Advanced Statistical Inference
Refresher on Probabilities

Motonobu Kanagawa
motonobu.kanagawa@eurecom.fr

Data Science Department
EURECOM

February 18, 2024

Outline

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:
  - *Train* (my train departs on time?)
  - *Earthquake* (there is an earthquake?)
  - $\neg \mathit{Earthquake} \vee \mathit{Train}$

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 dice roll is one of $(1, 2, 3, 4, 5, 6)$

Prior Probabilities

- Prior or unconditional probabilities of propositions, e.g.,
  - $P(Train = True) = 0.9$ (also denoted by $P(Train)$)
  - $P(Weather = sunny) = 0.72$
- Correspond to belief prior to arrival of any (new) evidence

Probability distribution

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

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

Multiple variables - Joint distribution

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

Multiple variables - Joint distribution

- Sum of all values is 1

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	T	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

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 $\sum$

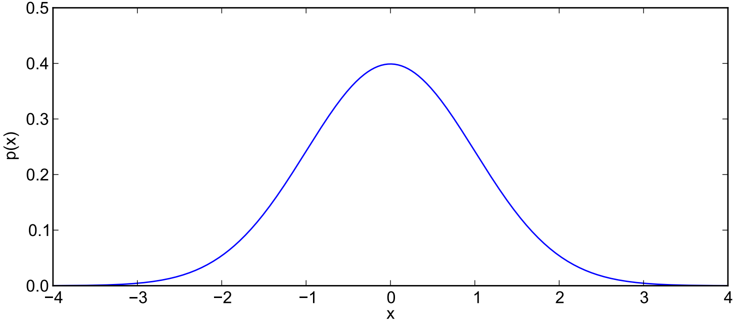
$$\sum_{v \in D} \mathbf{P}(X = v) = 1$$

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

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

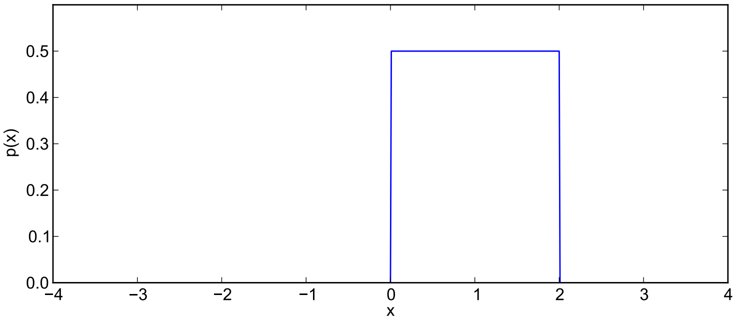
Probability density function - Examples

Gaussian probability density function



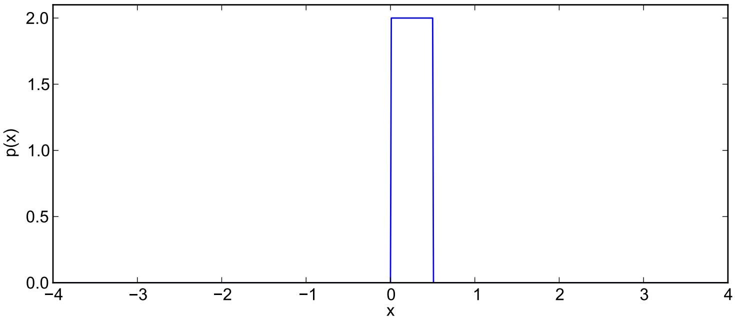
Probability density function - Examples

Uniform density function



Probability density function - Examples

Uniform density function - look at the y-axis!



Probability density vs distribution

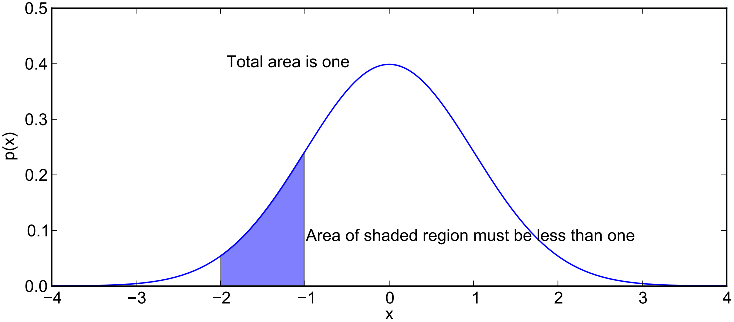
- Are probabilities not supposed to be between zero and one? Yes
- Is there anything wrong with the last plot? No
- The probability of an event being in a given interval I:

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

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. For example:
  - $P(Train = True | Weather = Rainy) = 0.5$.
  - $P(Train = True | Weather = Snowy) = 0.2$.

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	*T*	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

Conditional probabilities

- New evidence may be irrelevant, allowing simplification, e.g.,

$$
\begin{aligned}
& P(Train = True | Weather = Rainy, LotteryDraw) \\
= & P(Train = True | Weather = Rainy) = 0.5
\end{aligned}
$$

- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probabilities

Define:
Action A_t: leave home for airport t minutes before my 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

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

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

Making decisions under uncertainty

- Suppose I believe the following:

$$
\begin{aligned}
P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999
\end{aligned}
$$

- Which action to choose?

Making decisions under uncertainty

- Depends on my preferences for missing my 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 dice

$$P(Face = 1 \wedge Face = 2) = 0$$

- Possible outcomes are exhaustive:

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

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

Normalization for joint distributions

Example:

- Suppose that $\mathbf{P}(Train, Weather)$ is:

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	*T*	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

- Again, the 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 $P(Train = T)$
- Can we compute it from the joint distribution?

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	T	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

- Yes! - How?

Marginal distributions

- Sum the row corresponding to $Train = T$

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	T	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

- Therefore $P(Train = T) = 0.45 + 0.10 + 0.20 + 0.01 = 0.76$.
- Why?

Marginal distributions

- We are interested in $P(Train = T)$ regardless of any evidence about *Weather*
- Sum across all possible states of *Weather* ensures this

Outline

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

$$P(Train = T | Weather = Cloudy)$$

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	*T*	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

Conditional probability

- We need to ensure that

$$\sum_{Train=T,F} \mathbf{P}(Train|Weather = cloudy) = 1$$

Conditional probability

- In this example:

$$P(Weather = cloudy) = 0.20 + 0.05 = 0.25$$

$$\mathbf{P}(Train, Weather = cloudy) = (0.20, 0.05)$$

- Applying the definition of conditional probability

$$P(Train = T|Weather = cloudy) = \frac{0.20}{0.25} = 0.8$$

		Weather			
		sunny	*rainy*	*cloudy*	*snowy*
Train	*T*	0.45	0.10	0.20	0.01
	F	0.05	0.10	0.05	0.04

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

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

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:

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

Outline

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

- Suppose *Covid* as a cause and *Fever* as an effect.
- Suppose $P(Fever|Covid) = 0.8$, $P(Covid) = 0.05$ and $P(Fever) = 0.1$.

$$P(Covid|Fever) = \frac{P(Fever|Covid)P(Covid)}{P(Fever)}$$
$$= \frac{0.8 \times 0.05}{0.1} = 0.4.$$

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

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

$$\dots$$

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

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

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