02450: Introduction to Machine Learning and Data Mining

Measures of similarity, summary statistics and probabilities

Morten Mørup

DTU Compute, Technical University of Denmark (DTU)

Lecture Schedule

1. Introduction
   7 October: C1

Data: Feature extraction, and visualization

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

Supervised learning: Classification and regression

5. Decision trees and linear regression
   8 October: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   9 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   9 October: C14, C15
9. AUC and ensemble methods
   10 October: C16, C17

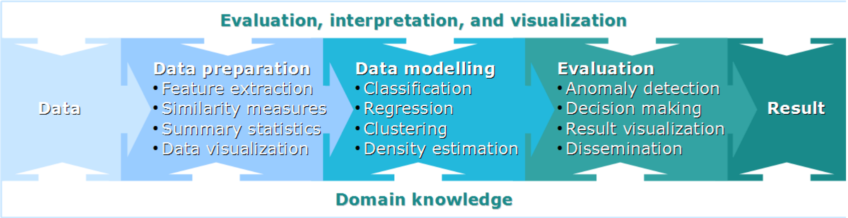
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    10 October: C18
11. Mixture models and density estimation
    11 October: C19, C20
12. Association mining
    11 October: C21

Recap

13. Recap
    11 October: C1-C21

Piazza online help: `https://piazza.com/dtu.dk/fall2019/october2019`



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

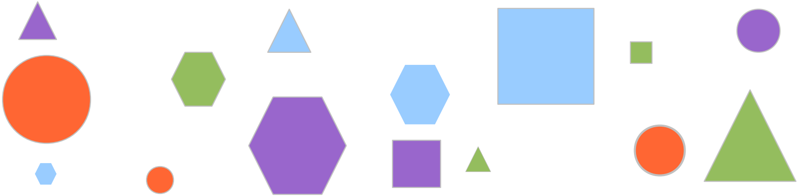
Similarity / Dissimilarity measures

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.

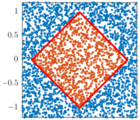
Outlier detection The observation most **dissimilar** to all other observations is an outlier



Dissimilarity measures

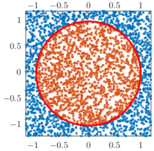
- General Minkowsky distance (p-distance) $d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}$
- One-norm ($p = 1$)

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



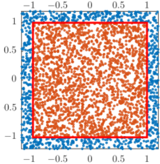
- Euclidean ($p = 2$)

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



- 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

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

K : Total number of attributes

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

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

Simple Matching Coefficient (SMC)

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

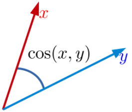
+ Symmetric

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.

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.

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 $x_k = y_k = 1$

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

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

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

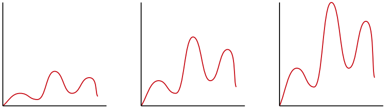
$$\mathrm{EJ}(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

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

Invariance

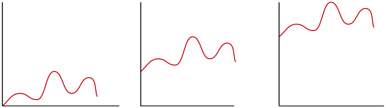
Scale invariance

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

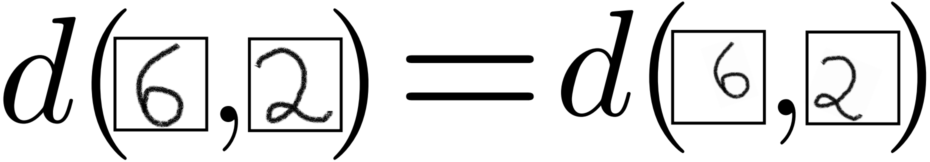


Translation invariance

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



General invariances



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

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

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

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

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

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

- Empirical covariance

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

- Empirical standard deviation

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

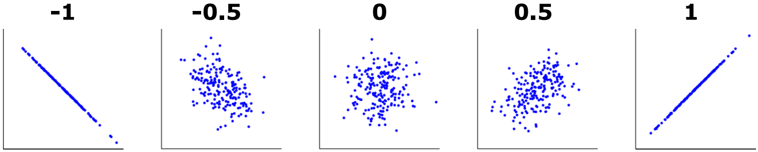
Correlation

- 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 $x_1, x_2, \ldots, x_N$. The **q'th quantile** is the value x_q of x such that a fraction q of the sample is **smaller** than 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:

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

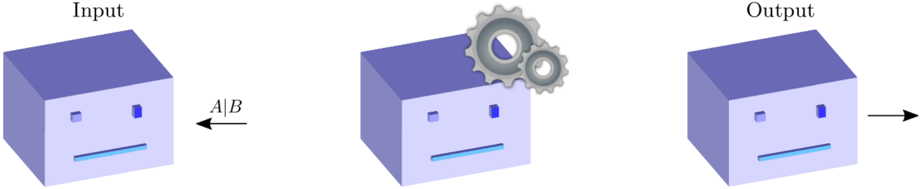
Probabilities

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



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

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

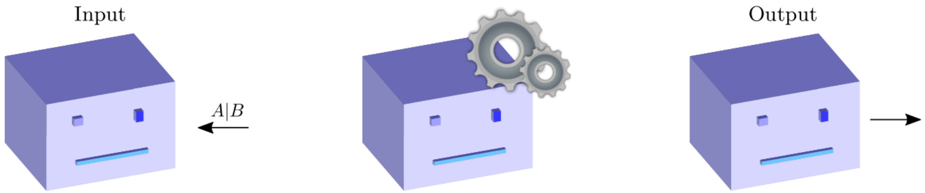
We reason about a proposition A in light of evidence B:

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

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**:

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

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

1 : *A proposition which is always true*

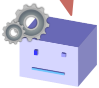
0 : *A proposition which is always false*

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

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

Rules of probability

$$\textit{The sum rule:} \quad P(A|C) + P(\overline{A}|C) = 1$$
$$\textit{The product rule:} \quad P(AB|C) = P(B|AC)P(A|C)$$



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:

$$\boxed{P(A|1) = P(A)}$$

$$p(A) + P(\overline{A}) = 1$$
$$p(AB) = P(A|B)P(B)$$

Remarkably, this is the mathematical basis for this course

Marginalization and Bayes' theorem

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

$$P(B|C) =$$

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

Bayes' theorem: $P(A|BC) =$

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

Marginalization and Bayes' theorem

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

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

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*, $\quad D$: *There was a positive DNA match*

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 exhaustive events

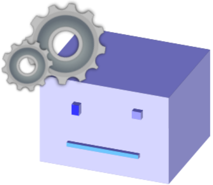
A_1 : *The side ⚀ face up.* A_2 : *The side ⚁ face up.*

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

$$P(A + B) =$$



- 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 **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:

$$X_x \longleftrightarrow X = x$$

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

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

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

$\longleftrightarrow$

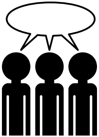
Sum rule $\sum_i P(x_i|z_k) = 1$

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 $p(y_j|x_i, z_k) = \dfrac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_j p(x_i|y_j, z_k)p(y_j|z_k)}$

Quiz 2, Medical diagnosis



A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text: *(x=Positive, y=0: no disease, y=1: Disease)*

  p(Positive|Disease)

  p(Positive|No Disease)

  p(Disease)

  p(No Disease)
- Use the basic rules of probability given to the right to find:

  p(Disease|Positive)

$$p(y) = \sum_x p(y,x)$$

$$= p(y|x)p(x) + p(y|\overline{x})p(\overline{x})$$

$$p(x,y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Independence

$$\textit{Independent:} \quad p(x_i, y_j) = p(x_i)p(y_j)$$
$$\textit{Conditionally independent given } z_k: \quad p(x_i, y_j | z_k) = p(x_i|z_k)p(y_j|z_k)$$

Expectations

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

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

Example: Uniform probability

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

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

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

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

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 Categorical 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 $[0, 1]$

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

The Bernoulli density, repeated events

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

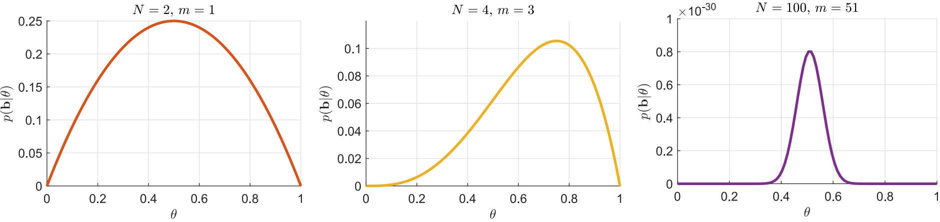
- Suppose we observe a sequence $b_1, \ldots, b_N$ of Bernoulli (binary) events.
- For instance, for N patients we record whether person 1 is well or ill ($b_1 = 0$ or $b_1 = 1$) and up to whether patient N is ill or well ($b_N = 0$ or $b_N = 1$)
- When we **know** θ (the chance a person is well or ill),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

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

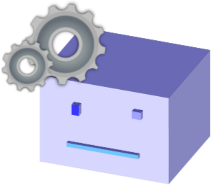
$$p(b_1, \cdots, b_N | \theta) =$$

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

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**

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

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

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)