Most quantities in machine learning are not fully deterministic.

- true randomness of events
  - a photon reaches a camera's CCD chip, is it detected or not?
    it depends on quantum effects, which -to our knowledge- are stochastic
- incomplete knowledge
  - what will be the next picture I take with my smartphone?
  - who will be in my floorball group this afternoon?
- insufficient representation
  - what material corresponds to that green pixel in the image?
    with RGB impossible to tell, with hyperspectral maybe possible

In practice, there is no difference between these!

Probability theory allows us to deal with this.

Random events, random variables

A **random variable** is a variable that randomly takes one of its possible values:

- the number of photons reaching a CCD chip
- the next picture I will take with my smartphone
- the names of all people in my floorball group

Some notation: we will write

- random variables with capital letters, e.g. X
- the set of possible values it can take with curly letters, e.g. $\mathcal{X}$
  for simplicity we only look at discrete $\mathcal{X}$ (otherwise, notation changes a little)
- any individual value it can take with lowercase letters, e.g. x

How likely each value $x \in \mathcal{X}$ is specified by a *probability distribution*:

- $p(X = x)$ is the probability that X takes the value $x \in \mathcal{X}$.
  If it's clear which variable we mean, we'll just write $p(x)$.
- for example, rolling a die, $p(X = 3) = p(3) = 1/6$
- we write $x \sim p(x)$ to indicate that the distribution of X is $p(x)$

Properties of probabilities

Some rules are needed for probabilities to make sense:

$$0 \leq p(x) \leq 1 \quad \text{for all } x \in \mathcal{X} \qquad \text{(positivity)}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1 \qquad \text{(normalization)}$$

If X has only two possible values, e.g. $\mathcal{X} = \{\texttt{true}, \texttt{false}\}$,

$$p(X = \texttt{false}) = 1 - p(X = \texttt{true})$$

Example: *PASCAL VOC2006* dataset

Define random variables

- X_{obj}: does a randomly picked image contain an object *"obj"*?
- $\mathcal{X}_{obj} = \{\texttt{true}, \texttt{false}\}$

$$p(X_{person} = \texttt{true}) = 0.254 \qquad p(X_{person} = \texttt{false}) = 0.746$$

$$p(X_{horse} = \texttt{true}) = 0.094 \qquad p(X_{horse} = \texttt{false}) = 0.916$$

Joint probabilities

Probabilities can be assigned to more than one random variable at a time:

- $p(X = x, Y = y)$ is the probability that $X = x$ and $Y = y$ (at the same time)

joint probability

Example: *PASCAL VOC2006* dataset

- $p(X_{person} = \texttt{true}, X_{horse} = \texttt{true}) = 0.050$
- $p(X_{dog} = \texttt{true}, X_{person} = \texttt{true}, X_{cat} = \texttt{false}) = 0.014$
- $p(X_{aeroplane} = \texttt{true}, X_{aeroplane} = \texttt{false}) = 0$

Marginalization

We can recover the probabilities of individual variables from the joint probability by summing over all variables we are not interested in.

- $p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)$
- $p(X_2 = z) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_3 \in \mathcal{X}_3} \sum_{x_4 \in \mathcal{X}_4} p(X_1 = x_1, X_2 = z, X_3 = x_3, X_4 = x_4)$

marginalization

Marginalization

We can recover the probabilities of individual variables from the joint probability by summing over all variables we are not interested in.

- $p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)$
- $p(X_2 = z) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_3 \in \mathcal{X}_3} \sum_{x_4 \in \mathcal{X}_4} p(X_1 = x_1, X_2 = z, X_3 = x_3, X_4 = x_4)$

marginalization

Example: *PASCAL VOC2006* dataset

- $p(X_{person} = \texttt{true}, X_{horse} = \texttt{true}) = 0.050$
- $p(X_{person} = \texttt{true}, X_{horse} = \texttt{false}) = 0.204$
- $p(X_{person} = \texttt{false}, X_{horse} = \texttt{true}) = 0.044$
- $p(X_{person} = \texttt{false}, X_{horse} = \texttt{false}) = 0.702$

	horse	no horse	
person	0.050	0.204	
no person	0.044	0.702	

Marginalization

We can recover the probabilities of individual variables from the joint probability by summing over all variables we are not interested in.

- $p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)$
- $p(X_2 = z) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_3 \in \mathcal{X}_3} \sum_{x_4 \in \mathcal{X}_4} p(X_1 = x_1, X_2 = z, X_3 = x_3, X_4 = x_4)$

marginalization

Example: *PASCAL VOC2006* dataset

- $p(X_{person} = \texttt{true}, X_{horse} = \texttt{true}) = 0.050$
- $p(X_{person} = \texttt{true}, X_{horse} = \texttt{false}) = 0.204$
- $p(X_{person} = \texttt{false}, X_{horse} = \texttt{true}) = 0.044$
- $p(X_{person} = \texttt{false}, X_{horse} = \texttt{false}) = 0.702$

	horse	no horse	$\sum$
person	0.050	0.204	0.254
no person	0.044	0.702	0.746
$\sum$	0.094	0.906	

- $p(X_{person} = \texttt{true}) = 0.050 + 0.204 = 0.254$
- $p(X_{horse} = \texttt{false}) = 0.204 + 0.702 = 0.906$

Conditional probabilities

One random variable can contain information about another one:

- $p(X = x \,|\, Y = y)$: **conditional probability**

  what is the probability of $X = x$, if we already know that $Y = y$?

- $p(X = x)$: **marginal probability**

  what is the probability of $X = x$, without any additional information?

- conditional probabilities can be computed from joint and marginal:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad \text{(not defined if } p(Y = y) = 0\text{)}$$

Conditional probabilities

One random variable can contain information about another one:

- $p(X = x \mid Y = y)$: **conditional probability**

  what is the probability of $X = x$, if we already know that $Y = y$?
- $p(X = x)$: **marginal probability**

  what is the probability of $X = x$, without any additional information?
- conditional probabilities can be computed from joint and marginal:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad \text{(not defined if } p(Y = y) = 0\text{)}$$

Example: *PASCAL VOC2006* dataset

- $p(X_{\textit{person}} = \texttt{true}) = 0.254$
- $p(X_{\textit{person}} = \texttt{true}|X_{\textit{horse}} = \texttt{true}) = \frac{0.050}{0.094} = 0.534$
- $p(X_{\textit{dog}} = \texttt{true}) = 0.139$
- $p(X_{\textit{dog}} = \texttt{true}|X_{\textit{cat}} = \texttt{true}) = \frac{0.002}{0.147} = 0.016$
- $p(X_{\textit{dog}} = \texttt{true}|X_{\textit{cat}} = \texttt{false}) = \frac{0.137}{0.853} = 0.161$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$
- $p(X_1 = 1) = 0.1 \quad p(X_1 = 2) = 0.2 \quad p(X_1 = 3) = 0.3$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$
- $p(X_1 = 1) = 0.1 \quad p(X_1 = 2) = 0.2 \quad p(X_1 = 3) = 0.3$
- $p(X_1 = 1, X_2 = 0) = 0.4 \quad p(X_1 = 1, X_2 = 1) = 0.6,$
  $p(X_1 = 2, X_2 = 0) = 0.2 \quad p(X_1 = 2, X_2 = 1) = 0.8,$
  $p(X_1 = 3, X_2 = 0) = 0.5 \quad p(X_1 = 3, X_2 = 1) = 0.5$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$
- $p(X_1 = 1) = 0.1 \quad p(X_1 = 2) = 0.2 \quad p(X_1 = 3) = 0.3$
- $p(X_1 = 1, X_2 = 0) = 0.4 \quad p(X_1 = 1, X_2 = 1) = 0.6,$
  $p(X_1 = 2, X_2 = 0) = 0.2 \quad p(X_1 = 2, X_2 = 1) = 0.8,$
  $p(X_1 = 3, X_2 = 0) = 0.5 \quad p(X_1 = 3, X_2 = 1) = 0.5$

True or false?

- $p(X = x, Y = y) \leq p(X = x)$ and $p(X = x, Y = y) \leq p(Y = y)$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$
- $p(X_1 = 1) = 0.1 \quad p(X_1 = 2) = 0.2 \quad p(X_1 = 3) = 0.3$
- $p(X_1 = 1, X_2 = 0) = 0.4 \quad p(X_1 = 1, X_2 = 1) = 0.6,$
  $p(X_1 = 2, X_2 = 0) = 0.2 \quad p(X_1 = 2, X_2 = 1) = 0.8,$
  $p(X_1 = 3, X_2 = 0) = 0.5 \quad p(X_1 = 3, X_2 = 1) = 0.5$

True or false?

- $p(X = x, Y = y) \leq p(X = x)$ and $p(X = x, Y = y) \leq p(Y = y)$
- $p(X = x|Y = y) \geq p(X = x)$

Quiz: Probabilities

X_1, X_2 random variables with $\mathcal{X}_1 = \{1, 2, 3\}$ and $\mathcal{X}_2 = \{0, 1\}$

What's wrong here?

- $p(X_1 = 1) = 1 \quad p(X_1 = 2) = 0 \quad p(X_1 = 3) = -1$
- $p(X_1 = 1) = 0.1 \quad p(X_1 = 2) = 0.2 \quad p(X_1 = 3) = 0.3$
- $p(X_1 = 1, X_2 = 0) = 0.4 \quad p(X_1 = 1, X_2 = 1) = 0.6,$
  $p(X_1 = 2, X_2 = 0) = 0.2 \quad p(X_1 = 2, X_2 = 1) = 0.8,$
  $p(X_1 = 3, X_2 = 0) = 0.5 \quad p(X_1 = 3, X_2 = 1) = 0.5$

True or false?

- $p(X = x, Y = y) \leq p(X = x)$ and $p(X = x, Y = y) \leq p(Y = y)$
- $p(X = x | Y = y) \geq p(X = x)$
- $p(X = x, Y = y) \leq p(X = x | Y = y)$

Dependence/Independence

Not every random variable is informative about every other.

- We say X **is independent of** Y if

  $$P(X = x, Y = y) = P(X = x)P(Y = y) \qquad \text{for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$

- equivalent (if defined):

  $$P(X = x|Y = y) = P(X = x), \qquad P(Y = y|X = x) = P(Y = y)$$

Dependence/Independence

Not every random variable is informative about every other.

- We say X **is independent of** Y if

  $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

- equivalent (if defined):

  $P(X = x|Y = y) = P(X = x), \quad P(Y = y|X = x) = P(Y = y)$

Example: Image datasets

- X_1: pick a random image from VOC2006. Does it show a `cat`?
- X_2: again pick a random image from VOC2006. Does it show a `cat`?
- $p(X_1 = \texttt{true}, X_2 = \texttt{true}) = p(X_1 = \texttt{true})p(X_2 = \texttt{true})$

Example: Video

- Y_1: does the first frame of a video show a `cat`?
- Y_2: does the second image of video show a `cat`?
- $p(Y_1 = \texttt{true}, Y_2 = \texttt{true}) \gg p(X_1 = \texttt{true})p(X_2 = \texttt{true})$

Expected value

We apply a function to (the values of) one or more random variables:

- $f(x) = \sqrt{x}$ or $f(x_1, x_2, \ldots, x_k) = \dfrac{x_1 + x_2 + \cdots + x_k}{k}$

The **expected value** or **expectation** of a function f with respect to a probability distribution is the weighted average of the possible values:

$$\mathbb{E}_{x \sim p(x)}[f(x)] := \sum_{x \in \mathcal{X}} p(x) f(x)$$

In short, we just write $\mathbb{E}_x[f(x)]$ or $\mathbb{E}[f(x)]$ or $\mathbb{E}[f]$ or $\mathbb{E}f$.

Example: rolling dice

Let X be the outcome of rolling a die and let $f(x) = x$

$$\mathbb{E}_{x \sim p(x)}[f(x)] = \mathbb{E}_{x \sim p(x)}[x] = \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6 = 3.5$$

Properties of expected values

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x, y) = x + y$

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

Properties of expected values

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x, y) = x + y$

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

The expected value has a useful property: it is *linear* in its argument.

- $\mathbb{E}_{x\sim p(x)}[f(x) + g(x)] = \mathbb{E}_{x\sim p(x)}[f(x)] + \mathbb{E}_{x\sim p(x)}[g(x)]$
- $\mathbb{E}_{x\sim p(x)}[\lambda f(x)] = \lambda \mathbb{E}_{x\sim p(x)}[f(x)]$

If a random variables does not show up in a function, we can ignore the expectation operation with respect to it

- $\mathbb{E}_{(x,y)\sim p(x,y)}[f(x)] = \mathbb{E}_{x\sim p(x)}[f(x)]$

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x, y) = x + y$

$$\begin{aligned}\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1 + x_2] \\ &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1] + \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_2] \\ &= \mathbb{E}_{x_1\sim p(x_1)}[x_1] + \mathbb{E}_{x_2\sim p(x_2)}[x_2] = 3.5 + 3.5 = \mathbf{7}\end{aligned}$$

The expected value has a useful property: it is *linear* in its argument.

- $\mathbb{E}_{x\sim p(x)}[f(x) + g(x)] = \mathbb{E}_{x\sim p(x)}[f(x)] + \mathbb{E}_{x\sim p(x)}[g(x)]$
- $\mathbb{E}_{x\sim p(x)}[\lambda f(x)] = \lambda \mathbb{E}_{x\sim p(x)}[f(x)]$

If a random variables does not show up in a function, we can ignore the expectation operation with respect to it

- $\mathbb{E}_{(x,y)\sim p(x,y)}[f(x)] = \mathbb{E}_{x\sim p(x)}[f(x)]$

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

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

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Answer 1: explicit calculation with dependent X_1 and X_2

$$p(x_1, x_2) = \begin{cases} \frac{1}{6} & \text{for combinations } (1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1) \\ 0 & \text{for all other combinations.} \end{cases}$$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \sum_{(x_1,x_2)} p(x_1, x_2)(x_1 + x_2)$$

$$= 0(1+1) + 0(1+2) + \cdots + \frac{1}{6}(1+6) + 0(2+1) + \cdots = 6 \cdot \frac{7}{6} = 7$$

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Answer 2: use properties of expectation as earlier

$$\begin{aligned}\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1 + x_2] \\ &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1] + \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_2] \\ &= \mathbb{E}_{x_1\sim p(x_1)}[x_1] + \mathbb{E}_{x_2\sim p(x_2)}[x_2] = 3.5 + 3.5 = 7\end{aligned}$$

The rules of probability take care of dependence, etc.