Machine Learning

Section 4: Bayesian networks

Stefan Harmeling

20. October 2021

Computational difficulties of probability theory

Computational difficulties of probability theory

The problem:

- The joint distribution of propositional variables $A, B, \ldots, Z$ has many free parameters.

$$
\begin{array}{rr}
[1] & p(A, B, \ldots, Z) = \ldots \\
[2] & p(\neg A, B, \ldots, Z) = \ldots \\
[3] & p(A, \neg B, \ldots, Z) = \ldots \\
\vdots & \\
[67108863] & p(\neg A, \neg B, \ldots, Z) = \ldots \\
[67108864] & p(\neg A, \neg B, \ldots, \neg Z) = 1 - \sum p(\ldots)
\end{array}
$$

- Requires a large memory and calculating $p(A)$ requires a lot of time.
- How can we specify the joint distribution with fewer numbers?
- Can we restrict how variables are relevant to each other.

An important note about notation

So far:

A represents a formula (or event):

$$p(A) = \text{probability that formula } A \text{ is true}$$
$$p(\neg A) = \text{probability that formula } \neg A \text{ is true}$$

From now on:

A is a (propositional) variable with values in $\{0, 1\}$, i.e. $p(A)$ is a function of two possible input values $A{=}1$ and $A{=}0$, i.e. with slightly unusual notation:

$$p(A{=}1) = \text{probability that proposition } A \text{ is true}$$
$$p(A{=}0) = \text{probability that proposition } A \text{ is false}$$

Stating that $p(A, B) = p(A)\, p(B)$ means:

$$p(A{=}1, B{=}1) = p(A{=}1)\, p(B{=}1)$$
$$p(A{=}1, B{=}0) = p(A{=}1)\, p(B{=}0)$$
$$p(A{=}0, B{=}1) = p(A{=}0)\, p(B{=}1)$$
$$p(A{=}0, B{=}0) = p(A{=}0)\, p(B{=}0)$$

Tracy, Jack and the wet grass (1) — joint prob.

from Barber 2012, 3.1.1

$$
\begin{aligned}
T &= \text{Tracey's grass is wet}\\
R &= \text{it rained last night}\\
S &= \text{Tracey's sprinkler was on last night}\\
J &= \text{grass of Tracey's neighbor Jack is wet}
\end{aligned}
$$

Joint probability

$$
\begin{aligned}
p(T, J, R, S) &= p(T, J, R|S)\, p(S)\\
&= p(T, J|R, S)\, p(R|S)\, p(S)\\
&= p(T|J, R, S)\, p(J|R, S)\, p(R|S)\, p(S)
\end{aligned}
$$

- apply three times product rule $p(A, B) = p(A|B)\, p(B)$

Tracy, Jack and the wet grass (2) — parameter counting

from Barber 2012, 3.1.1

$$\begin{aligned} T &= \text{Tracey's grass is wet} \\ R &= \text{it rained last night} \\ S &= \text{Tracey's sprinkler was on last night} \\ J &= \text{grass of Tracey's neighbor Jack is wet} \end{aligned}$$

Number of parameters of joint probability

$$p(T, R, S, J) = p(T|J, R, S)\; p(J|R, S)\; p(R|S)\; p(S)$$

- $p(T, R, S, J)$ requires 15 parameters.
- rewritten with product rule requires $8 + 4 + 2 + 1$ parameters.

Leave out irrelevant conditions (use domain knowledge)

$$p(T, J, R, S) = p(T|R, S)\; p(J|R)\; p(R)\; p(S)$$

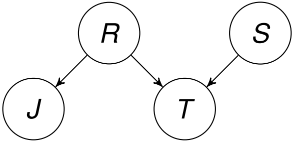
- only $4 + 2 + 1 + 1 = 8$ parameters!

Tracy, Jack and the wet grass (2) — representation

from Barber 2012, 3.1.1

$$p(T, J, R, S) = p(T|R, S)\ p(J|R)\ p(R)\ p(S)$$

Graphical representation



Conditional probability tables (CPTs)

$$p(R{=}1) = 0.2 \qquad p(S{=}1) = 0.1$$

$$p(J{=}1|R{=}1) = 1 \qquad p(J{=}1|R{=}0) = 0.2$$

$$p(T{=}1|R{=}1, S{=}0) = 1 \qquad p(T{=}1|R{=}1, S{=}1) = 1$$

$$p(T{=}1|R{=}0, S{=}1) = 0.9 \qquad p(T{=}1|R{=}0, S{=}0) = 0$$

Tracy, Jack and the wet grass (3) — inference

from Barber 2012, 3.1.1

Inference

- What is the probability that the sprinkler was on given that we observe that Tracey's grass is wet?

$$p(S{=}1|T{=}1) = \frac{p(S{=}1, T{=}1)}{p(T{=}1)} = \frac{\sum_{J,R} p(T{=}1, J, R, S{=}1)}{\sum_{J,R,S} p(T{=}1, J, R, S)}$$
$$= \ldots = 0.3382$$

- What is the probability that the sprinkler was on given that we observe that Tracey's and Jack's grass is wet?

$$p(S{=}1|T{=}1, J{=}1) = \frac{p(S{=}1, T{=}1, J{=}1)}{p(T{=}1, J{=}1)} = \frac{\sum_{R} p(T{=}1, J{=}1, R, S{=}1)}{\sum_{R,S} p(T{=}1, J{=}1, R, S)}$$
$$= \ldots = 0.1604$$

Jack's wet grass is *explaining away* the sprinkler as a reason for the wet grass of Tracey. Note: $S \perp\!\!\!\perp J$ but $S \not\perp\!\!\!\perp J \mid T$.

What is probabilistic reasoning?

Barber 2012, 1.2

1. identify all relevant variables, e.g. T, J, R, S
2. define joint probability $p(T, J, R, S)$
3. *evidence* fixes the values of certain variables, e.g. $T{=}1$
4. *inference* of the distribution of certain variables requires integrating out the rest, e.g. to calculate $p(S{=}1 | T{=}1)$

Bayesian networks aka Bayes nets, belief networks (1)

Typical definition from Barber 2012, 3.3 Belief networks; see also Pearl, 1988

Definition 4.1 (Bayesian network (version w/o explicit graph))

A Bayesian network is a distribution that can be written as

$$p(X_1, X_2, \dots, X_D) = \prod_{i=1}^{D} p(X_i | \mathrm{pa}(X_i))$$

Don't use this definition!

where $\mathrm{pa}(X)$ *are the parental variables of variable* X*. A Bayesian network can be represented as a Directed Acyclic Graph (DAG) with the propositional variables as nodes and arrows from parents to children.*

Problems of this definition:

- The graph is not unique! E.g.

$$p(X_1, X_2) = p(X_1)p(X_2|X_1) = p(X_2)p(X_1|X_2)$$

In both case p is a Bayesian network.

Bayesian networks aka Bayes nets, belief networks (2)

Compare Peters, Def 6.32 of causal graphical model

Better definition:

Definition 4.2 (Bayesian network)

A Bayesian network is a DAG $\mathcal{G}$ with vertices $X_1, \ldots, X_n$ and conditional probabilities $p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$ where $\mathrm{pa}_j^{\mathcal{G}}$ is the set of indices of the parents of X_j in $\mathcal{G}$ and $X_{\mathrm{pa}_j^{\mathcal{G}}}$ are the parent variables of X_j.
The $p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$ are also called conditional probability tables (CPTs).

Note that the conditional probabilities sum up to one in their first variable:

$$\sum_{X_j} p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}}) = 1$$

Note 4.3

A Bayesian network induces a joint distribution over $X_1, \ldots, X_n$:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$$

Bayesian networks aka Bayes nets, belief networks (3)

Compare Peters, Def 6.32 of causal graphical model

Note 4.4

The product rule for n variables

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j | x_1, \ldots, x_{j-1})$$

creates a factorization of the joint distribution for any variable ordering/permutation π*:*

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_{\pi(j)} | x_{\pi(1)}, \ldots, x_{\pi(j-1)})$$

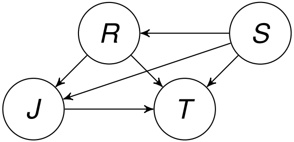
Thus any fully connected DAG together with any joint distribution forms a Bayesian network (which is not very interesting...).

E.g. ...

Bayesian networks aka Bayes nets, belief networks (3)

Without leaving out arrows it is also a Bayes net:

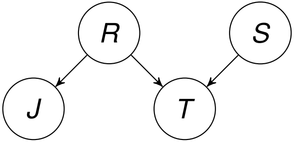
$$p(T, J, R, S) = p(T|J, R, S)\; p(J|R, S)\; p(R|S)\; p(S)$$



Thus any distribution can be written as a fully connected Bayes net for any variable ordering.

However, leaving out arrows if more efficient, but imposes constraints:

$$p(T, J, R, S) = p(T|R, S)\; p(J|R)\; p(R)\; p(S)$$



How can we characterize those constraints?

Measuring relevance between variables (1)

Definition 4.5 (independence)

Two variables A and B are independent, if and only if their joint distributions factorizes into so-called marginal distributions, i.e.

$$p(A, B) = p(A)\, p(B)$$

In that case $p(A|B) = p(A)$, which intuitively makes sense as well. Notation: $A \perp\!\!\!\perp B$. In words, information about B doesn't give information about A and vice versa.

Note that $p(R|S) = p(R)$ implies $p(R, S) = p(R)\, p(S)$.

Example:

- Two coins.

$$A = \text{coin 1 shows heads}$$
$$B = \text{coin 2 shows heads}$$

Then $A \perp\!\!\!\perp B$.

Measuring relevance between variables (2)

Definition 4.6 (conditional independence)

Two variables A and B are conditionally independent given variable C, if and only if their conditional distribution factorizes,

$$p(A, B|C) = p(A|C)\ p(B|C)$$

In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B doesn't tell us about A. Notation: $A \perp\!\!\!\perp B \mid C$

Example:

- Two coins and a bell.

$$A = \text{coin 1 shows heads}$$
$$B = \text{coin 2 shows heads}$$
$$C = \text{bell rings if both coins show the same result}$$

Then $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$,
but $A \not\perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp C \mid B$ and $B \not\perp\!\!\!\perp C \mid A$.

Measuring relevance between variables (3)

Definition 4.7 (conditional independence)

Two sets of variables $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given a set of variables $\mathcal{C}$, if and only if their conditional distribution factorizes,

$$p(\mathcal{A}, \mathcal{B} | \mathcal{C}) = p(\mathcal{A} | \mathcal{C}) \; p(\mathcal{B} | \mathcal{C})$$

where for $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, we define $p(\mathcal{A}) := p(A_1, A_2, \ldots, A_n)$. We write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$.

Note:

- The two previous definitions are special cases of the latter:

$$\begin{aligned} A \perp\!\!\!\perp B \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \\ A \perp\!\!\!\perp B \mid C \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \mid \{C\} \end{aligned}$$

Tracy, Jack and the wet grass — representation

from Barber 2012, 3.1.1

$$p(T, J, R, S) = p(T|R, S)\, p(J|R)\, p(R)\, p(S)$$

Conditional probability tables (CPTs)

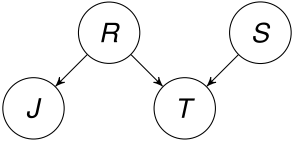
$$p(R{=}1) = 0.2 \qquad p(S{=}1) = 0.1$$

$$p(J{=}1|R{=}1) = 1 \qquad p(J{=}1|R{=}0) = 0.2$$

$$p(T{=}1|R{=}1, S{=}0) = 1 \qquad p(T{=}1|R{=}1, S{=}1) = 1$$

$$p(T{=}1|R{=}0, S{=}1) = 0.9 \qquad p(T{=}1|R{=}0, S{=}0) = 0$$

Graphical representation



What independencies can we infer only from the graph?

Conditional independencies in three variable networks

see also Barber 2012, 3.3.2

The four isolated paths in DAGs

(i)	$A \to B \to C$	$p(A,B,C) = p(C \mid B)\, p(B \mid A)\, p(A)$
(ii)	$A \leftarrow B \leftarrow C$	$p(A,B,C) = p(A \mid B)\, p(B \mid C)\, p(C)$
(iii)	$A \leftarrow B \to C$	$p(A,B,C) = p(A \mid B)\, p(C \mid B)\, p(B)$
(iv)	$A \to B \leftarrow C$	$p(A,B,C) = p(B \mid A, C)\, p(A)\, p(C)$

…imply the following independencies (with elementary proofs):

(i) $A \perp\!\!\!\perp C \mid B$ (ii) $A \perp\!\!\!\perp C \mid B$

(iii) $A \perp\!\!\!\perp C \mid B$ (iv) $A \perp\!\!\!\perp C$

However, they do not necessarily imply dependences, such as:

(i) $A \not\perp\!\!\!\perp C$ (ii) $A \not\perp\!\!\!\perp C$

(iii) $A \not\perp\!\!\!\perp C$ (iv) $A \not\perp\!\!\!\perp C \mid B$

Those might be true or wrong dependent on conditional probability tables.

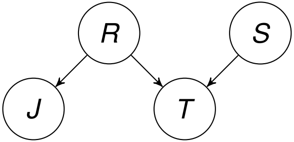
Tracy, Jack and the wet grass — cond. independencies

from Barber 2012, 3.1.1

Conditional independencies

(i) $A \rightarrow B \rightarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(ii) $A \leftarrow B \leftarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(iii) $A \leftarrow B \rightarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(iv) $A \rightarrow B \leftarrow C$	imply	$A \perp\!\!\!\perp C$

Graphical representation



What independencies can we infer only from the graph?

Answer: $J \perp\!\!\!\perp T \mid R$ and $R \perp\!\!\!\perp S$. But also $J \perp\!\!\!\perp S \mid R$, $J \perp\!\!\!\perp S$, $J \perp\!\!\!\perp S|R, T$ with the d-separation criterion (stay tuned).

A sophisticated criterion on graphs

copied from Peters, Def 6.1

Definition 4.8 (Pearl's d-separation)

Given a DAG $\mathcal{G}$.

1. *A path between nodes i_1 and i_m is **blocked by a set** S (with $i_1 \notin S$ and $i_m \notin S$), whenever there is a node i_k, such that one of the following two possibilities holds:*
   - *$i_k \in S$ and*

     $$\begin{aligned} & i_{k-1} \to i_k \to i_{k+1} \\ \text{or } & i_{k-1} \leftarrow i_k \to i_{k+1} \\ \text{or } & i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \end{aligned}$$

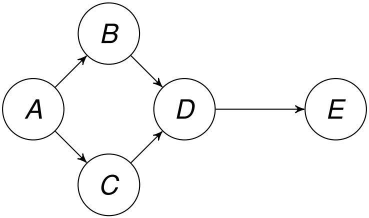
   - *neither i_k nor any of its descendents is in S and*

     $$i_{k-1} \to i_k \leftarrow i_{k+1}$$

2. *Two disjoint subsets of vertices A and B are **d-separated** by a third (also disjoint) subset S if every path between nodes in A and B is blocked by S. We write*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid S$$

Slightly more interesting example: diamont shape (1)

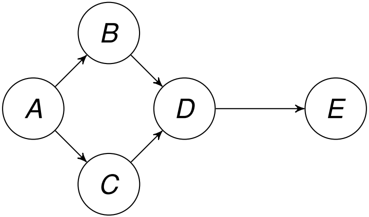


What independencies can we infer only from the graph?

Answers:

- $A \perp\!\!\!\perp D \mid B, C$: there are two paths from A to D: $A \to B \to D$ and $A \to C \to D$. The first is blocked by B, the second by C.
- $A \perp\!\!\!\perp E \mid B, C$: there are two paths ...
- $A \perp\!\!\!\perp E \mid D$: there are two paths ..., both are block by D.
- $B \perp\!\!\!\perp C \mid A$: there are two paths .... Note that D must not be observed, otherwise $B \to D \leftarrow C$ is open. Also E must not be observed (in def: "nor any of its descendents...").
- more: $A \perp\!\!\!\perp D \mid B, C, E$ and $A \perp\!\!\!\perp E \mid B, C, D$ and $C \perp\!\!\!\perp E \mid D$ (possibly with A and/or B), same for $B \perp\!\!\!\perp E \mid D$ ...

Slightly more interesting example: diamont shape (2)



What independencies can we infer only from the graph?

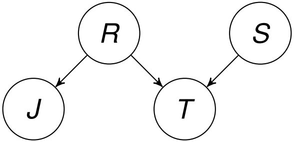
All answers: Let me know, if I missed one!

- $A \perp\!\!\!\perp D \mid B, C$ and $A \perp\!\!\!\perp D \mid B, C, E$
- $A \perp\!\!\!\perp E \mid B, C$ and $A \perp\!\!\!\perp E \mid B, C, D$ and $A \perp\!\!\!\perp E \mid D$ and $A \perp\!\!\!\perp E \mid B, D$ and $A \perp\!\!\!\perp E \mid C, D$
- $B \perp\!\!\!\perp C \mid A$
- $C \perp\!\!\!\perp E \mid D$ and $C \perp\!\!\!\perp E \mid D, A$ and $C \perp\!\!\!\perp E \mid D, B$ and $C \perp\!\!\!\perp E \mid D, A, B$
- $B \perp\!\!\!\perp E \mid D$ and $B \perp\!\!\!\perp E \mid D, A$ and $B \perp\!\!\!\perp E \mid D, B$ and $B \perp\!\!\!\perp E \mid D, A, B$

Tracy, Jack and the wet grass — cond. independencies

from Barber 2012, 3.1.1

Graphical representation



What independencies can we infer only from the graph?

Answer:

- $J \perp\!\!\!\perp T \mid R$: because the path from J to T is d-separated by observing R, so all paths between them are d-separated
- $R \perp\!\!\!\perp S$: because the path from R to S is d-separated, if we do not observe T, so all paths ...
- $J \perp\!\!\!\perp S \mid R$: because the path from J to S is d-separated by observing R, so all paths ...
- $J \perp\!\!\!\perp S$, because the path from J to S is d-separated by not observing T, so all paths ...
- $J \perp\!\!\!\perp S|R, T$, because the path from J to S is d-separated by observing R, so all paths ...

Linking graphs and distributions

Peters, Def 6.21

Definition 4.9

Given a DAG $\mathcal{G}$, a joint distribution p satisfies

1. *the **global Markov property** wrt. the DAG $\mathcal{G}$ if*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \implies A \perp\!\!\!\perp B \mid C$$

   for all disjoint vertex sets A, B and C and where $A \perp\!\!\!\perp B \mid C$ describes cond. ind. wrt. p.

2. *the **local Markov property** wrt. the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and*
3. *the **Markov factorization property** wrt. the DAG $\mathcal{G}$ if*

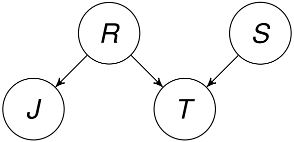
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_j \,|\, \mathrm{pa}_j^{\mathcal{G}})$$

Theorem 4.10 (Equivalence of Markov properties)

If some joint distribution has a density p then all Markov properties (from the previous def.) are equivalent.

Example

A distribution $p(R, S, T, J)$ is Markovian wrt to graph $\mathcal{G}$



if either (global Markov property)

$$J \perp\!\!\!\perp T \mid R$$
$$R \perp\!\!\!\perp S$$
$$J \perp\!\!\!\perp S \mid R$$
$$J \perp\!\!\!\perp S$$
$$J \perp\!\!\!\perp S|R, T$$

or if (Markov factorization property)

$$p(T, J, R, S) = p(T|R, S)\, p(J|R)\, p(R)\, p(S)$$

Summary

- A joint distribution, such as $p(A, B, C, \ldots, Z)$ requires lots of parameters, thus lots of memory.
- Exploit conditional independencies between variables.
- Factorize the joint distribution along a graph.
- There is a (somewhat complicated) criterion on graphs which corresponds to conditional independence

Main idea: combine probabilities and graphs