

# Lecture 2

Cezar Ionescu

18/05

DEPARTMENT FOR  
CONTINUING  
EDUCATION



**No lecture on the 1st of June!**

Homework from 2019-05-11 due **now!**

# Questions?

# Probability theory

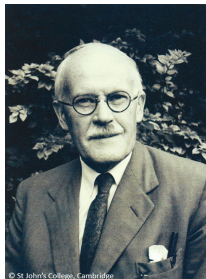
Richard von Mises (1883-1953):

*The term probability is meaningful for us only with regard to a clearly defined collective (or population).*



Sir Harold Jeffreys (1891-1989):

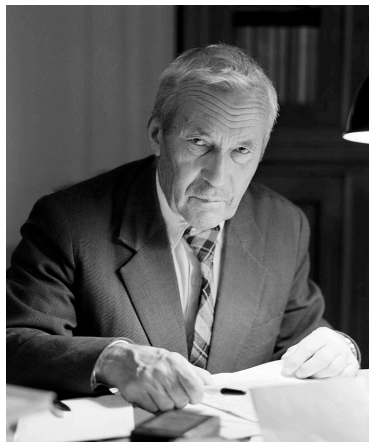
*[T]here is a valid primitive idea expressing the degree of confidence that we may reasonably have in a proposition.*



# Mathematical model

A. N. Kolmogorov (1903-1987):

- $\Omega$ , the set of all possible outcomes
- **Event**, the set of all events  
 $\text{Event} \subseteq \mathbb{P}(\Omega)$
- $p : \text{Event} \rightarrow [0, 1]$ , the probability measure, satisfying
  - 1  $p(\emptyset) = 0$ ,  $p(\Omega) = 1$
  - 2 for all events  $X_1, \dots, X_n, \dots$  pairwise disjoint,  
 $\sum p(X_i) = p(\bigcup X_i)$



# Examples

① Drawing a card from a standard deck.

- $\Omega = \{2\clubsuit, 2\diamonds, 2\heartsuit, 2\spadesuit, \dots\}$
- $\text{Event} = \mathbb{P}(\Omega)$

② Choosing a point of the unit disc.

- $\Omega = \{(x, y) \mid x, y \in \mathbb{R}, x^2 + y^2 \leq 1\}$
- We would like to have  $\text{Event} = \mathbb{P}(\Omega)$  ... but we can't
- We can take  $\text{Event} = \{X : \Omega \rightarrow \{0, 1\} \mid X \text{ computable}\}$

# The classical model of probability

$\Omega$  finite, **Event** =  $\mathbb{P}(\Omega)$ , and  $p$  defined by

$$p(X) = \text{card}(X) / \text{card}(\Omega)$$

Interpretations:

- Frequentist: every elementary outcome  $\omega \in \Omega$  turns up with approximately the same frequency as any other.
- Bayesian: we have no reason to prefer one elementary outcome over any other.

# Example

Rolling a die:

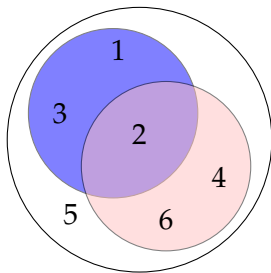
- $\Omega = \{1, 2, 3, 4, 5, 6\}$ ,  $\text{card}(\Omega) = 6$
- rolling an even number:  $\text{even} = \{2, 4, 6\}$ ,  $\text{card}(\text{even}) = 3$
- probability of rolling an even number:  
 $p(\text{even}) = \text{card}(\text{even})/\text{card}(\Omega) = 3/6 = 0.5$



# Operations with events

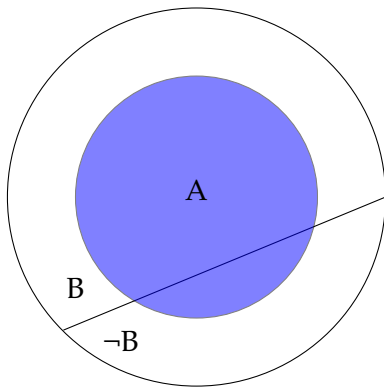
- union:  $A \cup B$
  - intersection:  $A \cap B$
  - complement:  $\neg A = \Omega - A$
- 

- result is even **and**  $\leq 3$
- result is even **or**  $\leq 3$
- result is **not**  $\leq 3$



# Exercises

- 1 Derive the formula for  $p(\neg A)$ .
- 2 Derive the formula for  $p(A \cup B)$
- 3 Assume that  $B_1, \dots, B_n$  are pairwise disjoint. Show that
$$p(A) = p(A \cap B_1) + \dots + p(A \cap B_n)$$



# Conditional probability

**Definition:** Let  $(\Omega, \text{Event}, p)$  as above,  $X \in \text{Event}$  s.t  $p(X) \neq 0$ . Let  $Y \in \text{Event}$ . The **conditional probability** of  $Y$  given  $X$  is defined by

$$p(Y \mid X) = p(Y \cap X) / p(X)$$

# Notational problems

- is  $|$  a set-theoretical operation?
- is  $|$  commutative?
- Kolmogorov's original notation:  $p_X(Y)$

# The law of total probability

$B_1, \dots, B_n$  pairwise disjoint,  $p(B_i) \neq 0$  for all  $i$ . Show that

$$p(A) = p(A \mid B_1) * p(B_1) + \dots + p(A \mid B_n) * p(B_n)$$

# Bayesian learning

- Set of hypotheses,  $H$
- Data,  $d \in D$
- Find the “best” hypothesis that fits the data
  - e.g., **Find-S**, “best” = the most specific hypothesis consistent with the data.
- Bayesian learning: “best” = *the most probable* hypothesis, given the data.

# Maximum a posteriori hypothesis

The aim of Bayesian learning is to determine the **maximum a posteriori (MAP) hypothesis**:

$$h_{\text{map}} = \operatorname{argmax}_h p(h \mid d)$$

What does  $p(h \mid d)$  mean?

# Bayes' theorem

$X, Y \in \text{Event}, p(X) * p(Y) \neq 0.$

$$p(X | Y) = p(X \cap Y) / p(Y)$$

$$\text{therefore } p(X \cap Y) = p(X | Y) * p(Y)$$

$$p(Y | X) = p(Y \cap X) / p(X)$$

$$\text{therefore } p(Y \cap X) = p(Y | X) * p(X)$$

$X \cap Y = Y \cap X \Rightarrow p(X \cap Y) = p(Y \cap X), \text{ therefore:}$

$$p(X | Y) = p(Y | X) * p(X) / p(Y)$$



# Example

Mitchell, pages 157-158:

- $H = \{\text{healthy}, \text{ill}\}$
- test results:  $\ominus$ ,  $\oplus$
- 0.008 of the population have the disease.
- patient ill  $\Rightarrow$  test positive in 98% of the cases
- patient healthy  $\Rightarrow$  test positive in 3% of the cases
- A patient's test has come back positive.

What is the MAP hypothesis?

# Translation to probability-speak

- $\Omega$  = the set of people in the population
- **Event** =  $\mathbb{P}(\Omega)$
- classical model for  $p$ :  $p(X) = \text{card}(X) / \text{card}(\Omega)$
- the events of interest: **Healthy, Ill, Pos, Neg**
- $p(\text{healthy}) = 0.992, p(\text{ill}) = 0.008, p(\oplus \mid \text{ill}) = 0.98,$   
 $p(\oplus \mid \text{healthy}) = 0.03$

Compute  $p(\text{ill} \mid \oplus)$ .

# Applying Bayes' theorem

$$\begin{aligned} & p(\text{ill} \mid \oplus) \\ = & p(\oplus \mid \text{ill}) * p(\text{ill}) / p(\oplus) \\ = & 0.98 * 0.008 / p(\oplus) \\ = & 0.00784 / p(\oplus) \end{aligned}$$

$$\begin{aligned} & p(\text{healthy} \mid \oplus) \\ = & p(\oplus \mid \text{healthy}) * p(\text{healthy}) / p(\oplus) \\ = & 0.03 * 0.992 / p(\oplus) \\ = & 0.02976 / p(\oplus) \end{aligned}$$

Therefore, the MAP hypothesis is healthy.

# A general argument

$$\begin{aligned}h_{\text{map}} &= \operatorname{argmax}_h p(h \mid d) \\&= \operatorname{argmax}_h (p(d \mid h) * p(h) / p(d)) \\&= \operatorname{argmax}_h (p(d \mid h) * p(h)) \text{ -- } p(d) \text{ is constant w.r.t. } h\end{aligned}$$

# Likelihood

If all hypotheses are equally likely:

$$\begin{aligned}h_{\text{map}} &= \operatorname{argmax}_h p(h \mid d) \\ &= \operatorname{argmax}_h p(d \mid h)\end{aligned}$$

$p(d \mid h)$  is called the **likelihood** of  $d$  given  $h$ .

A hypothesis that maximises the likelihood is a *maximum likelihood hypothesis (ML)*. Therefore

$$h_{\text{ml}} = \operatorname{argmax}_h p(d \mid h)$$

In general,  $h_{\text{map}} \neq h_{\text{ml}}$ !

# Exercise

What is  $h_{m1}$  in the example above?

# Homework

Apply Bayes' theorem to answer the following question (Elmer Mode 1966, page 53):

*A class in advanced mathematics contains 10 juniors, 30 seniors, and 10 graduate students. Three of the juniors, 10 of the seniors, and 5 of the graduate students received an A in the course. If a student is chosen at random from this class and is found to have earned an A, what is the probability that he is a graduate student?*

# Concept learning revisited

Concept:

$$c : X \rightarrow \{0, 1\}$$

$$h \in H \Rightarrow h : X \rightarrow \{0, 1\}$$

We assume  $c \in H$

The data:

$$d = \{(x_1, c(x_1)), \dots, (x_m, c(x_m))\}$$



# “Brute-force” Bayesian concept learning

```
given  $H$ ,  $p(h)$   
for every  $h \in H$ :  
    compute  $p(h \mid d) = p(d \mid h) * p(h) / p(d)$   
    return  $h_{\text{map}} = \operatorname{argmax}_h p(h \mid d)$ 
```

The problem is computing, for every  $h$ ,  $p(d \mid h)$ . We also want to compute  $p(d)$ .

# Computing the likelihood

Assumption: the data is completely correct, and hypotheses are not “noisy” functions.

$$p(d \mid h) = \begin{cases} 1 & \text{if for all } (x_i, b_i) \in d \text{ } h(x_i) = b_i \\ 0 & \text{else} \end{cases}$$

## Computing $p(d \mid h)$

We can now compute  $p(d)$  by applying the law of total probability:

$$p(d) = \sum_h p(d \mid h) * p(h)$$

In a certain sense, the data and the hypothesis must have the same nature.

# Consistent learners

Assume all hypotheses are equally likely. Then the brute-force Bayesian concept learning algorithm will return an ML hypothesis:

$$h_{\text{map}} = h_{\text{ml}} = \operatorname{argmax}_h p(d \mid h)$$

Since  $p(d \mid h) = 1$  if  $h$  is consistent with the data, and  $p(d \mid h) = 0$  otherwise, the algorithm will return any one of the hypotheses in  $H$  consistent with  $d$ .

# Bayesian concept learning and Find-S

**Find-S** returns a consistent hypothesis  $\Rightarrow$  it returns a MAP (and an ML) hypothesis!

Even if  $p$  is not constant w.r.t. hypotheses, **Find-S** can still return a MAP hypothesis, if there is a unique most specific hypothesis consistent with  $d$  in  $H$  and

$$h_i \subseteq h_j \Rightarrow p(h_i) \geq p(h_j)$$