = 0$ for $i \neq j$
- Consider any two events A and B

$$P(A+B) = P(A) + P(B) - P(AB)$$



- In general, for n mutually exclusive events

$$P(A_1 + A_2 + \cdots + A_n) = \sum_{i=1}^{n} P(A_i)$$

- A set of events is **exhaustive** if one has to be true: $A_1 + \cdots + A_n = 1$. Then:

$$\sum_{i=1}^{n} P(A_i) = P(A_1 + A_2 + \cdots + A_n) = 1$$

Exclusive and exhaustive events

A_1 : *The side* ⚀ *face up.* A_2 : *The side* ⚁ *face up.*

A_3 : *The side* ⚂ *face up.* A_4 : *The side* ⚃ *face up.*

A_5 : *The side* ⚄ *face up.* A_6 : *The side* ⚅ *face up.*

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_i A_j = 0$ for $i \neq j$
- Consider any two events A and B

$$\begin{aligned} P(A+B) &= 1 - P(\overline{A}\ \overline{B}) && = 1 - P(\overline{A}|\overline{B})P(\overline{B}) \\ &= 1 - \left[1 - P(A|\overline{B})\right] P(\overline{B}) && = P(B) + P(A\overline{B}) \\ &= P(B) + P(\overline{B}|A)P(A) && = P(B) + \left[1 - P(B|A)\right] P(A) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

- In general, for n mutually exclusive events $P(A_1 + A_2 + \cdots + A_n) = \sum_{i=1}^{n} P(A_i)$
- A set of events is **exchaustive** if one has to be true: $A_1 + \cdots + A_n = 1$. Then:|

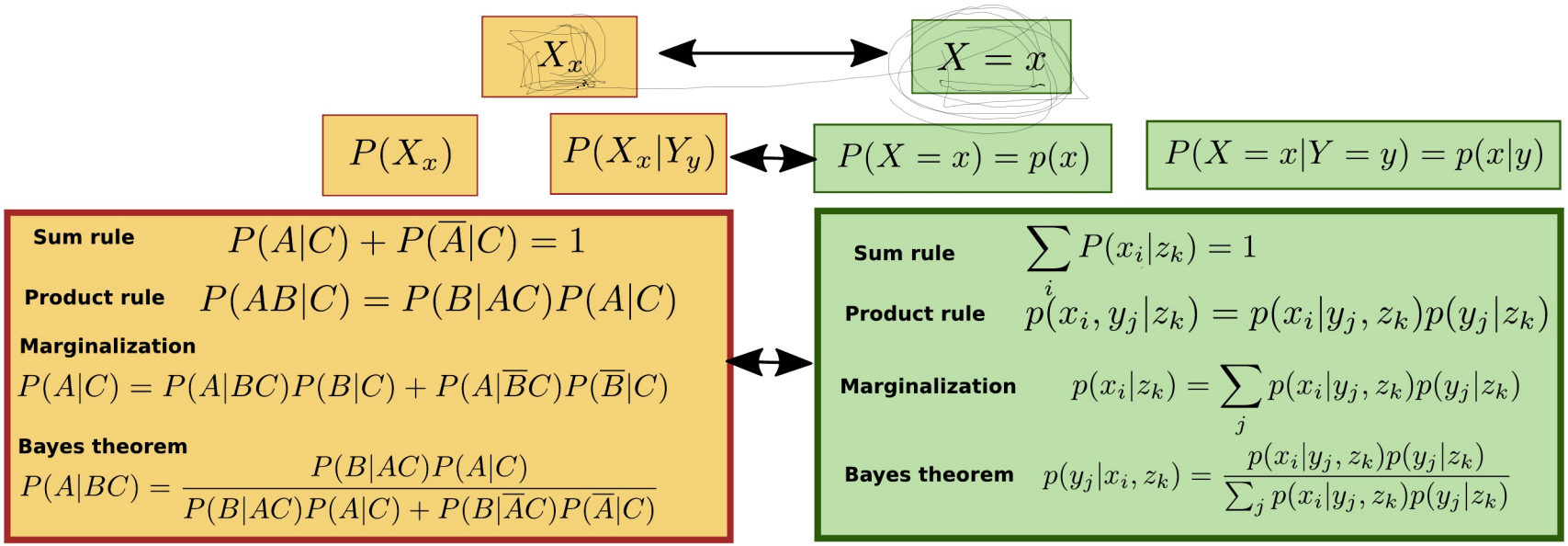
$$\sum_{i=1}^{n} P(A_i) = P(A_1 + A_2 + \cdots + A_n) = 1$$

Stochastic variables

- Often, we will measure numerical quantities (number of children, age of a patient, etc.)
- Suppose a quantity X (number of children) takes a value $x = 3$. We can write this as the binary event X_3 and in general:

$$X_x : \{\text{The binary event that } X \text{ is equal to the number } x\}$$

- Stochastic variable simplify this notation by the definition:



Quiz 3, Avila bible (Fall 2018)

$p(\tilde{x}_2, \tilde{x}_{10}\|y)$	$y=1$	$y=2$	$y=3$
$\tilde{x}_2=0,\ \tilde{x}_{10}=0$	0.19	0.3	0.19
$\tilde{x}_2=0,\ \tilde{x}_{10}=1$	0.22	0.3	0.26
$\tilde{x}_2=1,\ \tilde{x}_{10}=0$	0.25	0.2	0.35
$\tilde{x}_2=1,\ \tilde{x}_{10}=1$	0.34	0.2	0.2

Table 1: Probability of observing particular values of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on y.

We will consider a dataset based on the Avila bible. We wish to predict the copyist ($y = 0, 1, 2$) of a bible based on the two typographic attributes *upperm* and *mr/is*. We suppose the attributes have been binarized such that upperm corresponds to $\tilde{x}_2 = 0, 1$ and mr/is to $\tilde{x}_{10} = 0, 1$. Suppose the probability for each of the configurations of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on the copyist y are as given in Table 1. and the prior probability of the copyists is

$$p(y=1) = 0.316,\ p(y=2) = 0.356,\ p(y=3) = 0.328.$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

A. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.25$

B. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.313$

C. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.262$

D. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.298$

E. Don't know.

$$p(y=1 \mid x_2, x_{10}) = \frac{p(x_2, x_{10} \mid y=1)\,p(y=1)}{\sum_{k=1}^{3} p(x_2, x_{10} \mid y=k)\,p(y=k)}$$

Sum rule $\sum_i p(x_i|z_k) = 1$

Product rule $p(x_i, y_j|z_k) = p(x_i|y_j, z_k)p(y_j|z_k)$

Marginalization $p(x_i|z_k) = \sum_j p(x_i|y_j, z_k)p(y_j|z_k)$

Bayes theorem
$$p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_j p(x_i|y_j, z_k)p(y_j|z_k)}$$

The problem is solved by a simple application of Bayes' theorem:

$$
\begin{aligned}
&p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) \\
&= \frac{p(\tilde{x}_2 = 1, \tilde{x}_{10} = 0|y = 1)p(y = 1)}{\sum_{k=1}^{3} p(\tilde{x}_2 = 1, \tilde{x}_{10} = 0|y = k)p(y = k)}
\end{aligned}
$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_2 = 1, \tilde{x}_{10} = 0|y)$ in **??**. Inserting the values we see option D is correct.

Independence

Independent: $p(x_i, y_j) = p(x_i)p(y_j)$

Conditionally independent given z_k : $p(x_i, y_j|z_k) = p(x_i|z_k)p(y_j|z_k)$

Expectations

$$\textit{Expectation:} \quad \mathbb{E}[f] = \sum_{i=1}^{N} f(x_i)p(x_i). \tag{2}$$

$f(x) = x$ $\qquad$ $f(x) = (x-\mu)^2$

$$\text{mean: } \mathbb{E}[x] = \sum_{i=1}^{N} x_i p(x_i), \quad \text{Variance: } \operatorname{Var}[x] = \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2 p(x_i). \tag{3}$$

Example: Uniform probability

$$p(x_i) = \frac{1}{N}$$

$$\mathbb{E}[f] = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

$$\mathbb{E}[x] = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\operatorname{Var}[x] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models our of simpler building blocks (densities).
  In this course we will learn four:

Bernouilli density

The Catagorical density

The Beta density

The Multiviate normal density

The Bernoulli density

- Let $b = 0, 1$ denote a binary event.
- For instance, $b = 0$ corresponds to a person being well, and $b = 1$ that a person is ill.
- The probability of b is expressed using a parameter θ in the unit interval $[0, 1]$

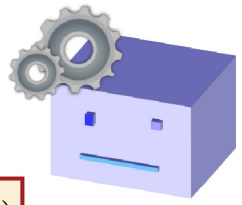
Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

$p(b=1) = p(b=1|\theta)$

$p(b=1) = \theta^1(1-\theta)^{1-1} = \theta$

$p(b=0) = \theta^0(1-\theta)^{1-0} = 1-\theta.$

The Bernoulli density, repeated events



Conditional independence $p(x_i, x_j|z_k) = p(x_i|z_k)p(x_j|z_k)$

- Suppose we observe a sequence $b_1, \ldots, b_N$ of Bernoulli (binary) events.
- For instance, for N patients we record whether person 1 is well or ill ($b_1 = 0$ or $b_1 = 1$) and up to whether patient N is ill or well ($b_N = 0$ or $b_N = 1$)
- When we **know** θ (the chance a person is well or ill), the events are **independent**

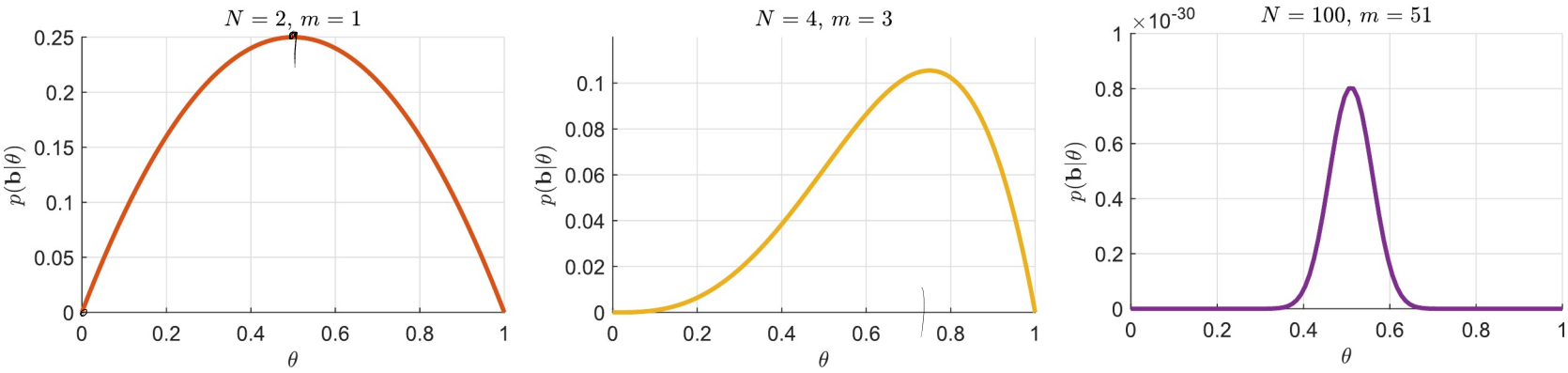
Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

conditional independence.

$$p(b_1, \cdots, b_N|\theta) = p(b_1|\theta) \times \cdots p(b_N|\theta) = \prod_{i=1}^{N} \theta^{b_i}(1-\theta)^{1-b_i} = \theta^{\sum_{i=1}^{N} b_i}(1-\theta)^{N-\sum b_i}$$

$$= \theta^m(1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N$$

The Bernoulli density, maximum likelihood



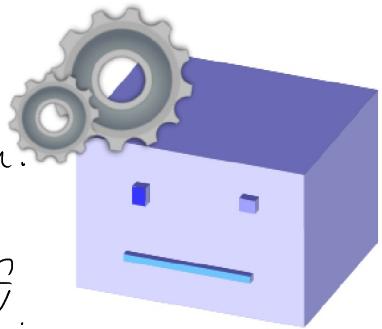
$$p(b_1, \ldots, b_N|\theta) = \theta^m(1-\theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

$$\theta^* = \arg\max_\theta p(b_1, \ldots, b_N|\theta)$$

algebra.

$= \frac{m}{N}$.



The value of θ according to which the data is most plausible

$D = (X, y)$ $\boxed{p(X, y \mid \theta)}$, $\theta^* = \arg\max_\theta p(X, y \mid \theta)$

Resources

https://02402.compute.dtu.d A more in-depth description of summary statistics (see chapter 1) (https://02402.compute.dtu.dk)

https://www.khanacademy.org An excellent introduction to probability theory which we recommend as a go-to resource (https://www.khanacademy.org/math/statistics-probability/probability-library)

http://citeseerx.ist.psu.edu A more in-depth discussion of Bayes in the court room (http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EFE0328140036A4C668BB5B9FC76C9BE?doi=10.1.1.599.8675&rep=rep1&type=pdf)