Probability background and basic results

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

Program

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation

Program

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation

Key rules of probability

- We have introduced the following two key rules for a pdf/pmf $p(\mathbf{x}, \mathbf{y})$

  (1) Product rule:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &= p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \end{aligned}$$

  (2) Sum rule (marginalisation):

$$p(\mathbf{y}) = \begin{cases} \sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) & \text{for discrete r.v.} \\ \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x} & \text{for continuous r.v.} \end{cases}$$

- We will keep using them in the course.

Example: joint pmf

- Two binary random variables:
  $w \in \{0, 1\}$ with 0≡sunny, 1≡rainy
  $u \in \{0, 1\}$, with 0≡no umbrella, 1≡umbrella
- Their joint pmf $p(w, u)$ can be represented as a 2D table:

	$u = 0$ (no)	$u = 1$ (yes)
$w = 0$ (sunny)	0.24	0.06
$w = 1$ (rainy)	0.07	0.63

- For example, the probability of the joint event that it is both rainy and that you carry an umbrella is 0.63.
- All entries are ≥ 0 and sum to 1 (normalisation condition).

Example: sum rule / marginalisation

- We use the sum rule / marginalisation to "reduce" the probabilities of joint events to probabilities of single events, e.g. to compute the probability whether it is raining or not.
- To marginalise out u, fix a w and sum "horizontally"; to marginalise out w, fix a u and sum "vertically".

	$u = 0$	$u = 1$	$p(w) = \sum_u p(w, u)$
$w = 0$	0.24	0.06	0.30
$w = 1$	0.07	0.63	0.70
$p(u) = \sum_w p(w, u)$	0.31	0.69	

- The notation $\sum_w$ means "sum over all values of w".
- In case of pdfs, the sum is replaced by an integral.

Example: product rule and conditioning

- Product rule: $p(w, u) = p(u|w)p(w)$, hence

$$p(u|w) = \frac{p(w, u)}{p(w)}$$

- Given it is rainy ($w = 1$), compute $p(u|w = 1)$:

$$p(u = 1|w = 1) = \frac{p(w = 1,\ u = 1)}{p(w = 1)} = \frac{0.63}{0.70} = 0.9$$

$$p(u = 0|w = 1) = \frac{p(w = 1,\ u = 0)}{p(w = 1)} = \frac{0.07}{0.70} = 0.1$$

- Note: $p(u = 1|w = 1) = 0.9$ while $p(u = 1) = 0.69$. Knowing that it rains changes the probability to see umbrellas. This means that the weather (w) and umbrellas (u) are statistically dependent (see later in the course).

Example: product rule and conditioning

- To condition on $w = 1$, keep only the row with $w = 1$ in the table for $p(w, u)$ and scale the terms to sum to 1.

	$u = 0$	$u = 1$
$w = 1$	0.07	0.63

$$\Downarrow$$

$u = 0$	$u = 1$
$\frac{0.07}{0.07+0.63}$	$\frac{0.63}{0.07+0.63}$

- Conditioning means restricting/filtering the outcome space to only retain those events that are compatible with the conditioning event, followed by renormalisation.

Notation

- When writing
$$p(w) = \sum_u p(w, u) \tag{1}$$
we overloaded notation and used the same symbol to denote the random variable and the variable we sum over.
- We also used the same symbol p to denote the pmf of the marginal $p(w)$ and the joint $p(w, u)$.
- The arguments and context are used to uniquely identify which quantity we refer to.
- This is often done, also for pdfs, but can lead to confusion.

Notation

- Cleaner but more complex notation uses subscripts, e.g. $p_W(w)$ and $p_{W,U}(w, u)$.
- Frees the symbols w, u from being needed as arguments, e.g. we can write $p_{W,U}$ or $p_{W,U}(\alpha, \beta)$ where α and β are arbitrary variable names.
- We can also have equations such as

$$p_W(\alpha) = \sum_{\beta} p_{W,U}(\alpha, \beta) \tag{2}$$

- We typically use the overloaded notation unless there is reason not to.

Program

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation

Representing pdfs/pmfs in terms of samples

- pdfs/pmfs $p(\mathbf{x}, \mathbf{y})$ may be represented by samples $(\mathbf{x}_i, \mathbf{y}_i) \sim p(\mathbf{x}, \mathbf{y}), i = 1, \ldots n$.
- Can be the main purpose (e.g. generated images or text) or used to approximate intractable computations.
- For example, $p(w, u)$ can be represented in terms of the data

i	w_i	u_i
1	0	0
2	0	1
3	1	1
$\vdots$	$\vdots$	$\vdots$
n	0	0

The frequency of occurrence of the four possible patterns (approximately) matches $p(w, u)$.

The empirical pdf/pmf

- We said that the samples approximately represented their pdf/pmf. We here formalise that relationship.
- We focus on discrete random variables $\mathbf{x} \in \{\mathbf{a}_1, \mathbf{a}_2, \ldots\}$ for simplicity. The argument extends to continuous random variables but is more technical.
- We will use indicator functions

$$\mathbb{1}(\mathbf{x} = \mathbf{a}) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{a} \\ 0 & \text{if } \mathbf{x} \neq \mathbf{a} \end{cases} \tag{3}$$

- Given n samples $\mathbf{x}_i$, each with weight $1/n$, we can represent them using the so-called empirical pmf

$$\hat{p}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\mathbf{x} = \mathbf{x}_i) \tag{4}$$

Equivalent to the table representation.

The empirical pdf/pmf

- We can group those samples with the same value together.
- Let n_k be the number of times that $\mathbf{x} = \mathbf{a}_k$ in the data set. Then

$$\hat{p}_n(\mathbf{x}) = \sum_k \frac{n_k}{n} \mathbb{1}(\mathbf{x} = \mathbf{a}_k) \tag{5}$$

- Tells us the fraction of times each possible value of $\mathbf{x}$ occurs among the samples. Corresponds to a histogram.
- As n increases $\hat{p}_n(\mathbf{x})$ becomes closer and closer to $p(\mathbf{x})$ since

$$\frac{n_k}{n} \to p(\mathbf{x} = \mathbf{a}_k) \tag{6}$$

The empirical pdf/pmf

- We constructed a pmf from samples, so we "learned" from data.
- We say $\hat{p}_n(\mathbf{x})$ is an estimate of $p(\mathbf{x})$.
- $\hat{p}_n(\mathbf{x})$ exactly represents the samples; there is no generalisation.
- It's rather "memorisation" than learning.
- Later in the course, we will discuss learning methods and approaches that do aim to generalise.

Marginalisation for samples

- Marginalising out variables means we are not interested in them, i.e. we ignore their values in the samples.
- This corresponds to dropping the variables that we marginalise over from the samples.
- Example:

$p(w,u)$		
i	w_i	u_i
1	0	0
2	0	1
3	1	1
$\vdots$	$\vdots$	$\vdots$
n	0	0

$\longrightarrow$

$p(w)$	
i	w_i
1	0
2	0
3	1
	$\vdots$
n	0

The frequency of 0's and 1's (approximately) matches $p(w)$.

Marginalisation for samples: the math

- Represent the samples $(w_i, u_i) \sim p(w, u)$ as

$$\hat{p}_n(w, u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}((w, u) = (w_i, u_i)) \tag{7}$$

- Marginalise out u

$$\begin{aligned}
\hat{p}_n(w) &= \sum_{u=0}^{1} \hat{p}_n(w, u) = \sum_{u=0}^{1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}((w, u) = (w_i, u_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}((w, 0) = (w_i, u_i)) + \mathbb{1}((w, 1) = (w_i, u_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(w = w_i)
\end{aligned}$$

  since each u_i is either 0 or 1.

- Equals the empirical pmf after dropping the u_i from the samples.
- Argument generalises to other pmfs and pdfs.

Conditioning for samples

- Conditioning means filtering the outcome space to only retain those events that are compatible with the conditioning event, followed by renormalisation.
- For samples, this means you only keep those samples that match the conditioning event.
- Example:

$p(w, u)$

i	w_i	u_i
1	0	0
2	0	1
3	1	1
$\vdots$	$\vdots$	$\vdots$
$n-1$	1	1
n	0	0

$\longrightarrow$

$p(u|w=1)$

i	u_i
3	1
$\vdots$	$\vdots$
$n-1$	1

- Shrinks the sample size.

Conditioning for samples: the math

- Represent the samples $(w_i, u_i) \sim p(w, u)$ as

$$\hat{p}_n(w, u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}((w, u) = (w_i, u_i)) \tag{8}$$

- Condition on $w = 1$:

$$\hat{p}_n(u|w = 1) = \frac{\hat{p}_n(w = 1, u)}{\hat{p}_n(w = 1)} \tag{9}$$

$$\propto \hat{p}_n(w = 1, u) \tag{10}$$

$$\propto \sum_{i=1}^{n} \mathbb{1}((w = 1, u) = (w_i, u_i)) \tag{11}$$

$$\propto \sum_{i:w_i=1} \mathbb{1}(u = u_i) \tag{12}$$

- This means we only retain those samples that match the $w = 1$ condition.

Program

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation

Product rule for conditionals

- The product rule

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y}) \tag{13}$$

  also applies to conditionals.
- We have

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{x}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{x}|\mathbf{y}, \mathbf{z}) \tag{14}$$

- And hence also

$$p(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{z})}{p(\mathbf{x}|\mathbf{z})} \tag{15}$$

- It's like for the "normal" product rule, but we have everywhere an additional $|\mathbf{z}$.

From product rule to chain rule

Iteratively applying the product rule allows us to factorise any joint pdf/pmf $p(\mathbf{x}) = p(x_1, x_2, \ldots, x_d)$ into product of conditional pdfs/pmfs.

$$
\begin{aligned}
p(\mathbf{x}) &= p(x_1)p(x_2, \ldots, x_d|x_1) \\
&= p(x_1)p(x_2|x_1)p(x_3, \ldots, x_d|x_1, x_2) \\
&= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4, \ldots, x_d|x_1, x_2, x_3) \\
&\vdots \\
&= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_d|x_1, \ldots x_{d-1}) \\
&= p(x_1) \prod_{i=2}^{d} p(x_i|x_1, \ldots, x_{i-1}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i)
\end{aligned}
$$

with $\mathrm{pre}_i = \mathrm{pre}(x_i) = \{x_1, \ldots, x_{i-1}\}$, $\mathrm{pre}_1 = \varnothing$ and $p(x_1|\varnothing) = p(x_1)$.

The chain rule can be applied to any ordering of the variables. For each x_i, we condition on all previous variables in the ordering.

Program

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation

Expectation

- Let $\mathbf{x}$ be a random variable with pdf/pmf $p(\mathbf{x})$. For some function $g(\mathbf{x})$, the expectation $\mathbb{E}[g(\mathbf{x})]$ is

$$\mathbb{E}[g(\mathbf{x})] = \begin{cases} \int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} & \text{for pdfs} \\ \sum_{\mathbf{x}} g(\mathbf{x})p(\mathbf{x}) & \text{for pmfs} \end{cases} \tag{16}$$

  (this is not a definition, but a basic result known as the "law of the unconscious statistician (lotus)")

- The expectation is linear, i.e.

$$\mathbb{E}[\alpha_1 g_1(\mathbf{x}) + \alpha_2 g_2(\mathbf{x})] = \alpha_1 \mathbb{E}[g_1(\mathbf{x})] + \alpha_2 \mathbb{E}[g_2(\mathbf{x})] \tag{17}$$

  due to the sum/integral being linear operations.

Expectation

- Assume $\mathbf{x}$ is discrete and two-dimensional, i.e. $\mathbf{x} = (x_1, x_2)$ but $g(\mathbf{x}) = x_1$ and hence only involves x_1.
- Then we only need $p(x_1)$ to compute the expectation:

$$\mathbb{E}[g(\mathbf{x})] = \sum_{x_1, x_2} x_1 p(x_1, x_2) \tag{18}$$

$$= \sum_{x_1, x_2} x_1 p(x_1) p(x_2|x_1) \tag{19}$$

$$= \sum_{x_1} \sum_{x_2} x_1 p(x_1) p(x_2|x_1) \tag{20}$$

$$= \sum_{x_1} x_1 p(x_1) \underbrace{\sum_{x_2} p(x_2|x_1)}_{1} \tag{21}$$

$$= \sum_{x_1} x_1 p(x_1) \tag{22}$$

- Generalises: you only need the distribution of those variables that actually occur in the expectation.

Notation

- Writing $\mathbb{E}\left[g(\mathbf{x})\right]$ assumes that context uniquely defines the $p(\mathbf{x})$ used in the expectation.
- To avoid ambiguity, subscripts can be used to indicate with respect to which distribution we take the expectation:

$$\mathbb{E}_{p(\mathbf{x})}\left[g(\mathbf{x})\right]$$

- The result from previous slide can thus be expressed as

$$\mathbb{E}_{p(x_1,x_2)}[x_1] = \mathbb{E}_{p(x_1)}[x_1] \tag{23}$$

- Unless needed for clarity, we use the simpler notation.
- Example where needed: we approximate $p(\mathbf{x})$ with $q(\mathbf{x})$ so that

$$\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \mathbb{E}_{q(\mathbf{x})}[g(\mathbf{x})] \tag{24}$$

Notation

- Other notation that is sometimes used is

$$\mathbb{E}_{\mathbf{x}}\left[g(\mathbf{x})\right]$$

- Indicates which random variable we take the expectation over. Assumes context makes it clear what pdf/pmf $\mathbf{x}$ has, or that their definition does not matter.
- In that notation, we have

$$\mathbb{E}_{x_1,x_2}[x_1] = \mathbb{E}_{x_1}[x_1] \tag{25}$$

Conditional expectation

- There a several notational conventions for conditional expectations. Let $g(\mathbf{x}, \mathbf{y})$ be a function of two random vectors $\mathbf{x}$ and $\mathbf{y}$.

$$\mathbb{E}[g(\mathbf{x}, \mathbf{y})|\mathbf{y}] \qquad \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[g(\mathbf{x}, \mathbf{y})] \qquad \mathbb{E}_{\mathbf{x}|\mathbf{y}}[g(\mathbf{x}, \mathbf{y})] \tag{26}$$

- They all denote

$$\begin{cases} \int g(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x} & \text{for pdfs} \\ \sum_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\mathbf{y}) & \text{for pmfs} \end{cases} \tag{27}$$

Product rule and expectations

- The product rule can be used to decompose expectations:

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[g(\mathbf{x},\mathbf{y})] = \sum_{\mathbf{x},\mathbf{y}} p(\mathbf{x},\mathbf{y}) g(\mathbf{x},\mathbf{y}) \tag{28}$$

$$= \sum_{\mathbf{x},\mathbf{y}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) g(\mathbf{x},\mathbf{y}) \tag{29}$$

$$= \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) g(\mathbf{x},\mathbf{y}) \tag{30}$$

$$= \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) g(\mathbf{x},\mathbf{y}) \tag{31}$$

$$= \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{y}|\mathbf{x})}[g(\mathbf{x},\mathbf{y})] \tag{32}$$

- This means we first take the expectation with respect to (wrt) $p(\mathbf{y}|\mathbf{x})$ then wrt $p(\mathbf{x})$.
- We can also swap the order of $\mathbf{x}$ and $\mathbf{y}$, so that

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})} = \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{y}|\mathbf{x})} = \mathbb{E}_{p(\mathbf{y})} \mathbb{E}_{p(\mathbf{x}|\mathbf{y})} \tag{33}$$

Chain rule and expectations

- Iteratively applying the product rule led to the chain rule. With $p(\mathbf{x}) = p(x_1, x_2, \ldots, x_d)$, we had

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i) \tag{34}$$

$$= p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)\ldots p(x_d|x_1, \ldots, x_{d-1}) \tag{35}$$

- As for the product rule, this decomposition leads to a decomposition of expectations:

$$\mathbb{E}_{p(\mathbf{x})} = \mathbb{E}_{p(x_1)}\mathbb{E}_{p(x_2|x_1)} \cdots \mathbb{E}_{p(x_d|x_1,\ldots,x_{d-1})} \tag{36}$$

- Note: this works with any ordering of the variables.

Chain rule and expectations

- To prove the statement, let $g(\mathbf{x})$ be some function of the d-dimensional $\mathbf{x}$.
- Assume $p(\mathbf{x})$ is a pmf for simplicity (in case of pdfs, use integrals instead). Then:

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] &= \sum_{\mathbf{x}} p(\mathbf{x})g(\mathbf{x}) \\
&= \sum_{\mathbf{x}} p(x_1)p(x_2|x_1)\dots p(x_d|x_1,\dots,x_{d-1})g(\mathbf{x}) \\
&= \sum_{x_1}\sum_{x_2}\dots\sum_{x_d} p(x_1)p(x_2|x_1)\dots p(x_d|x_1,\dots,x_{d-1})g(\mathbf{x}) \\
&= \sum_{x_1} p(x_1)\sum_{x_2} p(x_2|x_1)\dots\sum_{x_d} p(x_d|x_1,\dots,x_{d-1})g(\mathbf{x}) \\
&= \mathbb{E}_{p(x_1)}\mathbb{E}_{p(x_2|x_1)}\dots\mathbb{E}_{p(x_d|x_1,\dots,x_{d-1})}[g(\mathbf{x})]
\end{aligned}$$

- We first take the expectation wrt $p(x_d|x_1,\dots,x_{d-1})$, then $p(x_{d-1}|x_1,\dots,x_{d-2})$, going backwards until we take the expectation wrt $p(x_1)$.

Program recap

1. Sum and product rule
2. Samples
3. Chain rule
4. Expectation