Probabilistic Inference and Learning
Lecture 03
Continuous Variables

Philipp Hennig

26 April 2021

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Logistic Regression	8
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	9
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	10
6	05.05.	Gaussian Distributions		19	29.06.	Example: Topic Models	
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	11
8	12.05.	Understanding Deep Learning		21	06.07.	EM	
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	12
10	19.05.	An Example for GP Regression		23	13.07.	Example: Topic Models	
11	25.05.	Understanding Kernels	6	24	14.07.	Example: Inferring Topics	13
12	26.05.	Gauss-Markov Models		25	20.07.	Example: Kernel Topic Models	
13	08.06.	GP Classification	7	26	21.07.	Revision	

We need to talk about real numbers.

But first, we need to talk about probabilities of derived quantities

Recap from Lecture 1

Plausibility as a Measure

[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

Definition (σ-algebra, measurable sets & spaces)

Let Ω be a space of *elementary events*. Consider the power set 2^Ω, and let $\mathfrak{F} \subset 2^\Omega$ be a set of subsets of Ω. Elements of $\mathfrak{F}$ are called *random events*. If $\mathfrak{F}$ satisfies the following properties, it is called a **σ-algebra**.

1. $\Omega \in \mathfrak{F}$ II.
2. $(A, B \in \mathfrak{F}) \Rightarrow (A - B \in \mathfrak{F})$ I.
3. $(A_1, A_2, \dots \in \mathfrak{F}) \Rightarrow \left(\bigcup_{i=1}^{\mathbb{N}} A_i \in \mathfrak{F} \quad \wedge \quad \bigcap_{i=1}^{\infty} A_i \in \mathfrak{F}\right)$ I.

(this implies $\varnothing \in \mathfrak{F}$. If $\mathfrak{F}$ is a σ-algebra, its elements are called **measurable sets**, and $(\Omega, \mathfrak{F})$ is called a **measurable space** (or **Borel space**).

If Ω is countable, then 2^Ω is a σ-algebra, and everything is easy.

[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Definition (Measure & Probability Measure)

Let $(\Omega, \mathfrak{F})$ be a **measurable space** (aka. Borel space). A nonnegative real function $P : \mathfrak{F} \to \mathbb{R}_{0,+}$ (III.) is called a **measure** if it satisfies the following properties:

1. $P(\varnothing) = 0$
2. For any countable sequence $\{A_i \in \mathfrak{F}\}_{i=1,\dots,}$ of pairwise disjoint sets ($A_i \cap A_j = \varnothing$ if $i \neq j$), P satisfies **countable additivity** (aka. σ-**additivity**):

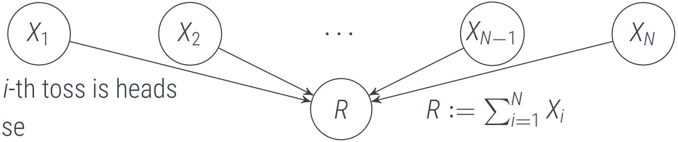
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \qquad \text{(V.)}$$

The measure P is called a **probability measure** if $P(\Omega) = 1$. IV.
(For probability measures, 1. is unnecessary). Then, $(\Omega, \mathfrak{F}, P)$ is called a **probability space**.

A hole in our theory?

What about derived quantities?

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

$$R := \sum_{i=1}^{N} X_i$$

- For $\boldsymbol{X} = [X_1, \dots, X_N]$, we have $\Omega = \{0, 1\}^N$.
- But what about $R \in [0, \dots, N] \subset \mathbb{N}$? It's not part of Ω.

Definition (Measurable Functions, Random Variables)

Let $(\Omega, \mathfrak{F})$ and $(\Gamma, \mathfrak{G})$ be two measurable spaces (i.e. spaces with σ-algebras). A function $X : \Omega \rightarrow \Gamma$ is called **measurable** if $X^{-1}(G) \in \mathfrak{F}$ for all $G \in \mathfrak{G}$. If there is, additionally, a probability measure P on $(\Omega, \mathfrak{F})$, then X is called a **random variable**.

Definition (Distribution Measure)

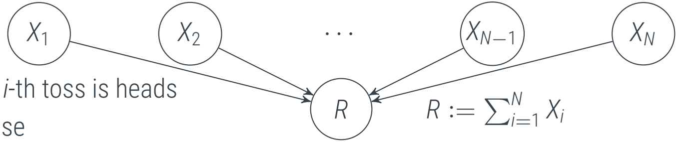
Let $X : \Omega \rightarrow \Gamma$ be a random variable. Then the **distribution measure** (or **law**) P_X of X is defined for any $G \subset \Gamma$ as

$$P_X(G) = P(X^{-1}(G)) = P(\{\omega \mid X(\omega) \in G\}).$$

Example: the Binomial Distribution

statistics of accumulated Bernoulli experiments

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

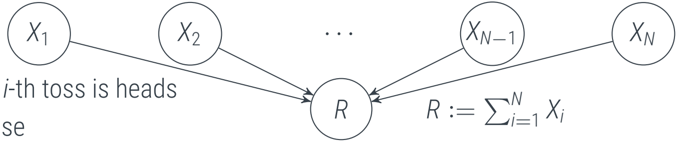
$$R := \sum_{i=1}^{N} X_i$$

$$P(R=r) = \sum_{\omega \in \{\boldsymbol{X} | R=r\}} \prod_{i=1}^{N} P(X_i) = \sum_{\omega \in \{\boldsymbol{X} | R=r\}} f^r \cdot (1-f)^{N-r} := P(r \mid f, N)$$

Example: the Binomial Distribution

statistics of accumulated Bernoulli experiments

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases} \qquad R := \sum_{i=1}^{N} X_i$$

$$P(R = r) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} \prod_{i=1}^{N} P(X_i) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} f^r \cdot (1-f)^{N-r} := P(r \mid f, N)$$

- original space: $\Omega = \{0; 1\}^N$ (countably finite)
- σ-algebra: 2^Ω (the power set)
- random variable $R = \sum_{i=1}^{N} X_i \in [0, \dots, N] =: \Gamma \subset \mathbb{N}$.
- distribution (measure) / law of R: ...

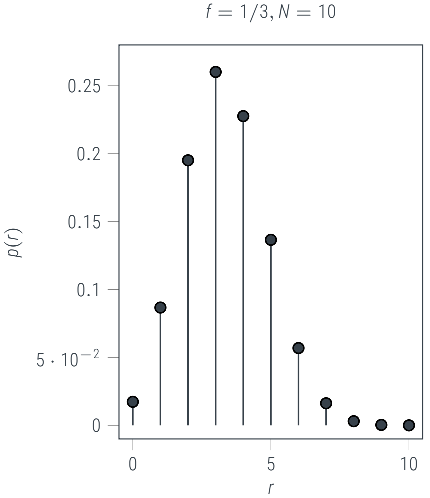
Example: the Binomial Distribution

statistics of accumulated Bernoulli experiments

The **distribution measure** of R is

$$P(r \mid f, N) = (\#\text{ ways to choose } r \text{ from } N) \cdot f^r \cdot (1-f)^{N-r}$$
$$= \frac{N!}{(N-r)! \cdot r!} \cdot f^r \cdot (1-f)^{N-r}$$
$$= \binom{N}{r} \cdot f^r \cdot (1-f)^{N-r}$$

Note: In the remainder of the course, will often **abuse notation** by writing $P(r)$ instead of $P(R = r)$ (recall again that $P(X) \neq P(Y)$!)



Now for the Real case ...

some complications for continous spaces

- in a countable space Ω, even 2^Ω is a σ-algebra.
- But in continous spaces, such as $\Omega = \mathbb{R}^d$, not all sets are measurable.
- However, $\mathbb{R}^d$ is a *topological space*

Definition (Topology)

Let Ω be a space and τ be a collection of sets. We say τ is a **topology** on Ω if

- $\Omega \in \tau$, and $\varnothing \in \tau$
- *any* union of elements of τ is in τ
- any intersection of *finitely many* elements of τ is in τ.

The elements of the topology τ are called **open sets**. In the Euclidean vector space $\mathbb{R}^d$, the canonical topology is that of all sets U that satisfy $x \in U :\Rightarrow \exists \varepsilon > 0 : ((\|y - x\| < \varepsilon) \Rightarrow (y \in U))$.

From topologies to σ-algebras

for topological spaces, it's easy to define σ-algebras

Note that a topology is *almost* a σ-algebra:

Definition (σ-algebra, measurable sets & spaces)

Let Ω be a space of *elementary events*. Consider the power set 2^Ω, and let $\mathfrak{F} \subset 2^\Omega$ be a set of subsets of Ω. Elements of $\mathfrak{F}$ are called *random events*. If $\mathfrak{F}$ satisfies the following properties, it is called a **σ-algebra**.

1. $\Omega \in \mathfrak{F}$ II.
2. $(A, B \in \mathfrak{F}) \Rightarrow (A - B \in \mathfrak{F})$ I.
3. $(A_1, A_2, \dots \in \mathfrak{F}) \Rightarrow$
   $(\bigcup_{i=1}^\infty A_i \in \mathfrak{F} \quad \wedge \quad \bigcap_{i=1}^\infty A_i \in \mathfrak{F})$ I.

(this implies $\varnothing \in \mathfrak{F}$. If $\mathfrak{F}$ is a σ-algebra, its elements are called **measurable sets**, and $(\Omega, \mathfrak{F})$ is called a **measurable space** (or **Borel space**).

Definition (Topology)

Let Ω be a space and τ be a collection of sets. We say τ is a **topology** on Ω if

- $\Omega \in \tau$, and $\varnothing \in \tau$
- *any* union of elements of τ is in τ
- any intersection of *finitely many* elements of τ is in τ.

The elements of the topology τ are called **open sets**. In the Euclidean vector space $\mathbb{R}^d$, the canonical topology is that of all sets U that satisfy $x \in U :\Rightarrow \exists \varepsilon > 0 : ((\|y - x\| < \varepsilon) \Rightarrow (y \in U))$.

Almost done!

Standard settings

Definition (Borel algebra)

Let (Ω, τ) be a topological space. The **Borel σ-algebra** is the σ-algebra *generated* by τ. That is by taking τ and completing it to include infinite intersections of elements from τ and all complements in Ω to elements of τ.

- In this lecture, we will almost exclusively consider (random) variables defined on discrete or Euclidean spaces. In the latter case, the σ-algebra will not be mentioned but assumed to be the Borel σ-algebra.
- Consider $(\Omega, \mathfrak{F})$ and $(\Gamma, \mathfrak{G})$. If both $\mathfrak{F}$ and $\mathfrak{G}$ are Borel σ-algebras, then any continuous function X is measurable (and can thus be used to define a random variable). This is because, for continuous functions, pre-images of open sets are open sets.

Now that we can define (Borel) σ-algebras on continous spaces, we can define probability distribution measures. They might just be a bit unwieldy.

- **Random Variables** allow us to define derived quantities from atomic events
- **Borel σ-algebras** can be defined on all topological spaces, allowing us to define probabilities if the elementary space is continuous.

Probability Densities

a convenient way to write things down

Definition (Probability Density Functions (pdf's))

Let $\mathfrak{B}$ be the Borel σ-algebra in $\mathbb{R}^d$. A probability measure P on $(\mathbb{R}^d, \mathfrak{B})$ has a **density** p if p is a non-negative (Borel) measurable function on $\mathbb{R}^d$ satisfying, for all $B \in \mathfrak{B}$

$$P(B) = \int_B p(x)\,dx =: \int_B p(x_1, \dots, x_d)\,dx_1 \,\dots\, dx_d$$

- In other words: P has a density if $P(B)$ can be written as an integral over B, for all B.
- not all measures have densities (e.g. measures with point masses)

Cumulative Distributions

Connecting probabilities to integration

Definition (Cumulative Distribution Function (CDF))

For probability measures P on $(\mathbb{R}^d, \mathfrak{B})$, the **cumulative distribution function** is the function

$$F(\boldsymbol{x}) = P\left(\prod_{i=1}^{d}(X_i < x_i)\right).$$

(In particular for the univariate case $d = 1$, we have $F(x) = P((-\infty, x]))$.
If F is sufficiently differentiable, then P has a density, given by

$$p(\boldsymbol{x}) = \left.\frac{\partial^d F}{\partial x_1 \cdots \partial x_d}\right|_{\boldsymbol{x}}.$$

and, for $d = 1$,

$$P(a \leq X < b) = F(b) - F(a) = \int_a^b f(x)\,dx.$$

Densities Satisfy the Laws of Probability Theory

because integrals are linear operators

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

without proof. Cf. Matthias Hein's lecture

- For probability densities p on $(\mathbb{R}^d, \mathfrak{B})$ we have

$$P(E) \overset{(IV)}{=} 1 = \int_{\mathbb{R}^d} p(x)\, dx.$$

- Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a random variable with density p_X on $\mathbb{R}^2$. Then the **marginal densities** of X_1 and X_2 are given by the **sum rule**

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_2, \qquad p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_1$$

- The **conditional density** $p(x_1 \mid x_2)$ (for $p(x_2) > 0$) is given by the **product rule**

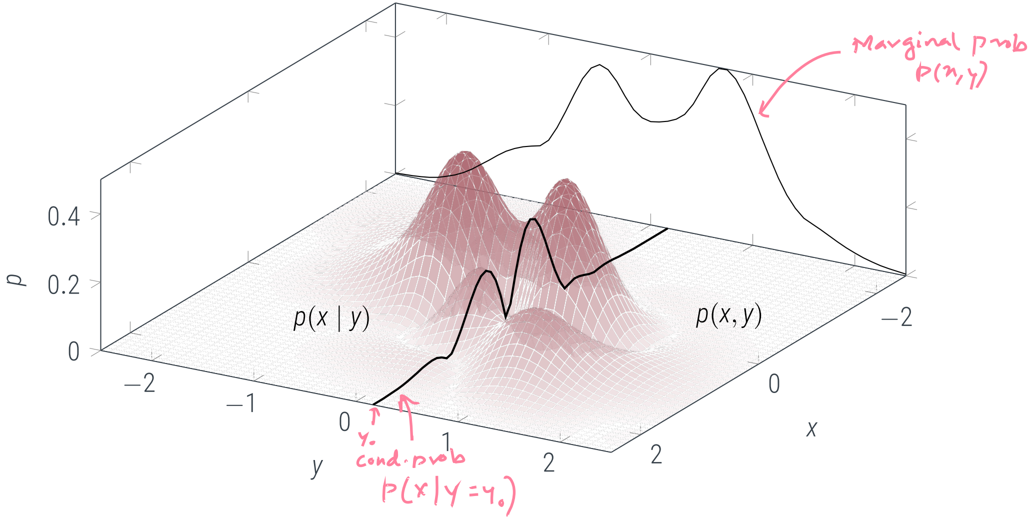
$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

- **Bayes' Theorem** holds:

$$p(x_1 \mid x_2) = \frac{p(x_1) \cdot p(x_2 \mid x_1)}{\int p(x_1) \cdot p(x_2 \mid x_1)\, dx_1}.$$

A Graphical View

sketch



Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $p_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$p_Y(y) = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right| = p_X(v(y)) \cdot \left|\frac{du(x)}{dx}\right|^{-1}.$$

Proof: for $u'(X) > 0$: $\forall\, d_1 = u(c_1) < y < u(c_2) = d_2$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \leq v(y)) = \int_{c_1}^{v(y)} p(x)\,dx$$

$$p_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = p_X(v(y)) \cdot \frac{dv(y)}{dy} = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right|$$

(2, 3, 4)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $p_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$p_Y(y) = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right| = p_X(v(y)) \cdot \left|\frac{du(x)}{dx}\right|^{-1}.$$

Proof: for $u'(X) < 0$: $\forall\ d_2 = u(c_2) < y < u(c_1) = d_1$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \geq v(y)) = 1 - P(X \leq v(y)) = 1 - \int_{c_1}^{v(y)} p(x)\, dx$$

$$p_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = -p_X(v(y)) \cdot \frac{dv(y)}{dy} = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right|$$

Theorem (Transformation Law, general)

Let $X = (X_1, \dots, X_d)$ have a joint density p_X. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be continously differentiable and injective, with non-vanishing Jacobian J_g. Then $Y = g(X)$ has density

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| & \text{if } y \text{ is in the range of } g, \\ 0 & \text{otherwise.} \end{cases}$$

The Jacobian J_g is the $d \times d$ matrix with

$$[J_g(x)]_{ij} = \frac{\partial g_i(x)}{\partial x_j}.$$

- **Probability Density Functions (pdf's)** distribute probability across continuous domains.
- they satisfy "the rules of probability":

$$\int_{\mathbb{R}^d} p(x)\, dx = 1$$

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_2 \qquad \textbf{sum rule}$$

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \qquad \textbf{product rule}$$

$$p(x_1 \mid x_2) = \frac{p(x_1) \cdot p(x_2 \mid x_1)}{\int p(x_1) \cdot p(x_2 \mid x_1)\, dx_1} \qquad \textbf{Bayes' Theorem.}$$

- Not every measure has a density, but all pdfs define measures
- Densities transform under continuously differentiable, injective functions $g : x \mapsto y$ with non-vanishing Jacobian as

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| & \text{if } y \text{ is in the range of } g, \\ 0 & \text{otherwise.} \end{cases}$$

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses

An example

Based on a very famous argument

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere

If we sample independently, what is the likelihood for a positive or a negative observation?

$$p(X = 1 \mid \pi) = \pi; \qquad p(X = 0 \mid \pi) = 1 - \pi$$

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0,1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0,1]$, i.e. $p(\pi) = 1$, zero elsewhere

If we sample independently, what is the likelihood for a positive or a negative observation?

$$p(X = 1 \mid \pi) = \pi; \qquad p(X = 0 \mid \pi) = 1 - \pi$$

What is the posterior after n positive, m negative observations?

$$p(\pi \mid n, m) = \frac{\pi^n (1-\pi)^m \cdot 1}{\int \pi^n (1-\pi)^m \cdot 1\, d\pi} = \frac{\pi^n (1-\pi)^m}{B(n+1, m+1)}$$

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DEMO

La probabilité de la plupart des événemens simples, est inconnue; en la considérant à priori, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais sie l'on a observé un résultat composé de plusieurs de ces événemens, la manière dont ils y entrent, rend quelques-unes de ces valeurs plus probables que les autres. Ainsi à mesure que les résultat observé se compose par le développement des événemens simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui se reserrant sans cesse, finiraient par coïncider, si le nombre des événemens simples devenait infini.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 363
Translated by a Deep Network, assisted by a human

The probability of most simple events is unknown. Considering it a priori, it seems susceptible to all values between zero and unity. But if one has observed a result composed of several of these events, the way they enter them makes some of these values more probable than the others. Thus, as the observed results are composed by the development of simple events, their real possibility becomes more and more known, and it becomes more and more probable that it falls within limits that constantly tighten, would end up coinciding if the number of simple events became infinite.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 363
Translated by a Deep Network, assisted by a human

Let’s be more careful with notation!
(but only once more, then we’ll be sloppy)

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

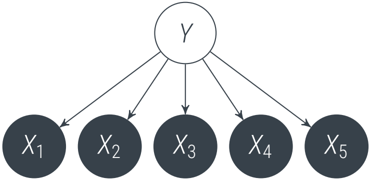
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

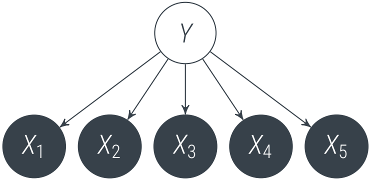
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Generative model and joint probability

- we abbreviate $Y = \pi$ as π, $X_i = x_i$ as x_i
- $p(\pi)$ is the prior of Y, written fully $p(Y = \pi)$
- $p(x_i|\pi)$ is the likelihood of observation x_i
- note that the likelihood is a function of π

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi \mid \text{“nothing”}) = p(\pi)$$

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = \frac{p(x_1|\pi)p(\pi)}{\int p(x_1|\pi)p(\pi)\,d\pi} = Z_1^{-1}p(x_1|\pi)p(\pi)$$

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = \frac{p(x_1|\pi)p(\pi)}{\int p(x_1|\pi)p(\pi)\,d\pi} = Z_1^{-1}p(x_1|\pi)p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi|x_1,x_2) = Z_2^{-1}p(x_2|x_1,\pi)p(x_1|\pi)p(\pi) = Z_2^{-1}p(x_2|\pi)p(x_1|\pi)p(\pi)$$

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = \frac{p(x_1|\pi)p(\pi)}{\int p(x_1|\pi)p(\pi)\,d\pi} = Z_1^{-1}p(x_1|\pi)p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi|x_1,x_2) = Z_2^{-1}p(x_2|x_1,\pi)p(x_1|\pi)p(\pi) = Z_2^{-1}p(x_2|\pi)p(x_1|\pi)p(\pi)$$

...

Probability of wearing glasses after five observations

$$p(\pi|x_1,x_2,x_3,x_4,x_5) = Z_5^{-1}\left(\prod_{i=1}^{5}p(x_i|\pi)\right)p(\pi)$$

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1 | \pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1 - \pi & \text{for } x_1 = 0 \end{cases}$$

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1-\pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV N for the number of observations being 1 (with values n)
- RV M for the number of observations being 0 (with values m)

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1-\pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV N for the number of observations being 1 (with values n)
- RV M for the number of observations being 0 (with values m)

Probability of wearing glasses after five observations

$$\begin{aligned} p(\pi|x_1, x_2, x_3, x_4, x_5) &= Z_5^{-1}\left(\prod_{i=1}^{5} p(x_i|\pi)\right) p(\pi) \\ &= Z_5^{-1}\pi^n(1-\pi)^m p(\pi) \\ &= p(\pi|n,m) \end{aligned}$$

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1}$$

with parameters $a > 0, b > 0$

*the Beta **distribution** with parameter a and b*

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1}$$

with parameters $a > 0, b > 0$

*the Beta **distribution** with parameter a and b*

Let's give the normalization factor Z of the beta distribution a name!

$$B(a,b) = \int_0^1 \pi^{a-1}(1-\pi)^{b-1}d\pi$$

*the Beta **function** with parameters a and b*

Quand les valeurs de x, considérées indépendamment du résultat observé, ne sont pas également possibles; en nommant z la fonction de x qui exprime leur probabilité; il est facile de voir, par ce qui a été dit dans le premier chaptire de ce Livre, qu'en changeant dans la formule (1), y dans $y \cdot z$, on aura la probabilité que la valeur de x est comprise dans les limites $x = \theta$ and $x = \theta'$. Cela revient à supposer toutes les valeurs de x également possible à priori, et à considérer le résultat observé, comme étant formé de deux résultats indépendans, dont les probabilités sont y et z. On peut donc ramener ainsi tous les case à celui ou l'on suppose à priori, avant l'événement, une égal possibilité aux différentes valeurs de x, et par cette raison, nous adopterons cette hypothèse dans ce qui va suivre.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 364
Translated by a Deep Network, assisted by a human

When the values of x, considered independently of the observed result, are not equally possible; if we name z the function of x which expresses their probability; it is easy to see, by what has been said in the first chapter of this Book, that by changing in formula (1), y in $y \cdot z$, we will have the probability that the value of x is within the limits $x = \theta$ and $x = \theta'$. This amounts to assuming all the values of x equally possible a priori, and to considering the observed result as being formed by two independent results, whose probabilities are y and z. We can thus reduce all the cases to the one where we assume a priori, before the event, an equal possibility to the different values of x, and by this reason, we will adopt this hypothesis in what follows.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 364
Translated by a Deep Network, assisted by a human

- **Random Variables** allow us to define derived quantities from atomic events
- **Borel σ-algebras** can be defined on all topological spaces, allowing us to define probabilities if the elementary space is continuous.
- **Probability Density Functions (pdf's)** distribute probability across continuous domains.
  - they satisfy "the rules of probability" (integrate to one, sum rule, product rule, hence Bayes' Theorem)
  - Not every measure has a density, but all pdfs define measures
  - Densities transform under continuously transformations
- Probabilistic inference can even be used to infer probabilities!