Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González

Probability Overview

- Events

  Discrete random variables, continuous random variables, compound events
- Axioms of probability

  What defines a reasonable theory of uncertainty
- Conditional probabilities
- Chain rule
- Bayes rule
- Joint probability distribution
- $P(Y|X)$: Facing practical problems
  - Conditional independencies for model simplification
  - Principles of parameter estimation

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment
- Examples
  - A : True if a randomly drawn person from our class is female
  - A : The hometown of a randomly drawn person from our class
  - A : True if two randomly drawn persons from our class have same birthday
- Define P(A) as "the fraction of possible worlds in which A is true" or "the fraction of times A holds, in repeated runs of the random experiment"
  - The set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S

A bit of formalism

More formally, we have:

- a sample space S, a.k.a. the set of possible worlds

  E.g., set of students in our class

- a random variable is a function defined over the sample space
  Gender: $S \rightarrow \{m, f\}$

  Height: $S \rightarrow \mathbb{R}$

- an event is a subset of S
  E.g., the subset of S for which Gender=f

  E.g., the subset of S for which (Gender=m) AND (eyeColor=blue)

- we are often interested in probabilities of specific events
- and of specific events conditioned on other specific events

A bit of formalism

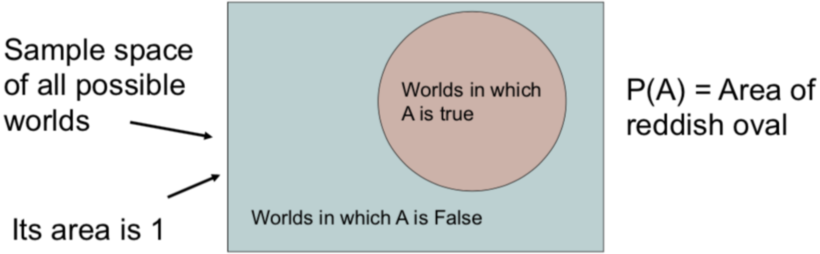
- X: random variable
  (UPPERCASE)
- $\boldsymbol{X} = (X_1, \ldots, X_n)$: random vector
  (BOLD UPPERCASE)
- Ω_X: possible values of random variable X.
- $x \in \Omega_X$: a possible value of random variable X: $X = x$
  (lowercase)
- $\boldsymbol{x} = (x_1, \ldots, x_n)$: a possible (tuple of) value of a random vector $\boldsymbol{X}$: $(X_1 = x_1, \ldots, X_n = x_n), \forall X_i : x_i \in \Omega_{X_i}$
  (**bold lowercase**)

A bit of formalism

- X: random variable
  (UPPERCASE)
- $\boldsymbol{X} = (X_1, \ldots, X_n)$: random vector
  (BOLD UPPERCASE)
- Ω_X: possible values of random variable X.
- $x \in \Omega_X$: a possible value of random variable X: $X = x$
  (lowercase)
- $\boldsymbol{x} = (x_1, \ldots, x_n)$: a possible (tuple of) value of a random vector $\boldsymbol{X}$: $(X_1 = x_1, \ldots, X_n = x_n), \forall X_i : x_i \in \Omega_{X_i}$
  (**bold lowercase**)

  *** Instantiation**

Visualizing A



A bit of formalism

The Axioms of Probability:

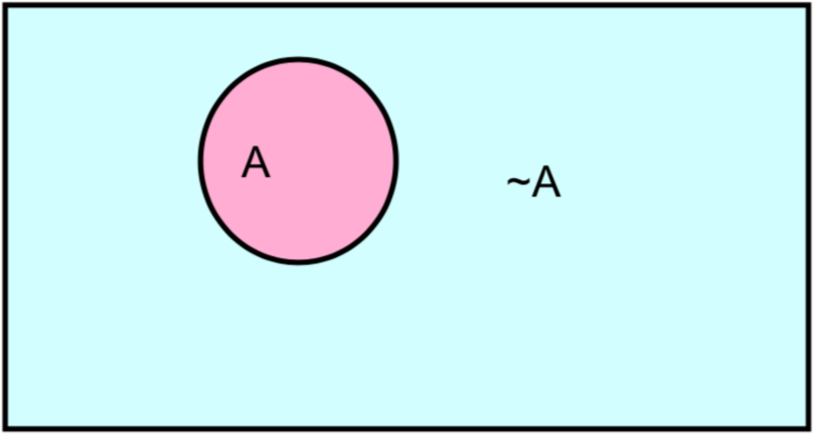
- $0 \leq P(A) \leq 1$
- $P(True) = 1$
- $P(False) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

[di Finetti 1931]:

when gambling based on "uncertainty formalism A" you can be exploited by an opponent
iff
your uncertainty formalism A violates these axioms

Elementary probability in pictures

$$P(\neg A) + P(A) = 1$$



A useful theorem

$$0 \leq P(A) \leq 1, P(\mathit{True}) = 1, P(\mathit{False}) = 0$$

$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$$

$$A = [A \text{ and } (B \text{ or } \neg B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \neg B)]$$

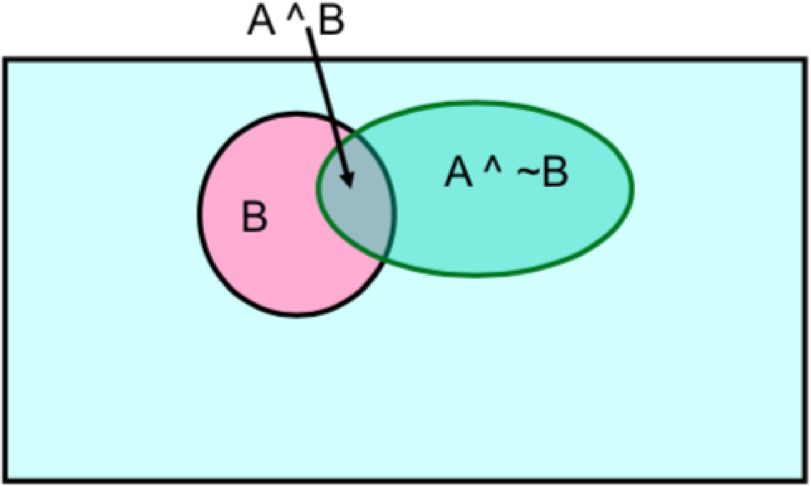
$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B) - P((A \text{ and } B) \text{ and } (A \text{ and } \neg B))$$

$$= P(A \text{ and } B) + P(A \text{ and } \neg B) \color{red}{- P(A \text{ and } B \text{ and } A \text{ and } \neg B)}$$

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B)$$

Elementary probability in pictures

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B) \qquad [\text{and} \equiv \wedge]$$

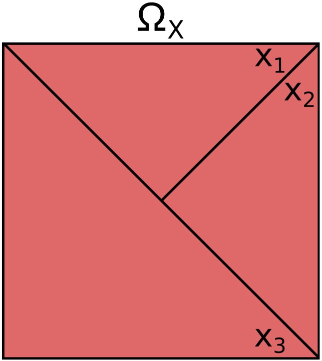


Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x)\right) = 1$$



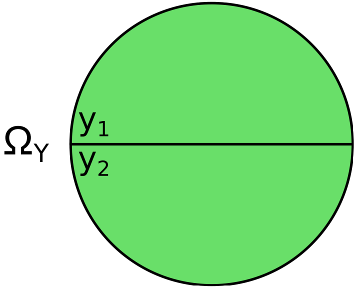
$$p(X = x_1) = 1/4$$
$$p(X = x_2) = 1/4$$
$$p(X = x_3) = 1/2$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x)\right) = 1$$



$p(Y = y_1) = 1/2$
$p(Y = y_2) = 1/2$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x)\right) = 1$$

$p(\boldsymbol{X})$: joint probabilities

$$0 \leq p(\boldsymbol{X} = \boldsymbol{x}) \leq 1, \forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}$$

$$\left(\sum_{\boldsymbol{x} \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}} p(\boldsymbol{X} = \boldsymbol{x})\right) = 1$$

$$p(\boldsymbol{X}) = p(X_1, \ldots, X_n) = p(X_1 \cap \cdots \cap X_n)$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x)\right) = 1$$

$p(\boldsymbol{X})$: joint probabilities

$$0 \leq p(\boldsymbol{X} = \boldsymbol{x}) \leq 1, \forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}$$

$$\left(\sum_{\boldsymbol{x} \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}} p(\boldsymbol{X} = \boldsymbol{x})\right) = 1$$

$$p(\boldsymbol{X}) = p(X_1, \ldots, X_n) = p(X_1 \cap \cdots \cap X_n)$$

Marginalization

$$p(Y) = \sum_{x \in \Omega_X} p(Y|X = x) \cdot p(X = x)$$

Concepts of probability

$p(X)$: marginal probabilities

$$0 \leq p(X = x) \leq 1, \forall x \in \Omega_X$$

$$\left(\sum_{x \in \Omega_X} p(X = x)\right) = 1$$

$p(\boldsymbol{X})$: joint probabilities

$$0 \leq p(\boldsymbol{X} = \boldsymbol{x}) \leq 1, \forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}$$

$$\left(\sum_{\boldsymbol{x} \in \Omega_{X_1} \times \cdots \times \Omega_{X_n}} p(\boldsymbol{X} = \boldsymbol{x})\right) = 1$$

$$p(\boldsymbol{X}) = p(X_1, \ldots, X_n) = p(X_1 \cap \cdots \cap X_n)$$

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x)\right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$

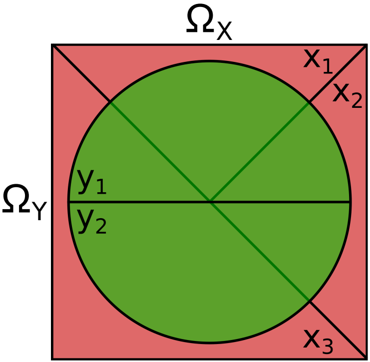
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y=y|X=x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y=y|X=x)\right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$p(Y=y_1|X=x_1) =$
$p(Y=y_2|X=x_1) =$
$p(Y=y_1|X=x_2) =$
$p(Y=y_2|X=x_2) =$
$p(Y=y_1|X=x_3) =$
$p(Y=y_2|X=x_3) =$

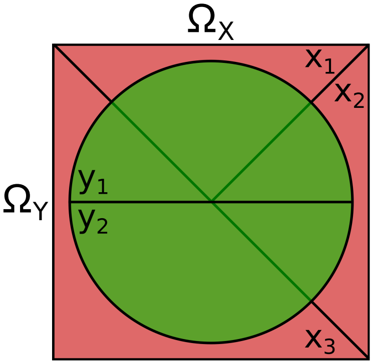
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x)\right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$p(Y = y_1|X = x_1) = 1$
$p(Y = y_2|X = x_1) = 0$
$p(Y = y_1|X = x_2) =$
$p(Y = y_2|X = x_2) =$
$p(Y = y_1|X = x_3) =$
$p(Y = y_2|X = x_3) =$

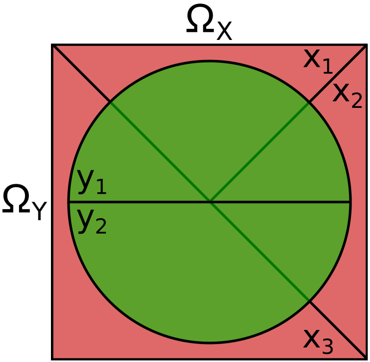
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x)\right) = 1, \forall x \in \Omega_X$$

$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$$
\begin{aligned}
p(Y = y_1|X = x_1) &= 1\\
p(Y = y_2|X = x_1) &= 0\\
p(Y = y_1|X = x_2) &= 1/2\\
p(Y = y_2|X = x_2) &= 1/2\\
p(Y = y_1|X = x_3) &=\\
p(Y = y_2|X = x_3) &=
\end{aligned}
$$

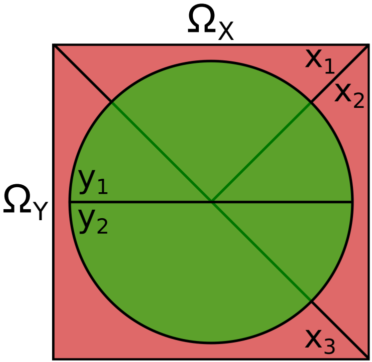
Concepts of probability

$p(Y|X)$: conditional probabilities

$$0 \leq p(Y = y|X = x) \leq 1, \forall x \in \Omega_X \wedge y \in \Omega_Y$$

$$\left(\sum_{y \in \Omega_Y} p(Y = y|X = x)\right) = 1, \forall x \in \Omega_X$$

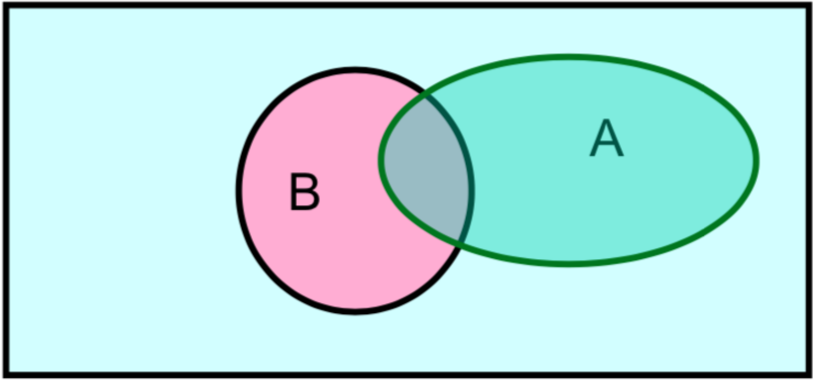
$$p(Y|X) = \frac{p(X,Y)}{p(X)} \text{ with } p(X) \neq 0$$



$p(Y = y_1|X = x_1) = 1$
$p(Y = y_2|X = x_1) = 0$
$p(Y = y_1|X = x_2) = 1/2$
$p(Y = y_2|X = x_2) = 1/2$
$p(Y = y_1|X = x_3) = 1/4$
$p(Y = y_2|X = x_3) = 3/4$

Definition of Conditional Probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$



Definition of Conditional Probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

The Chain Rule

$$P(A, B) = P(A|B) \cdot P(B)$$

Chain rule

Chain rule

For all $\boldsymbol{x}$ we have that

$$p(\boldsymbol{x}) = p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

** it holds for any ordering of $X_1, ..., X_n$

- Joint distribution as a product of conditional probabilities
- Example:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$

Exercise

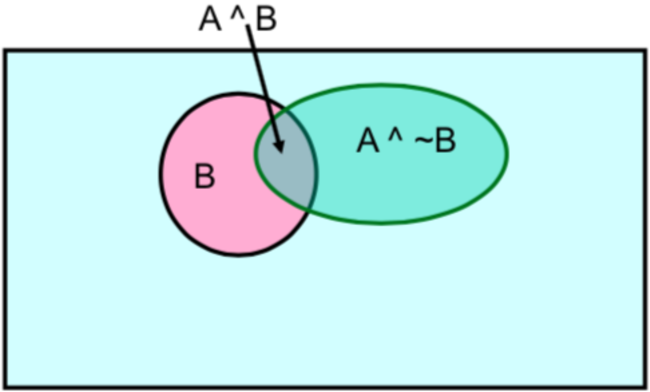
Chain rule

Let be $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$, show that for all $\boldsymbol{x}$ we have that $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$

** Remember that $p(x_i|x_1, ..., x_{i-1}) = \frac{p(x_1,...,x_i)}{p(x_1,...,x_{i-1})}$

Bayes rule

2 expressions for $P(A, B)$



Bayes rule

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

We call $P(A)$ the "prior"

and $P(A|B)$ the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

This is also the Bayes rule

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

$$P(A|B, C) = \frac{P(B|A, C) \cdot P(A, C)}{P(B, C)}$$

Exercise

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

A : you have a flu, B : you just coughed

Assume:

- $P(A) = 0{,}05$
- $P(B|A) = 0{,}80$
- $P(B|\neg A) = 0{,}2$

what is you have a flu given that you just coughed?

Independent events

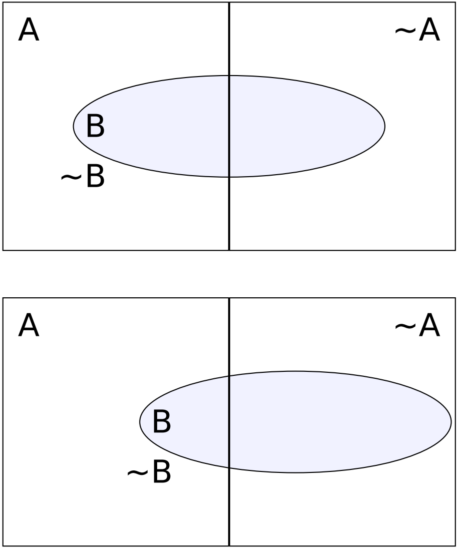
Definition

Two events A and B are independent if

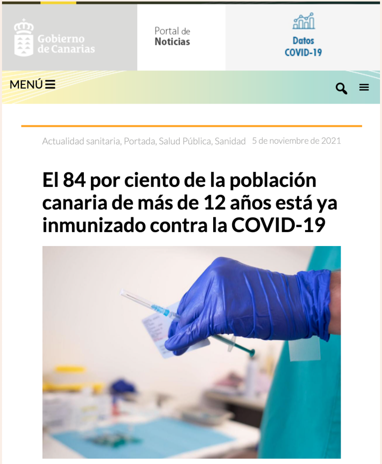
$$P(A, B) = P(A) \cdot P(B)$$

** Intuition **

Knowing A tells us nothing about the value of B (and vice versa)

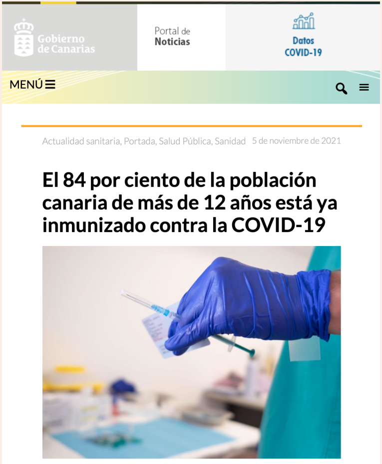


Exercise



- Can you tell which is the probability of ending up in the hospital if you get COVID19 and you are not vaccinated?

Exercise



- Can you tell which is the probability of ending up in the hospital if you get COVID19 and you are not vaccinated?
- Can say anything?

What does all this have to do with function approximation?

instead of $F : X \rightarrow Y$,

learn $P(Y|X)$

Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values
   (M variables $\rightarrow 2^M$ combinations)

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values
   (M variables $\rightarrow 2^M$ combinations)
2. Say how probable each combination is

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

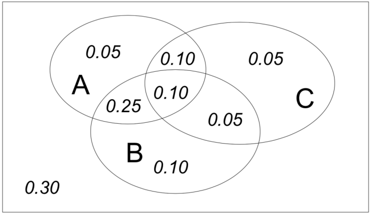
Joint distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all possible combinations of values
   (M variables $\rightarrow 2^M$ combinations)
2. Say how probable each combination is

   Subscribed to the axioms of probability if sum to 1

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Exercise

Using the joint distribution

You can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{r: rows matching } E} P(r)$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Exercise

Using the joint distribution

$$P(E) = \sum_{\text{r: rows matching } E} P(r)$$

$$P(\text{Poor} \wedge \text{Male}) = ?$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Exercise

Using the joint distribution

$$P(E) = \sum_{\text{r: rows matching } E} P(r)$$

$$P(\text{Poor}) = ?$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Inference with the joint distribution

You can ask for the probability of any logical expression involving a subset of attributes given another expression involving other attributes

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{\text{r: rows matching } E_1 \text{ \& } E_2} P(r)}{\sum_{\text{o: rows matching } E_2} P(o)}$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Exercise

Inference with the joint distribution

You can ask for the probability of any logical expression involving a subset of attributes given another expression involving other attributes

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{r:\text{ rows matching } E_1 \text{ \& } E_2} P(r)}{\sum_{o:\text{ rows matching } E_2} P(o)}$$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(Male|Poor) = ?$$

Learning and the joint distribution

Suppose we want to learn the function $f :< G, H >\rightarrow W$

Equivalently, $P(W|G, H)$

Solution:

- Learn joint distribution from train data
- Calculate $P(W|G, H)$ for test data

  E.g., given a *female* patient of 39 years old:

  $\arg\,\text{máx}_{w\in\{rich,poor\}} P(W = w|G = female, H = 40{,}5-)$

Solution?

$P(Y|X)$ sounds like a nice alternative solution to function $F : X \to Y$

Are we done?

Solution?

$P(Y|X)$ sounds like a nice alternative solution to function $F : X \to Y$

Are we done?

Main problem

Learning $P(Y|X)$ may require more data than we have

Solution?

$P(Y|X)$ sounds like a nice alternative solution to function $F : X \to Y$

Are we done?

Main problem

Learning $P(Y|X)$ may require more data than we have

E.g., consider learning the joint distribution for 100 binary variables

- # of rows in this table?
- # of data samples to learn faithfully?
- # of rows never observed?

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
   - maximum likelihood estimates
   - maximum a posteriori estimates

Facing practical problems

What to do?

1. **Be smart about how to represent joint distributions**
   - Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
   - maximum likelihood estimates
   - maximum a posteriori estimates

From definition to representation

Joint distribution

Let $\boldsymbol{X}_V$ be a set of variable, then

$$\forall \boldsymbol{x}_V, p(\boldsymbol{x}_V) = p(x_1, ..., x_v)$$

- A mapping: $\boldsymbol{x}_V \mapsto [0, 1]$
- $\sum_{\boldsymbol{x}_V} p(\boldsymbol{x}_V) = 1$
- Number of free parameters: $\prod_{i=1}^{v} r_i - 1 = r_V - 1$
- Exponential in the number of variables, v

From definition to representation

Marginal distribution

Let A and B be a partition of V. Then,

$$\forall \boldsymbol{x}_A, \boldsymbol{x}_B, p(\boldsymbol{x}_A) = \sum_{\boldsymbol{x}_B} p(\boldsymbol{x}_A, \boldsymbol{x}_B)$$

- Number of free parameters: $\prod_{i \in A} r_i - 1 = r_A - 1$

From definition to representation

Conditional distribution

Let A and B be a partition of V

$$\forall \boldsymbol{x}_A, p(\boldsymbol{x}_A|\boldsymbol{x}_B) = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{p(\boldsymbol{x}_B)} = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{\sum_{\boldsymbol{x}_A} p(\boldsymbol{x}_A, \boldsymbol{x}_B)}$$

- Family of distribution:
  A marginal distribution for each value assignment $\boldsymbol{X}_B = \boldsymbol{x}_B$
  $\sum_{\boldsymbol{x}_A} p(\boldsymbol{x}_A|\boldsymbol{x}_B) = 1$
- Number of free parameters:
  $(\prod_{i\in A} r_i - 1) \cdot (\prod_{j\in B} r_j) = (r_A - 1) \cdot r_B$

Exercise

X_1	X_2	$p(X_1, X_2)$
-	banana	0.1
-	apple	0.3
-	pear	0.2
$+$	banana	0.1
$+$	apple	0.2
$+$	pear	0.1

- Determine the domains of X_1 and X_2.
- Obtain the marginal distributions $p(X_1)$ y $p(X_2)$
- Obtain the conditional distributions $p(X_1|X_2 = apple)$ y $p(X_2|X_1 = +)$
- Compute the number of free parameters $p(X_1)$, $p(X_1|X_2)$ y $p(X_1, X_2)$

Conditional independence

A *qualitative* relationship between random variables

Let A, B, C be disjoint subsets of $V = \{1, ..., v\}$. We say that $\boldsymbol{X}_A$ is independent from $\boldsymbol{X}_B$ given $\boldsymbol{X}_C$ if and only if for all $(\boldsymbol{x}_A, \boldsymbol{x}_B, \boldsymbol{x}_C)$ we have that $p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$.

- Denoted by $X_A \perp\!\!\!\perp X_B|X_C$
- $p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$: Knowing/observing/fixing $\boldsymbol{x}_C$, the value $\boldsymbol{x}_B$ does not modify the probability of $\boldsymbol{x}_A$
- Exercise: Prove that
  $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C) \cdot p(\boldsymbol{x}_B|\boldsymbol{x}_C)$

Conditional independence

- Allow to simplify the factorization given by the chain rule
- Choose an appropriate ordering that allows to apply the independence over a conditional distribution

Example

- $\boldsymbol{X} = X_1, ..., X_5$
- $3 \perp\!\!\!\perp 4 | 1, 5$
- Ordering: $1, 4, 5, 3, 2$

$$
\begin{aligned}
p(\boldsymbol{X}) &= p(\boldsymbol{X}_{1,4,5})p(X_3|\boldsymbol{X}_{1,4,5})p(X_2|\boldsymbol{X}_{1,3,4,5}) \\
&= p(\boldsymbol{X}_{1,4,5})p(X_3|\boldsymbol{X}_{1,5})p(X_2|\boldsymbol{X}_{1,3,4,5})
\end{aligned}
$$

Exercise

True or false:

- $X_A \perp\!\!\!\perp X_B \Rightarrow p(\boldsymbol{x}_A|\boldsymbol{x}_B) = p(\boldsymbol{x}_A)$
- $X_A \perp\!\!\!\perp X_B \Rightarrow p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$
- $X_A \perp\!\!\!\perp X_B \Rightarrow p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)$
- $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)p(\boldsymbol{x}_B|\boldsymbol{x}_C)$
- $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\boldsymbol{x}_A|\boldsymbol{x}_B, \boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)p(\boldsymbol{x}_B|\boldsymbol{x}_C)$
- $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)p(\boldsymbol{x}_B|\boldsymbol{x}_A, \boldsymbol{x}_C)$

Counting the parameters

Conditional independences reduce the number of (free) parameters

- $X_A \perp\!\!\!\perp X_B$
  - $\forall \boldsymbol{x}$: $p(\boldsymbol{x}_A, \boldsymbol{x}_B) = p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)$
  - From $r_A \cdot r_B - 1$ to $r_A - 1 + r_B - 1$

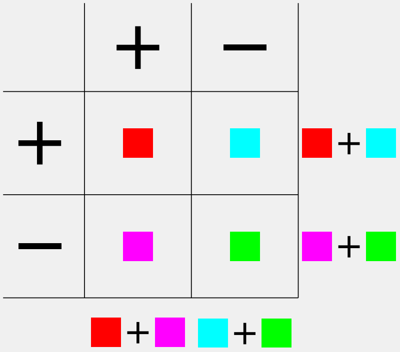
    E.g., $r_A = 3$, $r_B = 4$
    From $3 \times 4 - 1$ to $2 + 3$

- $X_A \perp\!\!\!\perp X_B | X_C$
  - $\forall \boldsymbol{x}$: $p(\boldsymbol{x}_A, \boldsymbol{x}_B | \boldsymbol{x}_C) = p(\boldsymbol{x}_A | \boldsymbol{x}_C)p(\boldsymbol{x}_B | \boldsymbol{x}_C)$
  - From $(r_A \cdot r_B - 1) \cdot r_C$ to $(r_A - 1 + r_B - 1) \cdot r_C$

    E.g., $r_A = 3$, $r_B = 4$, $r_C = 3$
    From $(3 \times 4 - 1) \times 3$ to $(2 + 3) * 3$

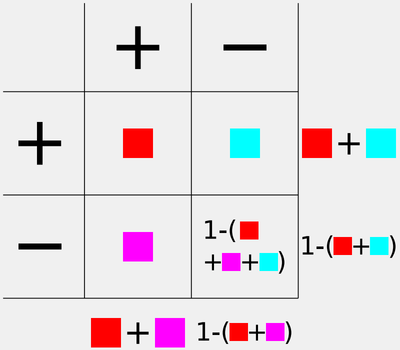
Counting the parameters, without independence

$p(A, B)$: three parameters



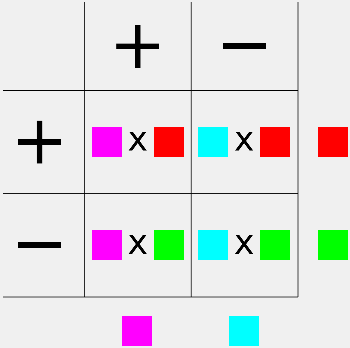
Counting the parameters, without independence

$p(A, B)$: three parameters



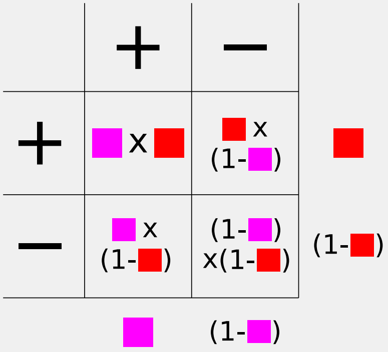
Counting the parameters, with independence

$p(A, B)$: two parameters



Counting the parameters, with independence

$p(A, B)$: two parameters



Conditional independence

Discarding parameters

The complexity of a statistical model can be understood as its flexibility for learning or its requirement of memory

Conditional independences:

- Reduce the complexity (i.e., # parameters) of the statistical model represented by the (simplified) chain rule
- Allow for avoiding to model (irrelevant) parameters associated to soft conditional dependences
- Help to deal with the trade-off between the complexity of the statistical model and amount of train data
  This has crucial implications in statistical models, e.g., overfitting

Exercise

Let be $r_A = 6, r_B = 4, r_C = 8$.

- $\forall \boldsymbol{x}, p(\boldsymbol{X}) = p(\boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{X}_C)$
- $\forall \boldsymbol{x}, p(\boldsymbol{X}) = p(\boldsymbol{X}_A)p(\boldsymbol{X}_B)p(\boldsymbol{X}_C)$
- $\forall \boldsymbol{x}, p(\boldsymbol{X}) = p(\boldsymbol{X}_A|\boldsymbol{X}_B)p(\boldsymbol{X}_B)p(\boldsymbol{X}_C)$
- $\forall \boldsymbol{x}, p(\boldsymbol{X}) = p(\boldsymbol{X}_A|\boldsymbol{X}_C)p(\boldsymbol{X}_B|\boldsymbol{X}_C)p(\boldsymbol{X}_C)$
- $\forall \boldsymbol{x}, p(\boldsymbol{X}) = p(\boldsymbol{X}_A)p(\boldsymbol{X}_B|\boldsymbol{X}_A)p(\boldsymbol{X}_C|\boldsymbol{X}_A, \boldsymbol{X}_B)$

1. Calculate the number of free parameters
2. Read conditional independences from factorizations

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. Be smart about how we estimate probabilities from *sparse* data
   - maximum likelihood estimates
   - maximum a posteriori estimates

Facing practical problems

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. **Be smart about how we estimate probabilities from *sparse* data**
   - maximum likelihood estimates
   - maximum a posteriori estimates

You should know

- Events

  Discrete random variables, continuous random variables, compound events
- Axioms of probability

  What defines a reasonable theory of uncertainty
- Conditional probabilities
- Chain rule
- Bayes rule
- Joint probability distribution

  How to calculate other quantities from the joint distribution
- $P(Y|X)$: Facing practical problems
  - Conditional independencies for model simplification
  - Principles of parameter estimation

Notation

- Indices $V = \{1, \ldots, n\}$, subsets of indices $A, B, C \subseteq V$
- Random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{X}_A = (X_i : i \in A)$
- Domain of X_i, Ω_i con $|\Omega_i| = r_i$
- Instance/case/sample: $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{x}_A = (X_i : i \in A)$
- Data set: $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}$
- Probability distribution: $\boldsymbol{X}$ distributed according to $p(\boldsymbol{X})$
- Probability of observing $\boldsymbol{x}$: $p(\boldsymbol{x})$

Probability overview

Probabilistic Graphical Models

Jerónimo Hernández-González