

Advanced Statistical Inference

Refresher on Probabilities

Maurizio Filippone
`Maurizio.Filippone@eurecom.fr`

Department of Data Science
EURECOM

Outline

Syntax of Probabilities

Sum Rule

Product Rule

Bayes' Rule

Random variables

A random variable [...] refers to a “part” of the world whose “status” is initially unknown. [...]

S.Russell, P.Norvig, “Artificial Intelligence. A Modern Approach”,
Prentice Hall (2003)

Boolean random variables

- ▶ Propositional or **Boolean** random variables
- ▶ Two possible values: *True* or *False*
- ▶ Examples:
 - ▶ *Cavity* (do I have a cavity?)
 - ▶ *Earthquake* (is there an earthquake?)
 - ▶ $\neg \text{Earthquake} \vee \text{Cavity}$

Multivalued variables

- ▶ Discrete or **Multivalued** random variables
- ▶ Values must be exhaustive and mutually exclusive
- ▶ Examples:
 - ▶ *Weather* is one of (*sunny*, *rainy*, *cloudy*, *snowy*)
 - ▶ *Face* of a die roll is one of (1, 2, 3, 4, 5, 6)

Prior Probabilities

- ▶ **Prior** or **unconditional probabilities** of propositions, e.g.,
 - ▶ $P(\text{Cavity} = T) = 0.1$ (also denoted by $P(\text{Cavity})$)
 - ▶ $P(\text{Weather} = \text{sunny}) = 0.72$
- ▶ Correspond to belief prior to arrival of any (new) evidence

Probability distribution

- ▶ **Probability distribution** gives values for all possible assignments:
 - ▶ $\mathbf{P}(\textit{Weather}) = (0.72, 0.1, 0.08, 0.1)$
 - ▶ Array of numbers with one index: $array[index]$
- ▶ **Normalized**, i.e., sums to 1

$$P(\textit{sunny} \vee \textit{rainy} \vee \textit{cloudy} \vee \textit{snowy}) = 1$$

Multiple variables - Joint distribution

- ▶ **Joint probability distribution** for a set of variables gives values for each possible assignment to all the variables
 - ▶ $\mathbf{P}(\textit{Cavity}, \textit{Weather}) =$ a 2×4 array
 - ▶ Array of numbers with multiple indices:
 $array[index1, index2, \dots]$

Multiple variables - Joint distribution

- Sum of all values is 1

		<i>Weather</i>			
		<i>sunny</i>	<i>rainy</i>	<i>cloudy</i>	<i>snowy</i>
Cavity	<i>T</i>	0.10	0.20	0.05	0.05
	<i>F</i>	0.25	0.15	0.02	0.18

Continuous variables

- ▶ What about continuous variables?
- ▶ How do we specify an array of infinite numbers?

Answer: **Probability Density Function**

Discrete vs continuous variables

- ▶ If X takes values in D discrete
- ▶ Sum expressed by the symbol \sum

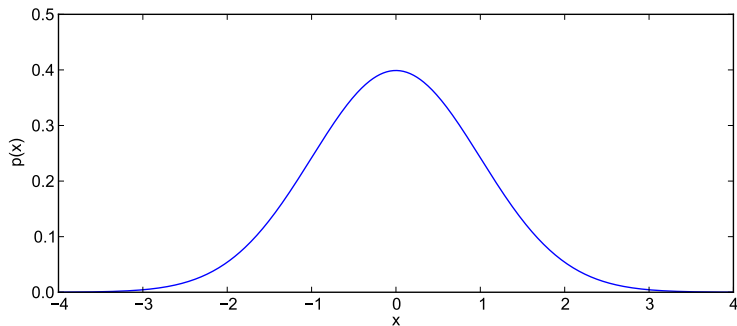
$$\sum_{v \in D} \mathbf{P}(X = v) = 1$$

- ▶ If X takes values in C continuous
- ▶ Sum expressed by the symbol \int

$$\int_C p(x) dx = 1$$

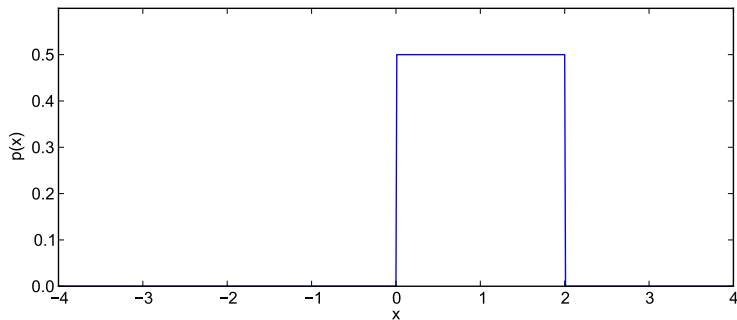
Probability density function - Examples

Gaussian probability density function



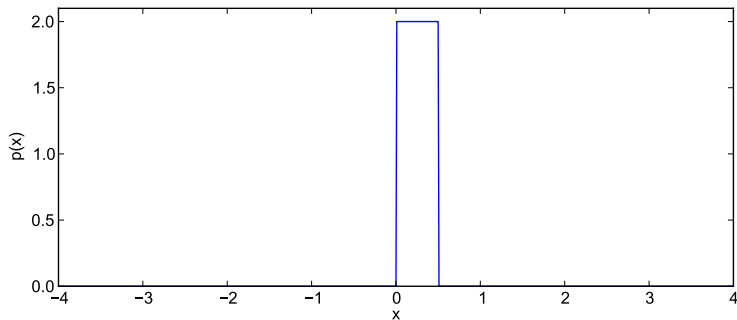
Probability density function - Examples

Uniform density function



Probability density function - Examples

Uniform density function - look at the y -axis!



Probability density vs distribution

- ▶ Are probabilities not supposed to be between zero and one?

Yes

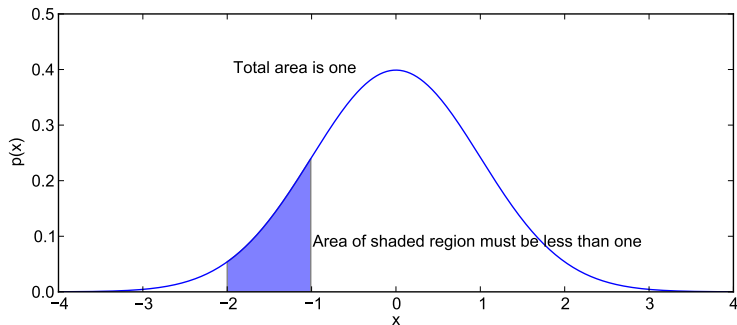
- ▶ Is there anything wrong with the last plot? No
- ▶ The probability of an event being in a given interval I :

$$0 \leq \int_I p(x) dx \leq 1$$

Why?

Probability density vs distribution

Let's see for the Gaussian



Conditional probabilities

- ▶ **Conditional** or **posterior** probabilities
- ▶ Denoted by the “|” symbol
- ▶ Belief updated **after** evidence is gathered
 - ▶ $P(\textit{Cavity}|\textit{Toothache}) = 0.8$, i.e., given that *Toothache* is all I know
 - ▶ $\mathbf{P}(\textit{Weather}|\textit{Earthquake})$ = two arrays of 4 elements (one for each state of *Earthquake*)

Conditional probabilities

- ▶ New evidence may be irrelevant, allowing simplification, e.g.,
 - ▶ $P(\text{Cavity} | \text{Toothache}, \text{LotteryDraw}) = P(\text{Cavity} | \text{Toothache}) = 0.8$
- ▶ This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probabilities

Define:

Action A_t : leave for airport t minutes before flight

Question:

Will A_t get me there on time?

Conditional probabilities

- ▶ A_{25} will get me to the airport on time with probability 0.04:

$$P(A_{25}) = 0.04$$

- ▶ Probabilities change with new **evidence**

$$P(A_{25} | \text{no reported accidents}) = 0.06$$

$$P(A_{25} | \text{no reported accidents, 5 a.m.}) = 0.15$$

Making decisions under uncertainty

- ▶ Suppose I believe the following:

$$P(A_{25} \text{ gets me there on time} | \dots) = 0.04$$

$$P(A_{90} \text{ gets me there on time} | \dots) = 0.70$$

$$P(A_{120} \text{ gets me there on time} | \dots) = 0.95$$

$$P(A_{1440} \text{ gets me there on time} | \dots) = 0.9999$$

- ▶ Which action to choose?

Making decisions under uncertainty

- ▶ Depends on my **preferences** for missing flight vs. airport cuisine, ...
- ▶ **Utility theory** is used to represent and infer preferences

Decision theory = utility theory + probability theory

- ▶ We are not covering this in ASI...

From prior to posterior probabilities

Question:

How can we **update** our belief about a random variable?

Answer:

Bayesian inference

Outline

Syntax of Probabilities

Sum Rule

Product Rule

Bayes' Rule

Normalization of probabilities

- ▶ Possible outcomes of a random variable are mutually exclusive
- ▶ For example, in the case of the roll of a die

$$P(\text{Face} = 1 \wedge \text{Face} = 2) = 0$$

- ▶ Possible outcomes are exhaustive:

$$\text{Face} = 1 \vee \dots \vee \text{Face} = 6 \quad \text{is true}$$

$$\text{hence} \quad \sum_i P(\text{Face} = i) = 1$$

Normalization for joint distributions

Example:

- ▶ Suppose that $\mathbf{P}(\text{Cavity}, \text{Toothache})$ is:

		<i>Toothache</i>	
		<i>T</i>	<i>F</i>
<i>Cavity</i>	<i>T</i>	0.04	0.06
	<i>F</i>	0.01	0.89

- ▶ Again, sum over all possible outcomes is 1

Marginal distributions

- ▶ Given a joint distribution over a set of variables we can compute the distribution for a subset of variables
- ▶ This is sometimes called **marginal** distribution

Marginal distributions

- ▶ Suppose we are interested in $P(\text{Cavity} = T)$
- ▶ Can we compute it from the joint distribution?

		<i>Toothache</i>	
		<i>T</i>	<i>F</i>
<i>Cavity</i>	<i>T</i>	0.04	0.06
	<i>F</i>	0.01	0.89

- ▶ Yes! - How?

Marginal distributions

- ▶ Sum the row corresponding to $Cavity = T$

		<i>Toothache</i>			
		<i>T</i>		<i>F</i>	
<i>Cavity</i>	<i>T</i>	0.04	+	0.06	= 0.10
	<i>F</i>	0.01		0.89	

- ▶ Therefore $P(Cavity = T) = 0.1$
- ▶ Why?

Marginal distributions

- ▶ We are interested in $P(\text{Cavity} = T)$ **regardless** of any evidence about *Toothache*
- ▶ Sum across all possible states of *Toothache* ensures this

Outline

Syntax of Probabilities

Sum Rule

Product Rule

Bayes' Rule

Conditional probability

- ▶ Definition of conditional probability:

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad \text{if} \quad P(B) \neq 0$$

- ▶ **Why?** It's about normalization

Conditional probability

Example:

- ▶ Suppose we are interested in

$$P(\text{Cavity} = T \mid \text{Toothache} = T)$$

		<i>Toothache</i>	
		<i>T</i>	<i>F</i>
<i>Cavity</i>	<i>T</i>	0.04	0.06
	<i>F</i>	0.01	0.89

Conditional probability

- ▶ We need to ensure that

$$\sum_{Cavity=T,F} \mathbf{P}(Cavity | Toothache = T) = 1$$

Conditional probability

- ▶ In this example:

$$P(\textit{Toothache} = T) = 0.05$$

$$\mathbf{P}(\textit{Cavity}, \textit{Toothache} = T) = (0.04, 0.01)$$

- ▶ Applying the definition of conditional probability

$$P(\textit{Cavity} = T | \textit{Toothache} = T) = \frac{0.04}{0.05} = 0.8$$

Product rule

- ▶ The definition of conditional probability

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad \text{if} \quad P(B) \neq 0$$

- ▶ Implies the so called **product rule**:

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

- ▶ Easy to remember: product turns “|” into “,”

Conditional probability

- ▶ General version holds for whole distributions

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

Viewed as a 4×2 set of equations

Conditional probability - Chain Rule

- ▶ The same idea applies to multiple variables

$$P(A, B, C) = P(A, B|C)P(C)$$

$$P(A, B|C) = P(A|B, C)P(B|C)$$

Conditional probability - Chain Rule

- **Chain rule** is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \dots, X_n) &= \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \mathbf{P}(X_1, \dots, X_{n-1}) \\ &= \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \times \\ &\quad \mathbf{P}(X_{n-1} | X_1, \dots, X_{n-2}) \mathbf{P}(X_1, \dots, X_{n-2}) \\ &= \dots \\ &= \prod_{i=1}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1})\end{aligned}$$

Outline

Syntax of Probabilities

Sum Rule

Product Rule

Bayes' Rule

Bayes' Rule

- ▶ Product rule

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

- ▶ Implies **Bayes' rule**

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

- ▶ Why is this useful?

Bayes' Rule

- For assessing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Bayes' Rule

Example

- ▶ Let M be meningitis, S be stiff neck:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

- ▶ Note: posterior probability of meningitis still very small!

Bayes' Rule - Normalization

- ▶ Suppose we wish to compute a posterior distribution over A given $B = b$, and suppose A has possible values $a_1 \dots a_m$
- ▶ We can apply Bayes' rule for each value of A

$$P(A = a_1 | B = b) = \frac{P(B = b | A = a_1)P(A = a_1)}{P(B = b)}$$

...

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

Bayes' Rule - Normalization

- ▶ Adding these up, and noting that:

$$\sum_i P(A = a_i | B = b) = 1,$$

we obtain:

$$\frac{1}{P(B = b)} = \frac{1}{\sum_i P(B = b | A = a_i) P(A = a_i)}$$

- ▶ This is the **normalization factor**, constant wrt i , denoted α :

$$\mathbf{P}(A | B = b) = \alpha \mathbf{P}(B = b | A) \mathbf{P}(A)$$

Bayes' Rule - Normalization

- ▶ Typically compute an unnormalized distribution, normalize at end
- ▶ For example, suppose

$$\mathbf{P}(B = b|A)\mathbf{P}(A) = (0.4, 0.2, 0.2)$$

then

$$\begin{aligned}\mathbf{P}(A|B = b) &= \alpha(0.4, 0.2, 0.2) \\ &= \frac{1}{0.4 + 0.2 + 0.2}(0.4, 0.2, 0.2) \\ &= (0.5, 0.25, 0.25)\end{aligned}$$