Review of Probability Theory

Javier Larrosa

UPC Barcelona Tech

.

Random Variables

Random Variable

A **random variable** A can take a finite set of **values** $\{a_1, a_2, \ldots, a_m\}$. They are mutually exclusive (i.e, only one value can be taken at a time) and exhaustive (i.e, there is no other value that the variable can take).

- $A = a_i$ means that the variable A takes (or will take) value a_i
- We can think of $A = a_i$ as an **event** that either is or is not true

Random Variables

Example

Let A be the temperature (in Celsius) in Barcelona next Tuesday at noon. Its domain values are for example [10..30]

Example

Let A be the position in which Barça will end up the football league. Its domain values are [1..16]

Example

Let A be the day of the week in which I was born. Its domain values are $\{$ M,T,W,...S$\}$

Univariate probability distributions

Univariate probability distributions

A **probability distribution** of random variable A is a set of numbers $P(A) = \{P(A = a_1), P(A = a_2), \cdots, P(A = a_m)\}$ such that:

- for all a_i, $0 \leq P(A = a_i) \leq 1$
- $\sum_{i=1}^{m} P(A = a_i) = 1$

- We can think of $P(A = a_i)$ (also written $P(a_i)$) as our **belief** that variable A takes value a_i
- $P(a_i) = 1$ means that we are completely sure that $A = a_i$
- $P(a_i) = 0$ means that we are completely sure that $A \neq a_i$
- $0 < P(a_i) < 1$ means that we have no certainty about the event
  - $P(a_i) < P(a_j)$ means that event a_j is more likely than event a_i

Probability Distributions

Example

Let A be the temperature (in Celsius) in Barcelona next Tuesday at noon. Its domain values are for example [10..30]

$$P(A = 20) = 0.3, P(A = 30) = 0.001, \cdots$$

Example

Let A be the position in which Barça will end up the football league. Its domain values are [1..16]

$$P(A = 1) = .3, P(A = 2) = .3, P(A = 3) = .15, P(A = 4) = .05, \ldots$$

Parenthesis (What are probabilities?)

Events can be different things:

- will a rolling dice get a 3?
- will Barça win the next league?
- I was born on Sunday?
- is there intelligent life out of Earth?

Frequentist Interpretation

relative frequency with which an event is true,

Subjective or Bayesian Interpretation

degree of belief that an agent attaches to the likelihood that a is true

Joint probabilities

The real power in probabilistic modeling comes from modeling sets of random variables.

multivariate prob. distributions

Let A and B two random variables taking values in $\{a_1, a_2, \ldots, a_m\}$ and $\{b_1, b_2, \ldots, b_{m'}\}$, respectively. Their join distribution is

$$P(A, B) = \{P(a_i, b_j) |\ 1 \leq i \leq m, 1 \leq j \leq m'\}$$

with,

- for all a_i, b_j, $0 \leq P(a_i, b_j) \leq 1$
- for all a_i, b_j, $\sum_{i=1}^{m} \sum_{j=1}^{m'} P(a_i, b_j) = 1$

- We can think of the pair of variables A, B as a "super- variable" AB defined as the cross-product of individual variables A and B
- AB has $(m \cdot m')$ values in its domain
- It can be generalized to any number of variables $P(A, B, C, D)$

Join Probabilities Example

Let A and B be the temperature (in Celsius) in Barcelona and Hong Kong next Tuesday at noon. Its domain values are for example [10..30]. As an example,

- $P(A = 20) = .3$
- $P(B = 24) = .25$
- $P(A = 20, B = 24) = .16$

From Join Probabilities to Marginals and Vice-versa

How to compute a probability $P(a)$ from a joint distribution

Law of Total Probability (marginalization)

$$P(a) = \sum_{b} P(a, b)$$

It can be generalized,

$$P(a, b) = \sum_{c,d} P(a, b, c, d)$$

NOTATION ON VARIABLES: An upper case X means free-variable, lower-case x means $X = x$ (i.e, the variable is instantiated with a value).

Parenthesis (irrationality)

Humans are very bad dealing with probabilities

Experiment by D. Kahenan

Ask two different groups of people about how much they would pay for an insurance with a $100,000$ dollars premium,

1. in case of death
2. in case of death caused by a terrorist attack

The experiment shows that people pay more money for the second one.

A lot of people take advantage of this (gambling houses, insurance companies,...)

Conditional probabilities

Conditional prob. distributions

Let A and B two random variables taking values in $\{a_1, a_2, \ldots, a_m\}$ and $\{b_1, b_2, \ldots, b_{m'}\}$, respectively. Their **conditional probability** $P(A|B) = \{P(a_i|b_j)|\ 1 \leq i \leq m, 1 \leq j \leq m'\}$.
The value $P(a_i|b_j)$ represents the probability of the event $A = a_i$ given that $B = b_j$ is known to be true.
Each $P(A|b_j)$ is a prob. distribuion. So must satisfy that:

- for all a_i, b_j, $0 \leq P(a_i|b_j) \leq 1$
- for all b_j, $\sum_{i=1}^{m} P(a_i|b_j) = 1$

- A conditional probability distribution $P(A|B)$ gives us a quantitative way to represent how B provides information about A.
- It can be generalized to any number of variables $P(A, B|C, D)$

Conditional Probabilities: Example

- $B \in \{b, \neg b\}$, Barça wins the league
- $A \in \{a, \neg a\}$, Pedri is injured
- $C \in \{c, \neg c\}$, Courtois is injured
- $J \in \{j, \neg j\}$, I am injured

$P(b) = .5,\ P(b|a) = .3,\ P(b|c) = .6,\ P(b|a, c) = .4,\ P(b|j) = .5$

From Join Probabilities to Conditional Probabilities and Vice-versa

How to compute a conditional probability $P(a|b)$,

Property

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

Which can also be written as,

Property

$$P(a,b) = P(a|b)P(b)$$

Wraping up Example

Consider the outcome of **rolling a dice**. Let A be a variable representing the **outcome**, thus taking values in $\{1, 2, \cdots, 6\}$. Let B be a variable representing whether the outcome is going to be an **even number**, thus taking values $\{even, odd\}$.

It is reasonable to associate to these variables the following distributions:

- $P(A) = \{P(A = i) = 1/6 | \ 1 \leq i \leq 6\}$
- $P(B) = \{P(B = even) = .5, P(B = odd) = .5\}$

Wraping up Example

The conditional distribution $P(A|B = even)$ is,

- $P(A = 1|B = even) = 0$
- $P(A = 2|B = even) = 1/3$
- $P(A = 3|B = even) = 0$
- $P(A = 4|B = even) = 1/3$
- $P(A = 5|B = even) = 0$
- $P(A = 6|B = even) = 1/3$

Wraping up Example

The conditional distribution $P(B|A = 1)$ is,

- $P(B = even|A = 1) = 0$
- $P(B = odd|A = 1) = 1$

Wraping up Example

The join distribution $P(A, B)$ is,

- $P(A = 1, B = even) = 0$
- $P(A = 2, B = even) = 1/6$
- $P(A = 3, B = even) = 0$
- ...
- $P(A = 1, B = odd) = 1/6$
- $P(A = 2, B = odd) = 0$
- $P(A = 3, B = odd) = 1/6$
- ...

Remember: Marginals and Conditionals

Law of Total Probability (marginalization)

$$P(a) = \sum_b P(a, b)$$

Property

$$P(a|b) = \frac{P(a, b)}{P(b)}$$

Check the formulas on the rolling dice example

Some new formulas (not really)

Chain Rule

$$P(a,b,c,d) = P(a|b,c,d)P(b|c,d)P(c|d)P(d)$$

$$P(a,b,c,d) = P(a|b,c,d)P(b,c,d) = P(a|b,c,d)P(b|c,d)P(c,d) =$$

$$= P(a|b,c,d)P(b|c,d)P(c|d)P(d)$$

Some new formulas (not really)

Marginalization

$$P(a) = \sum_{b} P(a|b)P(b)$$

Because $P(a, b) = P(a|b)P(b)$
It can be generalized (with the chain rule) to,

$$P(a, b) = \sum_{c,d} P(a, b, c, d) = \sum_{c,d} P(a, b|c, d)P(c, d)$$

Or even to,

$$P(a) = \sum_{b,c,d} P(a, b, c, d) = \sum_{b,c,d} P(a|b, c, d)P(b|c, d)P(c|d)P(d)$$

Typical Queries

Consider a set of variables $X = \{X_1, X_2, ..., X_n\}$ and a probabilistic model of how they behave, modeled as a probability distribution over a set of variables $P(X_1, X_2, \cdots, X_n)$

- **Prior Marginals**: Our beliefs about the variables when we have no information
$$P(X_i)$$
- **Posterior Marginals**: Our updated beliefs after getting some evidence e (where $E \subset X$)
$$P(X_i|e)$$
- **Maximum A Posteriori Hypothesis (MAP)**: The most probable explanation w.r.t. $Y \subset X$ of what we have observed.
$$y^* = argmax_y P(y|e)$$

Typical Queries

Consider a system modeled as a probability distribution over a set of variables $P(X_1, X_2, \cdots, X_n)$

- **Prior Marginals**: Our beliefs about the variables when we have no information

$$P(X_i) = \sum_{x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n} P(x_1, x_2, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)$$

- **Posterior Marginals**: Our updated beliefs after getting some evidence e

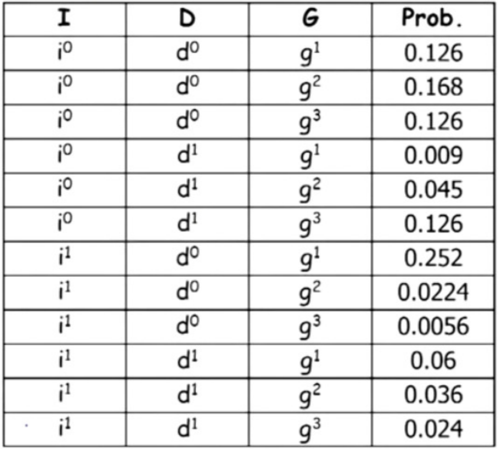
$$P(X_i|e) = \frac{P(X_i, e)}{P(e)}$$

- **Maximum A Posteriori Hypothesis (MAP)**: The most probable explanation of what we have observed.

$$y^* = argmax_y P(y|e) = argmax_y \frac{P(y, e)}{P(e)} = argmax_y P(y, e)$$

Example: Intelligence, Difficulty, Grade

Supose that FIB gives us data about grades (good g^1, average g^2, fail g^3) obtained by students (smart i^1 and not-so-smart i^0) in courses (easy d^0 and hard d^1)



I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Example: Intelligence, Difficulty, Grade

Compute:

- $P(i^0) = \sum_d \sum_g P(i^0, d, g) = 0.6$
- $P(I) = (P(i^0), P(i^1)) = (0.6, 0.4)$
- $P(G) = (P(g^1), P(g^2), P(g^3)) = (0.447, 0.2714, 0.2816)$
- $P(I, G) =$ $(P(i^0, g^1), P(i^0, g^2), P(i^0, g^3), P(i^1, g^1), P(i^1, g^2), P(i^1, g^3))$
- $P(i^0|g^1) = \frac{P(i^0,g^1)}{P(g^1)} = \frac{0.135}{0.447} = 0.302$
- $P(i^0|g^1, d^1) = \frac{P(i^0,g^1,d^1)}{P(g^1,d^1)} = \frac{0.009}{0.069} = 0.13$
- $P(I|g^1, d^1) = (P(i^0|g^1, d^1), P(i^1|g^1, d^1)) = (0.13, 0.87)$
- $MAP(I|g^1, d^1) = argmax_i P(i|g^1, d^1) = i^1$

Typical Queries (example)

Consider words of length 3 in English. Let X_i be their $i-th$ letter.
By going through our library we can compute $P(X_1, X_2, X_3)$

$X_1X_2X_3$	$P(X_1, X_2, X_3)$
cat	.20
bat	.05
rat	.25
red	.21
bed	.10
bet	.19

$X_1 \in \{c, b, r\}, X_2 \in \{a, e\}, X_3 \in \{t, d\}$

Typical Queries (example cont.)

- **Prior Marginals** Probability distr. of the first letter

$$P(X_1) = \sum_{x_2, x_3} P(X_1, X_2 = x_2, X_3 = x_3)$$

X_1	$P(X_1)$
c	.20
b	.34
r	.46

Typical Queries (example cont.)

- **Posterior Marginals** Probability distr. of the first letter given that the second letter is an *e*

$$P(X_1|X_2 = e) = \frac{P(X_1, X_2 = e)}{P(X_2 = e)}$$

X_1	$P(X_1\|X_2 = e)$
c	.0
b	.58
r	.42

Typical Queries (example cont.)

- $MAP(X_1|X_2 = e)$

$$x_1^* = argmax_{x_1} P(X_1 = x_1|X_2 = e) = b$$

The most probable first letter given the evidence is $x_1^* = b$

- $MAP(X_3|X_2 = e)$

$$x_3^* = argmax_{x_1} P(X_3 = x_3|X_2 = e) = d$$

The most probable third letter given the evidence is $x_3^* = d$

Typical Queries (example cont.)

- $MAP(X1, X_3 | X_2 = e)$

$$(x_1, x_3)^* = argmax_{x_1,x_3} P(X_1 = x_1, X_3 = x_3 | X_2 = e)$$

$$(x_1, x_3)^* = argmax\{P(X_1 = c, X_3 = t | X_2 = e), P(X_1 = c, X_3 = d | X_2 =$$

X_1	X_3	$P(X_1, X_3 \| X_2 = e)$
r	d	.42
b	d	.20
b	t	.38

The most probable combination of first and third letters given the evidence is $(x_1, x_3)^* = (r, d)$

Queries (example)

Consider a medical database of **symptoms** and **diseases**. For simplicity, we will assume two boolean symptoms (Temperature, Headache) and two diseases (e.g. Flu, Cold). By processing the data base of 1000 cases we get the following probabilities,

T	H	F	C	$P(T, H, F, C)$
0	0	0	0	.700
0	0	0	1	.002
0	0	1	0	.007
1	1	0	1	.070
1	1	0	0	.005
1	0	1	0	.135
1	0	0	0	.019
0	1	0	0	.060
0	1	0	1	.025

Queries (example)

- **Prior Marginals** Probability of a random patient having Flu

$$P(F) =$$

- **Posterior Marginals** Probability of a patient having flu given that he has headache

$$P(F|H = 1) =$$

- **Posterior Marginals** Probability of a patient having flu given that he has headache and does not have high temperature

$$P(F|H = 1, T = 0) =$$

Queries (example)

- $MAP(F|H=1)$ (Does a patient with headache have flu?)

$$(f)^* = argmax_f P(f|H=1) =$$

- $MAP(C|H=1)$ (Does a patient with headache have a cold?)

$$(c)^* = argmax_c P(c|H=1) =$$

- $MAP(F,C|H=1)$ (Why a patient has headache and high temperature?)

$$(f,c)^* = argmax_{f,c} P(f,c|H=1,T=1) =$$

Event Independence

Two events are **independent** if knowing one does not affect our believes about the other.

Independence

Two events a and b are **independent**, noted $a \perp b$, iff $P(a, b) = P(a) \cdot P(b)$.

Alternative equivalent definition is that $P(a|b) = P(a)$ (or, equivalently that $P(b|a) = P(b)$).

Can you see any independence in slide 11 (injuries in Barça)? (yes, $b \perp j$)
Can you see any independence in slides 12-15 (rolling dice)? (no)

Variable Independence

Two **variables** are **independent** if knowing one (no matter which value) does not affect our believes about the other.

Variable Independence

two variables A and B are **independent**, noted $A \perp B$, iff forall a_i, b_j, $P(a_i, b_j) = P(a_i) \cdot P(b_j)$.

Alternative Definition

two variables A and B are **independent**, noted $A \perp B$, iff forall a_i, b_j, $P(a_i | b_j) = P(a_i)$.

Variable Independence is stronger than event independence

Example

A	B	C	P(A,B,C)
0	0	0	.162
0	0	1	.072
0	1	0	.030
0	1	1	.036
1	0	0	.378
1	0	1	.288
1	1	0	.030
1	1	1	.004

- Check that $B \perp C$
  - $P(B) = (0.9, 0.1)$
  - $P(C) = (0.6, 0.4)$
  - $P(B, C) =$

Variable Independence means factorization

Variable Independence

If $A \perp B$, then $P(A, B) = P(A) \cdot P(B)$.

Variable Independence tells me that it is possible to **factorize** one big table $O(m \cdot m')$ into two smaller tables $O(m + m')$.

Example

Exercise:

Consider a robot manufacturer. Currently, he has sold $102,000$ units such that the 2 years warranty has expired. $30,000$ units have, at some point during the warranty, broken down.

Going deeper into the details, it turns out that the manufacturer has three different factories and each factory makes one third of the robots. The number of robots from factory 1, 2 and 3 that have broken down is $5,000$, $10,000$ and $15,000$, respectively.

Consider variables $B = \{b, \neg b\}$a robot breaks within the 2 years warranty, and $F = \{f_1, f_2, f_3\}$ factory where a robot has been made.

1. Identify independencies

Example (computing the data)

F	P(F)
f_1	1/3
f_2	1/3
f_3	1/3

B	F	$P(B\|F)$
$\neg b$	f_1	29/34
$\neg b$	f_2	24/34
$\neg b$	f_3	19/34
b	f_1	5/34
b	f_2	10/34
b	f_3	15/34

B	P(B)
$\neg b$	72/102
b	30/102

Example (independencies)

Check that:

- $f_2 \perp b$
- $f_2 \perp \neg b$
- Is not true that $B \perp F$

Conditional Independence

Two events are **independent given a** third, if knowing one does not affect our believes about the other as long as we already know the third.

Conditional Independence

two events a and b are independent given c, noted $a \perp b | c$, iff $P(a, b|c) = P(a|c) \cdot P(b|c)$.

Alternative equivalent definition is that $P(a|b, c) = P(a|c)$ (or, equivalently that $P(b|a, c) = P(b|c)$).

Variable Conditional Independence

Two **variables** are **independent given a** third, if knowing one does not affect our believes about the other as long as we already know the third.

Conditional Independence

two variables A and B are independent given variable C, noted $A \perp B | C$, iff forall a, b, c, $P(a, b|c) = P(a|c) \cdot P(b|c)$.

Variable Conditional Independence is stronger than event conditional independence

Example

A	B	C	P(A,B,C)
0	0	0	.162
0	0	1	.378
0	1	0	.108
0	1	1	.252
1	0	0	.056
1	0	1	.024
1	1	0	.014
1	1	1	.006

- Note that $B \perp C | A$
  - $P(B|A)$
  - $P(C|A)$
  - $P(B,C|A)$

Independence

Property

- $a \perp b$ does not imply $a \perp b | c$
- $a \perp b | c$ does not imply $a \perp b$

Property

- $A \perp B$ does not imply $A \perp B | C$
- $A \perp B | C$ does not imply $A \perp B$

You can check the property in the two previous examples

Bayes Rule

Bayes' rule expresses how we can reverse our knowledge

Bayes Rule

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

The proof of this equation is direct. Since,

- $P(a,b) = P(a|b)P(b)$
- $P(a,b) = P(b|a)P(a)$

we know that,

$$P(a|b)P(b) = P(b|a)P(a)$$

Bayes Rule. Example Robots

Consider a robot manufacturer. Currently, he has sold $102,000$ units such that the 2 years warranty has expired. $30,000$ units have, at some point during the warranty, broken down.

Going deeper into the details, it turns out that the manufacturer has three different factories and each factory makes one third of the robots. The number of robots from factory 1, 2 and 3 that have broken down is $5,000$, $10,000$ and $15,000$, respectively.

Example

Some data is readily available:

F	P(F)
f_1	1/3
f_2	1/3
f_3	1/3

B	F	$P(B\|F)$
$\neg b$	f_1	29/34
$\neg b$	f_2	24/34
$\neg b$	f_3	19/34
b	f_1	5/34
b	f_2	10/34
b	f_3	15/34

But $P(F|\ B)$ is not so obvious

But using Bayes Rule: $P(F|B) = \frac{P(B|F)P(F)}{P(B)}$

Example

Let's compute: $P(f_1|b) = \frac{P(b|f_1)P(f_1)}{P(b)}$

- We already have $P(b|f_1)$ and $P(f_1)$

$$P(b|f_1)P(f_1) = \frac{5}{34}\frac{1}{3}$$

- We now that $P(b) = 30/102$
- Therefore, $P(f_1|b) = \frac{5}{102}/\frac{30}{102} = 1/6$

Bayes Rule: example medical diagnosis

Example: The probability of a random person having cancer is 0.01. To diagnose Cancer (C) we have a test (T) which gives false positives with probability 0.2 and false negatives with probability 0.1
We would like to know what is the probability of a patient having cancer if the test was positive ($P(c|t)$) and if the test was negative ($P(c|\neg t)$)

Example

Some data is readily available:

C	P(C)
c	0.01
$\neg c$	0.99

T	C	$P(T\|C)$
$\neg t$	$\neg c$	0.8
$\neg t$	c	0.1
t	$\neg c$	0.2
t	c	0.9

But $P(C|\ T)$ is not so obvious

Using Bayes Rule: $P(C|T) = \frac{P(T|C)P(C)}{P(T)}$

Example

Let's compute: $P(c|t) = \frac{P(t|c)P(c)}{P(t)}$

- We already have $P(t|c)$ and $P(c)$

$$P(t|c)P(c) = 0.9 \cdot 0.01 = 0.009$$
$$P(t|\neg c)P(\neg c) = 0.2 \cdot 0.99 = 0.198$$

- We need $P(t) =$ (law of total probability)

$$P(t) = P(t|c)P(c) + P(t|\neg c)P(\neg c) = 0.009 + 0.198 = 0.207$$

- Therefore,

$$P(c|t) = 0.009/0.207 = 0.04$$

Assignment: compute $P(c|\neg t)$

Bayes Rule: example traffic jam

We know that rain influences traffic jams in Barcelona. We know that the probability of rain is 0.05, the probability of traffic jam if it rains is 0.6 and if it does not rain 0.25

We wake up in the morning, turn on the radio and hear that there is a traffic jam. What is the probability of rain?