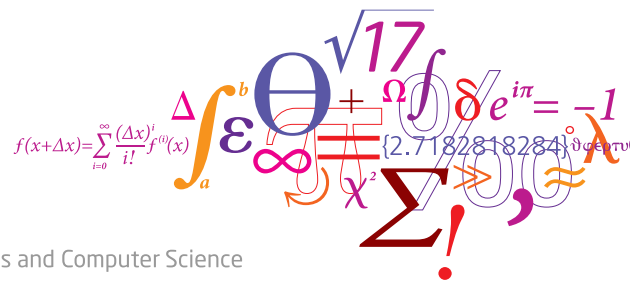
02450: Introduction to Machine Learning and Data Mining

Measures of similarity, summary statistics and probabilities

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

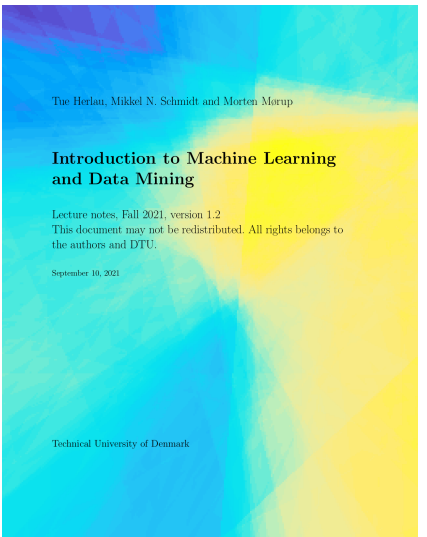
Today

Feedback Groups of the day:

Akshat Bhardwaj, Aldis Helga Bjorgvinsdottir, Alysha K Chamadia, Andreas Holmer Bigom, Andrew Boonto Blumensen, Anja Lykke Borre, Anna Brønd, Ashish Chawla, Aske Bruun-Ringgaard, August Hertz Bugge, August Valentin Nørgaard Birch, Belen Castellote Lopez, Benjamin Bogø, Bogdan Capsa, Calle Ryge Carlsen, Chinmay Bhalla, Christoffer Binzer Bjørner, Connor Ward Chewning, Dimitris Bokos-Zygouris, Frederik Bjørling Bornemann, Frederik Jakob Eskildsen Bruun, Fridtjof Cerup-Simonsen, Hanna Maria Börtin, Helga Þórey Björnsdóttir, Jakob Friis Christiansen, Joana Cardiff Aleu, Johan Matzen Brodt, Johan Stauner Bill, João Augusto César Moutinho, Juan Cervera Bustamante, Juliane Bjørn Budde, Julien François Per Chegaray, Jun Chen, Kian Bostani Nezhad, Kunt Alp Celebi, Lars Knuth Satoshi Boyens-Thiele, Lorenzo Capaldo, Ludovica Caccaro, Marion Josephine Isabelle Maëlys Cadet, Mikkel Wøidemann Klæbel Blomsterberg, Niels Karsten Bisgaard-Bohr, Oriol Cayon Domingo, Oscar Carpentier, Pau Carrascal Fabregat, Peter Sebastian Hein Bitsch, Peter Tønder Blendstrup, Philipp Bockshecker, Piet Johan Brochorst Christensen, Raquel Maria Casañ Crespo, Rares-Victor Botis, Ruixin Chen, Rune Yding Brogaard, Samy Chekkouri, Simon Buk-Mortensen, Sueanoi Choksawas, Tyme Chatupanyachotikul, Xenia Cohen Chime, Yuechen Chen, Zhe Chen

Reading material:

Chapter 4, Chapter 5



Lecture Schedule

1. Introduction
   31 August: C1

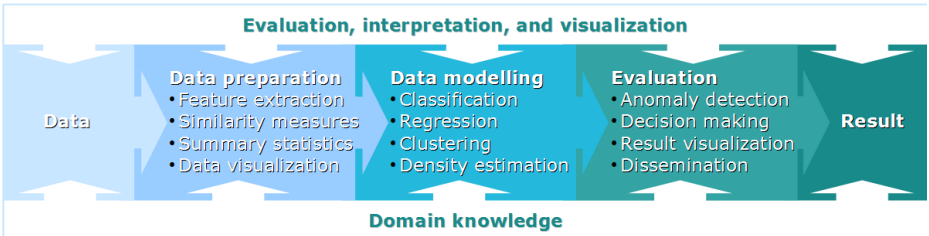
   Data: Feature extraction, and visualization
2. Data, feature extraction and PCA
   7 September: C2, C3
3. **Measures of similarity, summary statistics and probabilities**
   14 September: C4, C5
4. Probability densities and data visualization
   21 September: C6, C7

   Supervised learning: Classification and regression
5. Decision trees and linear regression
   28 September: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   5 October: C10, C12 **(Project 1 due before 13:00)**
7. Performance evaluation, Bayes, and Naive Bayes
   12 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   26 October: C14, C15
9. AUC and ensemble methods
   2 November: C16, C17

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

    Recap
13. Recap and discussion of the exam
    30 November: C1-C21

Online help: Forum on DTU Learn
Videos of lectures: `https://video.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)



Learning Objectives

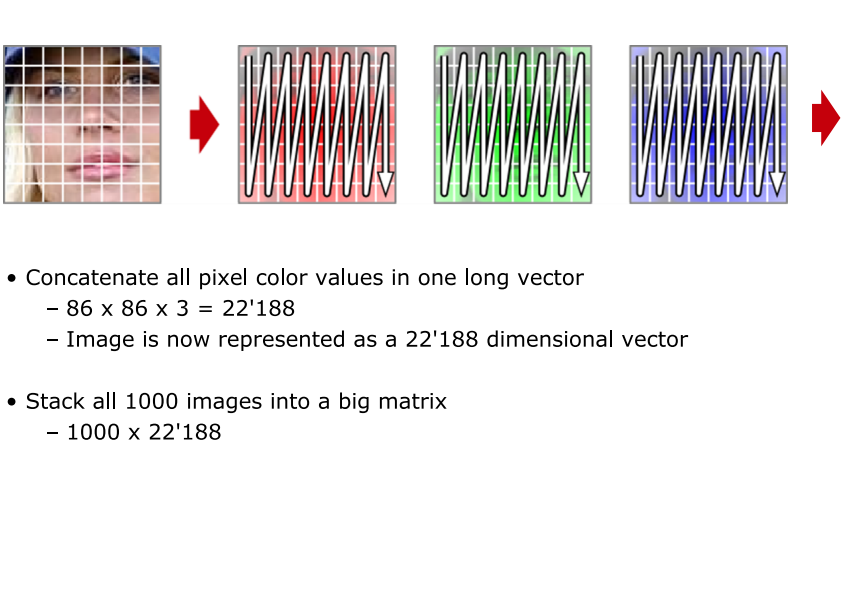
- Compute measures of similarity/dissimilarity (Lp distance, cosine distance, etc.)
- Understand basics of probability theory and stochastic variables in the discrete setting
- Understand probabilistic concepts such as expectations, independence and the Bernoulli distribution
- Understand the maximum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

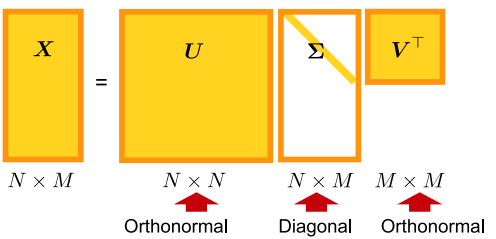
Pre-processing



- 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

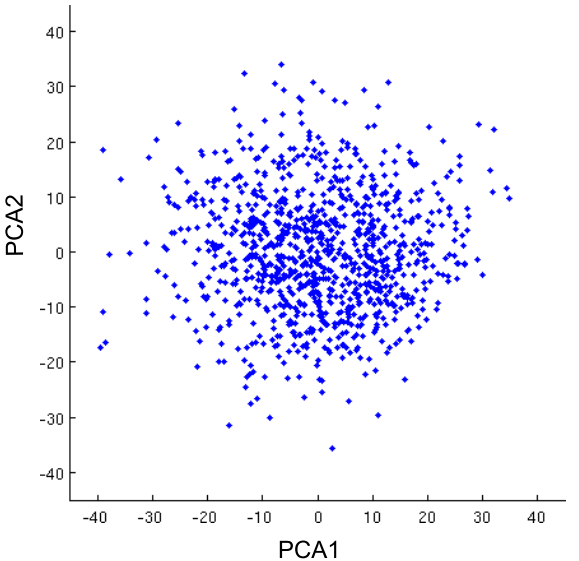
Principal component analysis (PCA)

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

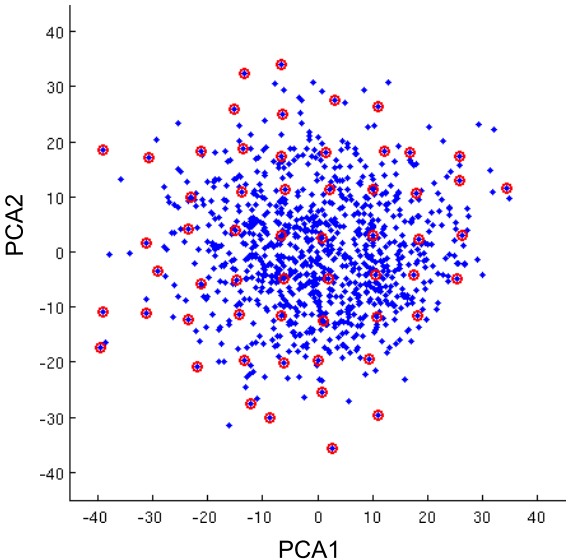


- **Plot data in the transformed coordinate system**
  – Corresponds to looking at data from an angle where it is most spread out

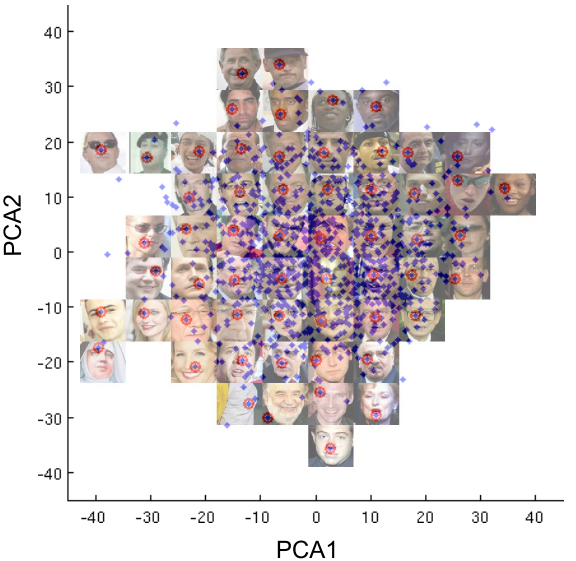
PCA on face images



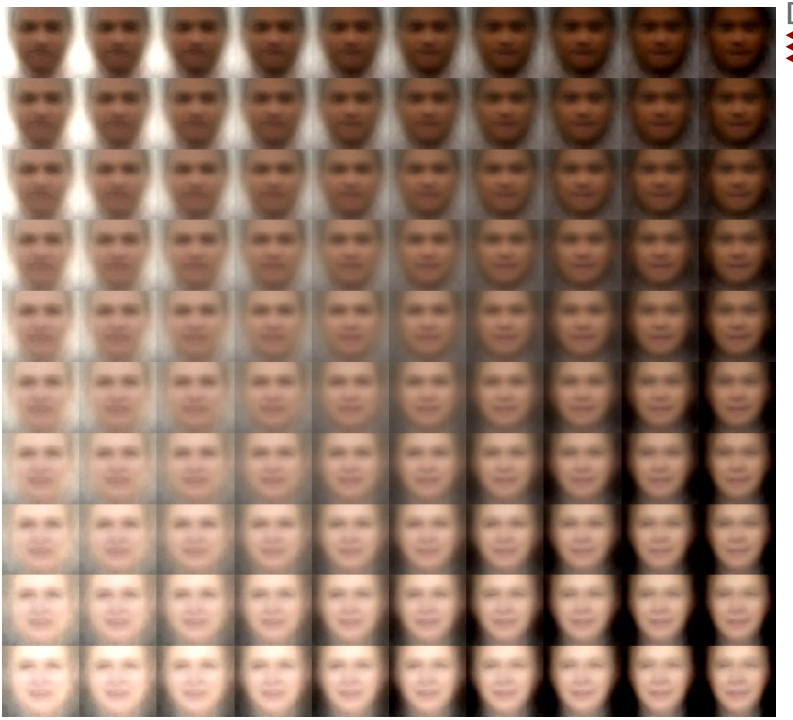
PCA on face images



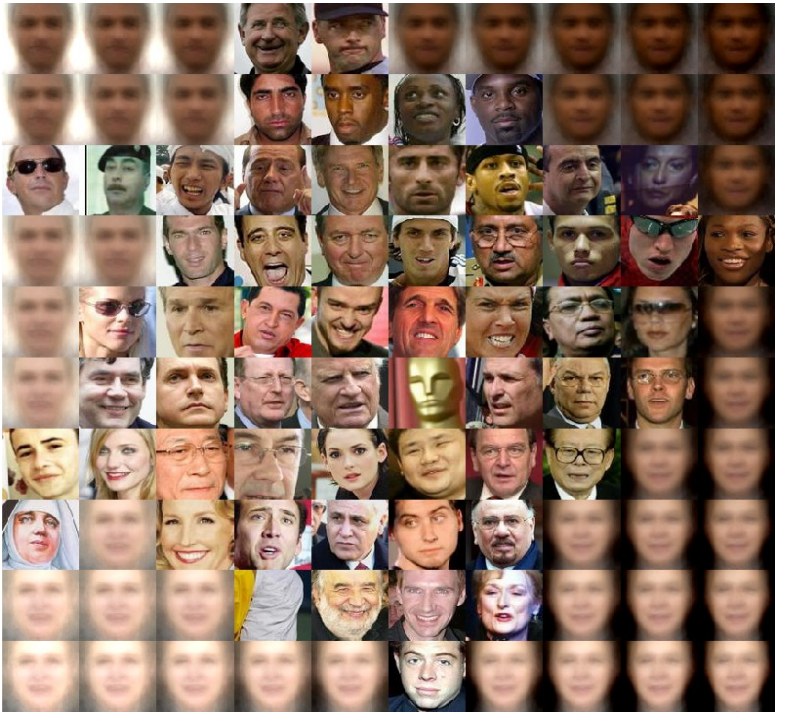
PCA on face images



- 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 it is **most similar**/**least dissimilar** to.

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

Dissimilarity measures

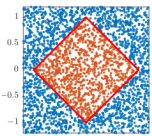
Minkowsky is the general rule: the other are derivations

- 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 ($p = 1$)

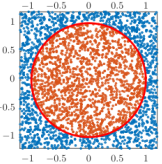
it'd be good if you don't want to be sensitive

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



- Euclidean ($p = 2$)

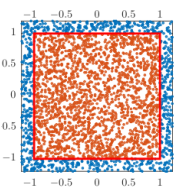
$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{M} (x_j - y_j)^2}$$



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

good choice for looking for outliers

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



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

Imagine X and Y binary vectors

Similarity measures

$\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$

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)

it's symmetric in 0 and 1, if I change the 0s with the 1s in the 2 vectors, the SMC value doesn't change: useful because sometimes 0s or 1s could be not significative

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

different vectors size means the number of non-0s entries in the vector

+ Symmetric

Jaccard Coefficient

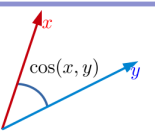
only focuses in positive matches

but if the two vectors are different size we might get high value -> solved with the COSINE

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

+ Positive matches

Cosine similarity



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

divide by the norm, and focuses on positive matches

works for both binary vectors but also for any other kind

+ 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 continuous data

||x^2|| counts the number of non-0s in the vector

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.

F00 = 1
F11 = 2
k = 5
x.T * y = 2
||x|| = 3^(1/2) = 0^2 +1^2 +1^2 .. = square root of non-0s entries

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
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.

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

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

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

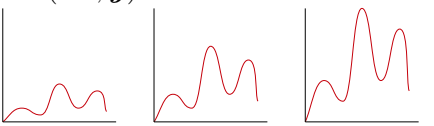
$$\mathrm{EJ}(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

Invariance

Scale invariance

the COSINE similarity is scale invariant

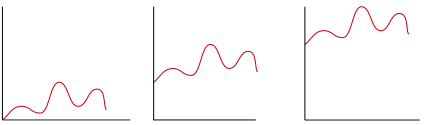
$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha \boldsymbol{x}, \boldsymbol{y})$$



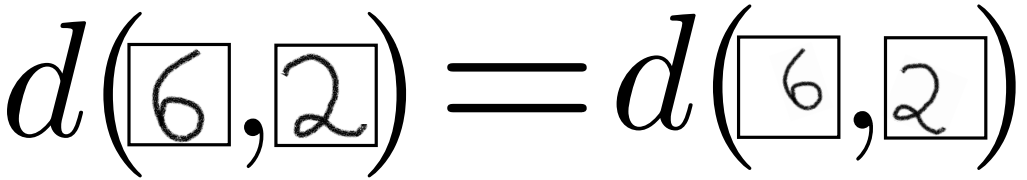
Translation invariance

required in image recognition for isntance, in I scale the image, or add some light, or rotate, I want my method to be robust

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



General invariances



Transformations

Standardization: Ensure a single attribute will not dominate:

IMPORTANT = the scale of the 3 attributes is really different, if I don't standarize them, all the differences will focus mainly on the income.
If I standarize when not needed I may lose some information

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

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

Combining heterogeneous attributes Transform measures and combine

- we first get a similarity measure between the education
- S = 1 / (1+D) : if I want to transform a Dissimilarity measure into a Similarity one.

$$s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

the choose of the function 'a' depends on the application

$$s_{\text{Age.}} = a\left(a + d_1(x_{\text{Age.}}, y_{\text{Age.}})\right)^{-1}, \ a = 1$$

to combine the two
- use the mean
- a particular weight

$$s(x, y) = \frac{1}{2}\left(s_{\text{Edu.}} + s_{\text{Age.}}\right)$$

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 $x_1, x_2, \ldots, x_N \in \mathbb{R}$ and $y_1, y_2, \ldots, y_N \in \mathbb{R}$:

- 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

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

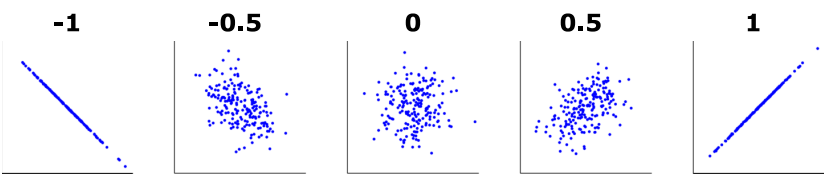
Correlation

- Measure of degree of linear relationship

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

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

$$x_k = a y_k + b$$



Quantiles

Given N observations of an attribute $x_1, x_2, \ldots, x_N \in \mathbb{R}$.

Quantiles describe the *points* that divide the underlying distribution into intervals of equal probably:

- The one 2-quantile (**median**) divides the distribution in two intervals.
- The three 4-quantiles (**quartiles**) divides the distribution in four intervals.
- The 99 100-quantiles (**percentiles**) divides the distribution in 100 intervals.

The **median** is the same as the 2nd quartile or the 50th percentile.

E.g., we can (approximately) find the **median** by

- Sort the observations in ascending order $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$

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

Probabilities

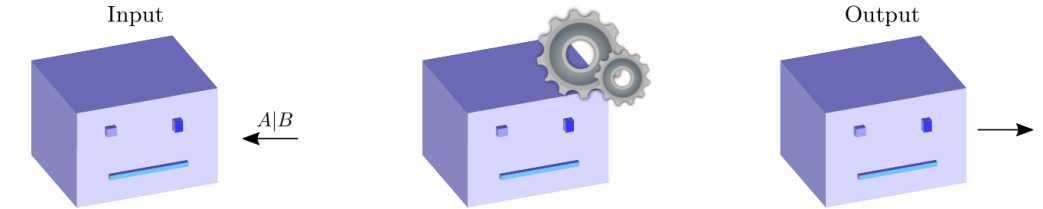
Continuous optimizazion properties: if I have an image classification, I want the output says that the one shown is a cat with a certain amount of probability, instead of saying only is a cat or dog (discrete optimizazion)

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

Algorithmically: **If** an 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:

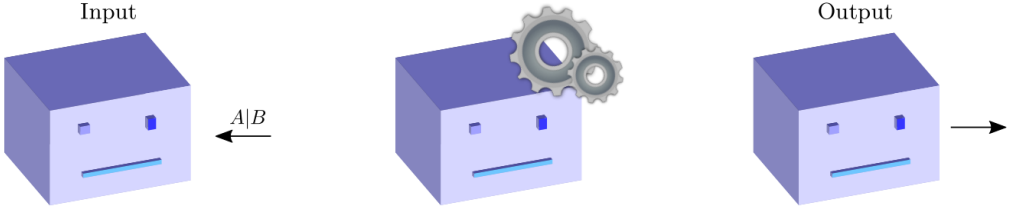
the probability of A given B is x -> the degree of probability that A is true knowing that B is true, is equal to x

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

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



G : *The accused is guilty*

E_1 : *A car similar to his was seen at the crime scene.*

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

the probability of A given B is x -> the degree of probability that A is true knowing that B is true, is equal to X

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

$$P(G|E) > P(G|E_2)$$ if we get more knowledge, we become more certain about somethng

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**:

$$AB \equiv \text{True if } A \text{ and } B \text{ are both true}$$
$$A + B \equiv \text{True if either } A \text{ or } B \text{ are true}$$
$$\overline{A} \equiv \text{True if } A \text{ is false}$$

We define two special propositions which is 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%?

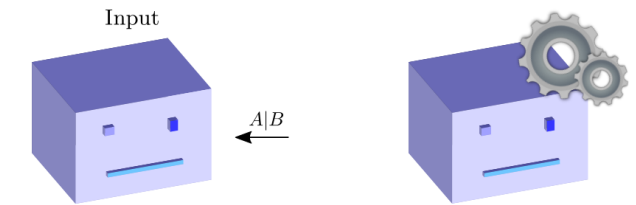
A. $P(R_1R_2R_3|F) > 0.9$

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.



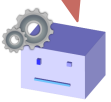
Assuming B is true, how plausible is it A is true?

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

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

Marginalization

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

Bayes theorem

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

p(AB) = p(B|A) * p(A)
p(AB) = p(A|B) * p(B)
p(A|B) * p(B) = p(B|A) * p(A) -- > p(A|B) = p(B|A) * p(A)

DNA

Bayes theorem

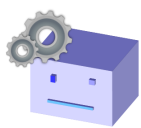
$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\overline{A}C)P(\overline{A}|C)}$$

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*



$$p(G|D) = \frac{p(D|G) * P(G)}{p(D)}$$

$$p(D) = p(GD) + p(!GD) = p(D|G) * p(G) + p(D|!G) * p(1 - !G) =$$
$$= 1 * 1/8000 \quad + \quad 10\text{^-}6 * (1 - 1/8000) = 99\%$$

$$p(D|G) = 1$$
$$p(G) = 1/8000$$

Exclusive and exhaustive events

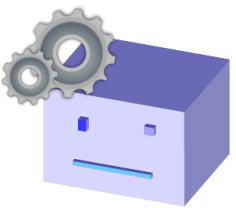
A_1 : *The side* ⚀ *face up.* A_2 : *The side* ⚁ *face up.*

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

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



- 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:

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

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:

$X_x \longleftrightarrow X = x$

$P(X_x)$ $\quad$ $P(X_x|Y_y) \longleftrightarrow P(X = x) = p(x)$ $\quad$ $P(X = x|Y = y) = p(x|y)$

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

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

Marginalization

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

Bayes theorem

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

$\longleftrightarrow$

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 $p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_{j'} p(x_i|y_{j'}, z_k)p(y_{j'}|z_k)}$

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.

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
$$p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_{j'} p(x_i|y_{j'}, z_k)p(y_{j'}|z_k)}$$

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

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

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

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 out of simpler building blocks (distributions (discrete variables see chapter 5) and densities (continuous variables see chapter 6)). In this course we will use four:

  Bernoulli distribution

  The Categorical distribution

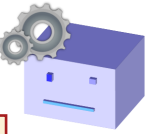
  The Beta density

  The Multiviate normal density

The Bernoulli distribution

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

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



The Bernoulli distribution, 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 ill or well ($b_1 = 0$ 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**

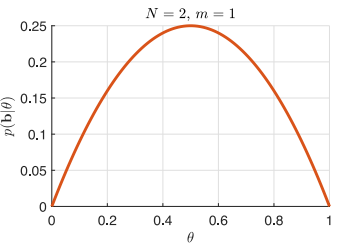
Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

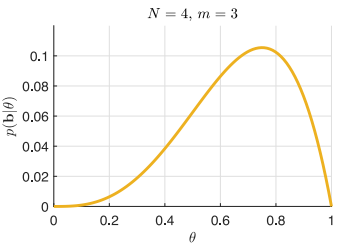
$$p(b_1, \cdots, b_N|\theta) = \prod_{i=1}^{N} p(b_i|\theta) = \prod_{i=1}^{N} \theta^{b_i}(1-\theta)^{1-b_i} = \theta^{\sum_{i=1}^{N} b_i}(1-\theta)^{N-\sum_{i=1}^{N} b_i}$$

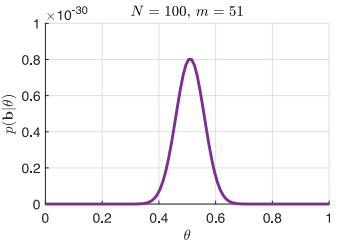
product of N bernoulli distribution

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

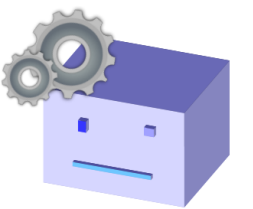
The Bernoulli distribution, maximum likelihood







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



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

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

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

Resources

https://bayes.wustl.edu Classical textbook which treats probabilities as states-of-knowledge and discuss many practical and philosophical issues (this book converted me to ML!)

(https://bayes.wustl.edu/etj/prob/book.pdf)

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)