Probabilistic Modelling and Reasoning
— Introduction —

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, University of Edinburgh

Spring semester 2018
Variability is part of nature

- Human heights vary
- Men are typically taller than women but height varies a lot

Data from U.S. CDC, adults ages 18-86 in 2007
Variability

- Our handwriting is unique
Variability

- Our handwriting is unique
- Variability leads to uncertainty: e.g. 1 vs 7 or 4 vs 9
Variability

- Variability leads to uncertainty
- Reading handwritten text in a foreign language
Example: Screening and diagnostic tests

- Early warning test for Alzheimer’s disease  (Scharre, 2010, 2014)
- Detects “mild cognitive impairment”

- Takes 10–15 minutes
- Freely available

---

7. Copy this picture:

8. Drawing test
- Draw a large face of a clock and place in the numbers
- Position the hands for 5 minutes after 11 o’clock

(Example from sagetest.osu.edu)
Example: Screening and diagnostic tests

- Early warning test for Alzheimer’s disease  (Scharre, 2010, 2014)
- Detects “mild cognitive impairment”

- Takes 10–15 minutes
- Freely available
- Assume a 70 year old man tests positive.
- Should he be concerned?

7. Copy this picture:

8. Drawing test
   - Draw a large face of a clock and place in the numbers
   - Position the hands for 5 minutes after 11 o’clock

(Example from sagetest.osu.edu)
Accuracy of the test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
Accuracy of the test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
- 80% correct for people with impairment

![Diagram showing accuracy of the test with sensitivity of 0.8 and specificity of 0.95]
Accuracy of the test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
- 95% correct for people w/o impairment
Variability implies uncertainty

- People of the same group do not have the same test results
  - Test outcome is subject to variability
  - The data are noisy
Variability implies uncertainty

- People of the same group do not have the same test results
  - Test outcome is subject to variability
  - The data are noisy
- Variability leads to uncertainty
  - Positive test $\equiv$ true positive ?
  - Positive test $\equiv$ false positive ?
Variability implies uncertainty

- People of the same group do not have the same test results
  - Test outcome is subject to variability
  - The data are noisy
- Variability leads to uncertainty
  - Positive test $\equiv$ true positive ?
  - Positive test $\equiv$ false positive ?
- What can we safely conclude from a positive test result?
- How should we analyse such kind of ambiguous data?
Probabilistic approach

- The test outcomes $y$ can be described with probabilities

  \[
  \text{sensitivity} = 0.8 \iff \Pr(y = 1| x = 1) = 0.8 \\
  \iff \Pr(y = 0| x = 1) = 0.2
  \]

  \[
  \text{specificity} = 0.95 \iff \Pr(y = 0| x = 0) = 0.95 \\
  \iff \Pr(y = 1| x = 0) = 0.05
  \]

- $\Pr(y|x)$: model of the test specified in terms of (conditional) probabilities

- $x \in \{0, 1\}$: quantity of interest (cognitive impairment or not)
Among people like the patient, $\Pr(x = 1) = 5/45 \approx 11\%$ have a cognitive impairment (plausible range: 3% – 22%, Geda, 2014)
Probabilistic model

- Reality:
  - properties/characteristics of the group of people like the patient
  - properties/characteristics of the test

- Probabilistic model:
  - $\Pr(x = 1)$
  - $\Pr(y = 1|x = 1)$ or $\Pr(y = 0|x = 1)$
  - $\Pr(y = 1|x = 0)$ or $\Pr(y = 0|x = 0)$

Fully specified by three numbers.
Probabilistic model

- **Reality:**
  - properties/characteristics of the group of people like the patient
  - properties/characteristics of the test
- **Probabilistic model:**
  - \( \Pr(x = 1) \)
  - \( \Pr(y = 1|x = 1) \) or \( \Pr(y = 0|x = 1) \)
  - \( \Pr(y = 1|x = 0) \) or \( \Pr(y = 0|x = 0) \)

  Fully specified by three numbers.

- A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.
If we tested the whole population

With impairment
\( p(x=1) \)

Without impairment
\( p(x=0) \)
If we tested the whole population

Fraction of people who are impaired and have positive tests:

\[ Pr(x = 1, y = 1) = \frac{4}{5} \cdot \frac{5}{45} = \frac{4}{45} \quad (4/5 = 0.8) \]
If we tested the whole population

Fraction of people who are impaired and have positive tests:

\[ Pr(x = 1, y = 1) = Pr(y = 1|x = 1) Pr(x = 1) = 4/45 \]  

(product rule)
If we tested the whole population

Fraction of people who are not impaired but have positive tests:

$$\Pr(x = 0, y = 1) = \frac{2}{40} \cdot \frac{40}{45} = \frac{2}{45} \quad (\frac{2}{40} = 0.05)$$
If we tested the whole population

Fraction of people who are not impaired but have positive tests:

$$\Pr(x = 0, y = 1) = \Pr(y = 1|x = 0) \Pr(x = 0) = \frac{2}{45} \quad \text{(product rule)}$$
If we tested the whole population

Fraction of people where the test is positive:

\[ \Pr(y = 1) = \Pr(x = 1, y = 1) + \Pr(x = 0, y = 1) = \frac{6}{45} \]  
(sum rule)
Putting everything together

▶ Among those with a positive test, fraction with impairment:

\[
\Pr(x = 1|y = 1) = \frac{\Pr(y = 1|x = 1) \Pr(x = 1)}{\Pr(y = 1)} = \frac{4}{6} = \frac{2}{3}
\]

▶ Fraction without impairment:

\[
\Pr(x = 0|y = 1) = \frac{\Pr(y = 1|x = 0) \Pr(x = 0)}{\Pr(y = 1)} = \frac{2}{6} = \frac{1}{3}
\]

▶ Equations are examples of “Bayes’ rule”.

Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%, or from 6% to 50%.

\[50\% \equiv \text{coin flip}\]
Putting everything together

- Among those with a positive test, fraction with impairment:

\[
\Pr(x = 1|y = 1) = \frac{\Pr(y = 1|x = 1) \Pr(x = 1)}{\Pr(y = 1)} = \frac{4}{6} = \frac{2}{3}
\]

- Fraction without impairment:

\[
\Pr(x = 0|y = 1) = \frac{\Pr(y = 1|x = 0) \Pr(x = 0)}{\Pr(y = 1)} = \frac{2}{6} = \frac{1}{3}
\]

- Equations are examples of “Bayes’ rule”.

- Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%,
Putting everything together

Among those with a positive test, fraction with impairment:

\[
\Pr(x = 1 | y = 1) = \frac{\Pr(y = 1 | x = 1) \Pr(x = 1)}{\Pr(y = 1)} = \frac{4}{6} = \frac{2}{3}
\]

Fraction without impairment:

\[
\Pr(x = 0 | y = 1) = \frac{\Pr(y = 1 | x = 0) \Pr(x = 0)}{\Pr(y = 1)} = \frac{2}{6} = \frac{1}{3}
\]

Equations are examples of “Bayes’ rule”.

Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%, or from 6% to 50%.
Putting everything together

- Among those with a positive test, fraction with impairment:

\[
\Pr(x = 1|y = 1) = \frac{\Pr(y = 1|x = 1) \Pr(x = 1)}{\Pr(y = 1)} = \frac{\frac{4}{6}}{\frac{2}{3}} = \frac{2}{3}
\]

- Fraction without impairment:

\[
\Pr(x = 0|y = 1) = \frac{\Pr(y = 1|x = 0) \Pr(x = 0)}{\Pr(y = 1)} = \frac{\frac{2}{6}}{\frac{1}{3}} = \frac{1}{3}
\]

- Equations are examples of “Bayes’ rule”.

- Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%, or from 6% to 50%.

- 50% ≡ coin flip
Probabilistic reasoning

- Probabilistic reasoning $\equiv$ probabilistic inference: Computing the probability of an event that we have not or cannot observe from an event that we can observe
  - Unobserved/uncertain event, e.g. cognitive impairment $x = 1$
  - Observed event $\equiv$ evidence $\equiv$ data, e.g. test result $y = 1$
Probabilistic reasoning

- Probabilistic reasoning ≡ probabilistic inference: Computing the probability of an event that we have not or cannot observe from an event that we can observe
  - Unobserved/uncertain event, e.g. cognitive impairment $x = 1$
  - Observed event ≡ evidence ≡ data, e.g. test result $y = 1$

- “The prior”: probability for the uncertain event before having seen evidence, e.g. $\Pr(x = 1)$

- “The posterior”: probability for the uncertain event after having seen evidence, e.g. $\Pr(x = 1|y = 1)$
Probabilistic reasoning

- Probabilistic reasoning ≡ probabilistic inference: Computing the probability of an event that we have not or cannot observe from an event that we can observe
  - Unobserved/uncertain event, e.g. cognitive impairment $x = 1$
  - Observed event ≡ evidence ≡ data, e.g. test result $y = 1$
- “The prior”: probability for the uncertain event before having seen evidence, e.g. $\Pr(x = 1)$
- “The posterior”: probability for the uncertain event after having seen evidence, e.g. $\Pr(x = 1|y = 1)$
- The posterior is computed from the prior and the evidence via Bayes’ rule.
Key rules of probability

(1) Product rule:

\[ \Pr(x = 1, y = 1) = \Pr(y = 1|x = 1) \Pr(x = 1) \]
\[ = \Pr(x = 1|y = 1) \Pr(y = 1) \]

(2) Sum rule:

\[ \Pr(y = 1) = \Pr(x = 1, y = 1) + \Pr(x = 0, y = 1) \]
Key rules of probability

(1) Product rule:

\[ \Pr(x = 1, y = 1) = \Pr(y = 1|x = 1) \Pr(x = 1) = \Pr(x = 1|y = 1) \Pr(y = 1) \]

(2) Sum rule:

\[ \Pr(y = 1) = \Pr(x = 1, y = 1) + \Pr(x = 0, y = 1) \]

Bayes’ rule (conditioning) as consequence of the product rule

\[ \Pr(x = 1|y = 1) = \frac{\Pr(x = 1, y = 1)}{\Pr(y = 1)} = \frac{\Pr(y = 1|x = 1) \Pr(x = 1)}{\Pr(y = 1)} \]

Denominator from sum rule, or sum rule and product rule

\[ \Pr(y = 1) = \Pr(y = 1|x = 1) \Pr(x = 1) + \Pr(y = 1|x = 0) \Pr(x = 0) \]
Key rules or probability

- The rules generalise to the case of multivariate random variables (discrete or continuous)
- Consider the conditional joint probability density function (pdf) or probability mass function (pmf) of $x, y$: $p(x, y)$

1. Product rule:

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

2. Sum rule:

$$p(y) = \begin{cases} \sum_x p(x, y) & \text{for discrete r.v.} \\ \int p(x, y)dx & \text{for continuous r.v.} \end{cases}$$
Probabilistic modelling and reasoning

- **Probabilistic modelling:**
  - Identify the quantities that relate to the aspects of reality that you wish to capture with your model.
  - Consider them to be random variables, e.g. $x, y, z$, with a joint pdf (pmf) $p(x, y, z)$.
Probabilistic modelling and reasoning

- Probabilistic modelling:
  - Identify the quantities that relate to the aspects of reality that you wish to capture with your model.
  - Consider them to be random variables, e.g. $x, y, z$, with a joint pdf (pmf) $p(x, y, z)$.

- Probabilistic reasoning:
  - Assume you know that $y \in \mathcal{E}$ (measurement, evidence)
  - Probabilistic reasoning about $x$ then consists in computing

$$p(x | y \in \mathcal{E})$$

or related quantities like $\text{argmax}_x p(x | y \in \mathcal{E})$ or posterior expectations of some function $g$ of $x$, e.g.

$$E [g(x) | y \in \mathcal{E}] = \int g(u)p(u | y \in \mathcal{E})du$$
Solution via product and sum rule

Assume that all variables are discrete valued, that $\mathcal{E} = \{y_o\}$, and that we know $p(x, y, z)$. We would like to know $p(x|y_o)$.

- **Product rule:** $p(x|y_o) = \frac{p(x,y_o)}{p(y_o)}$
- **Sum rule:** $p(x, y_o) = \sum_z p(x, y_o, z)$
- **Sum rule:** $p(y_o) = \sum_x p(x, y_o) = \sum_{x,z} p(x, y_o, z)$
- **Result:**

$$p(x|y_o) = \frac{\sum_z p(x, y_o, z)}{\sum_{x,z} p(x, y_o, z)}$$
What we do in PMR

\[ p(x|y_o) = \frac{\sum_z p(x,y_o,z)}{\sum_{x,z} p(x,y_o,z)} \]

Assume that \( x, y, z \) each are \( d = 500 \) dimensional, and that each element of the vectors can take \( K = 10 \) values.
What we do in PMR

\[
p(x|y_o) = \frac{\sum_z p(x,y_o,z)}{\sum_{x,z} p(x,y_o,z)}
\]

Assume that \(x, y, z\) each are \(d = 500\) dimensional, and that each element of the vectors can take \(K = 10\) values.

- **Issue 1:** To specify \(p(x, y, z)\), we need to specify \(K^{3d} - 1 = 10^{1500} - 1\) non-negative numbers, which is impossible.
\[ p(x|y_o) = \frac{\sum_z p(x,y_o,z)}{\sum_{x,z} p(x,y_o,z)} \]

Assume that \( x, y, z \) each are \( d = 500 \) dimensional, and that each element of the vectors can take \( K = 10 \) values.

**Issue 1:** To specify \( p(x, y, z) \), we need to specify \( K^{3d} - 1 = 10^{1500} - 1 \) non-negative numbers, which is impossible.

**Topic 1: Representation** What reasonably weak assumptions can we make to efficiently represent \( p(x, y, z) \)?
What we do in PMR

\[ p(x|y_o) = \frac{\sum_z p(x,y_o,z)}{\sum_{x,z} p(x,y_o,z)} \]

- **Issue 2:** The sum in the numerator goes over the order of \( K^d = 10^{500} \) non-negative numbers and the sum in the denominator over the order of \( K^{2d} = 10^{1000} \), which is impossible to compute.

- **Topic 2:** Exact inference
  Can we further exploit the assumptions on \( p(x,y_o,z) \) to efficiently compute the posterior probability or derived quantities?

- **Issue 3:** Where do the non-negative numbers \( p(x,y_o,z) \) come from?

- **Topic 3:** Learning
  How can we learn the numbers from data?

- **Issue 4:** For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

- **Topic 4:** Approximate inference and learning
What we do in PMR

\[ p(x|y_o) = \frac{\sum_z p(x, y_o, z)}{\sum_{x,z} p(x, y_o, z)} \]

- **Issue 2:** The sum in the numerator goes over the order of \( K^d = 10^{500} \) non-negative numbers and the sum in the denominator over the order of \( K^{2d} = 10^{1000} \), which is impossible to compute.

**Topic 2: Exact inference** Can we further exploit the assumptions on \( p(x, y, z) \) to efficiently compute the posterior probability or derived quantities?

- **Issue 4:** For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

**Topic 4: Approximate inference and learning**
What we do in PMR

\[ p(x|y_o) = \frac{\sum_z p(x, y_o, z)}{\sum_{x,z} p(x, y_o, z)} \]

- **Issue 2:** The sum in the numerator goes over the order of \( K^d = 10^{500} \) non-negative numbers and the sum in the denominator over the order of \( K^{2d} = 10^{1000} \), which is impossible to compute.

**Topic 2: Exact inference** Can we further exploit the assumptions on \( p(x, y, z) \) to efficiently compute the posterior probability or derived quantities?

- **Issue 3:** Where do the non-negative numbers \( p(x, y, z) \) come from?

**Topic 3: Learning** How can we learn the numbers from data?

- **Issue 4:** For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

**Topic 4: Approximate inference and learning**
What we do in PMR

\[ p(x|y_o) = \frac{\sum_z p(x,y_o,z)}{\sum_{x,z} p(x,y_o,z)} \]

- **Issue 2**: The sum in the numerator goes over the order of \( K^d = 10^{500} \) non-negative numbers and the sum in the denominator over the order of \( K^{2d} = 10^{1000} \), which is impossible to compute.

  **Topic 2: Exact inference** Can we further exploit the assumptions on \( p(x,y,z) \) to efficiently compute the posterior probability or derived quantities?

- **Issue 3**: Where do the non-negative numbers \( p(x,y,z) \) come from?

  **Topic 3: Learning** How can we learn the numbers from data?

- **Issue 4**: For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

  **Topic 4: Approximate inference and learning**