Naive Bayesian networks, extensions

Peter Antal antal@mit.bme.hu

Overview

- Basic concepts of probability theory
  - Joint distribution
  - Conditional probability
  - Bayes’ rule
  - Chain rule
  - Marginalization
  - General inference
  - Independence
    - Conditional independence
    - Contextual independence
    - Direct dependency
    - Independence model
      - Logical properties
- **Naive Bayesian networks**
  - **Definition**
  - **Inference**
  - Full Bayesian treatment
    - Specification
      - parameters
      - structures
    - Inference
    - Learning
- Extensions of N-BNs and related models
  - Tree-augmented N-BN
  - BN-augmented BN
  - Hierarchical BN
  - Context-sensitive independencies
  - Noisy-OR
  - Logistic regression

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain
- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  Cavity = false $\wedge$ *Toothache = false*

  Cavity = false $\wedge$ *Toothache = true*

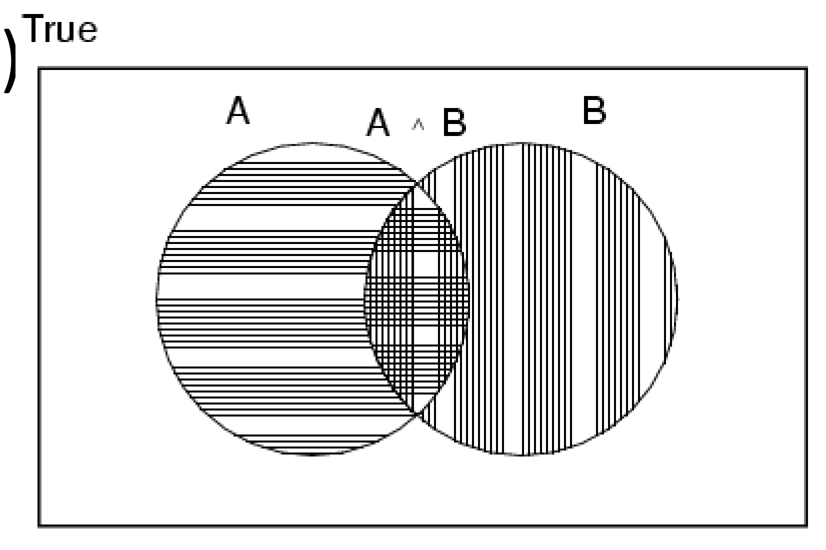
  Cavity = true $\wedge$ *Toothache = false*

  Cavity = true $\wedge$ *Toothache = true*

- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A, B*
-
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B)$
  -



Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

- Boolean random variables
- e.g., *Cavity* (do I have a cavity?)
-
- Discrete random variables
- e.g., *Weather* is one of <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive

- Elementary proposition constructed by assignment of a value to a
- random variable: e.g., *Weather = sunny*, *Cavity = false*
- (abbreviated as $\neg$*cavity*)

- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather = sunny* $\vee$ *Cavity = false*

Joint (probability) distribution

- Prior or unconditional probabilities of propositions
- e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence
- Probability distribution gives values for all possible assignments:
- **P**(*Weather*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
- **P**(*Weather,Cavity*) = a 4 × 2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Conditional probability

- Conditional or posterior probabilities
-

  e.g., P(*cavity* | *toothache*) = 0.8

  i.e., given that *toothache* is all I know

- (Notation for conditional distributions:
-

  P(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
-

  P(*cavity* | *toothache,cavity*) = 1

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

  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial
-

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
-
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
-
- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(\mathit{Weather},\mathit{Cavity}) = \mathbf{P}(\mathit{Weather} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})$
- (View as a set of 4 × 2 equations, not matrix mult.)
-

Bayes' rule

An algebraic triviality

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)} = \frac{p(Y \mid X)\, p(X)}{\sum_X p(Y \mid X)\, p(X)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Chain rule

- Chain rule is derived by successive application of product rule:
- $\mathbf{P}(X_1, \ldots, X_n)$ $= \mathbf{P}(X_1,\ldots,X_{n-1})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$

  $= \mathbf{P}(X_1,\ldots,X_{n-2})\ \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$

  $= \ldots$

  $= \pi\ \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

Marginalization

- ~Summing out/averaging out
- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Can also compute conditional probabilities:
-

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Denominator can be viewed as a normalization constant α
-

P(*Cavity | toothache*) = α, **P**(*Cavity,toothache*)
= α, [**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache,¬ catch*)]
= α, [<0.108,0.016> + <0.012,0.064>]
= α, <0.12,0.08> = <0.6,0.4>

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Any question about observable events in the domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** – **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_h\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
    1. Worst-case time complexity $O(d^n)$ where d is the largest arity
    2. Space complexity $O(d^n)$ to store the joint distribution
    3. How to find the numbers for $O(d^n)$ entries?

Independence, Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z defined as follows

for all x,y and z with $P(z)>0$: $P(x;y|z)=P(x|z)\,P(y|z)$

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Direct dependence: $D_P(X;Y|V/\{X,Y\})$

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

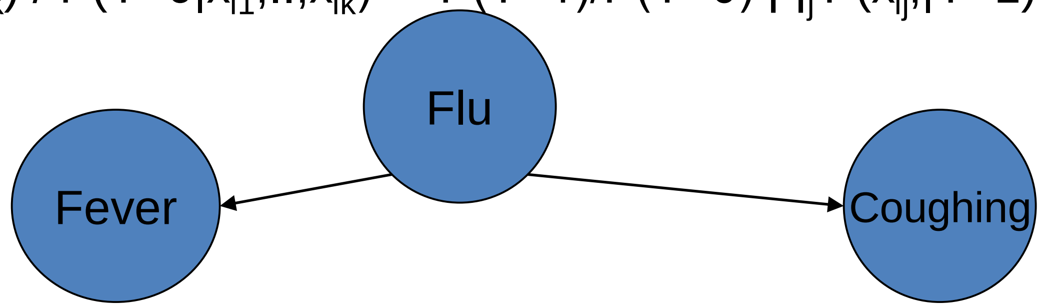
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \mid Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Summary

- Basic concepts of probability theory
  - On the use of probabilities: PDSS:2.1
  - The Bayesian framework: PDSS:2.2
  - LATER: Indepence models: PDSS:2.3

  https://www.mit.bme.hu/system/files/oktatas/targyak/9383/Antal_Valoszinusegi.pdf
- Naive Bayesian networks
  - Definition, Inference (PDSS:2.5.1)
  - Full Bayesian treatment: LATER
  - ➔IDA:9.2.5 (~9.2)

  https://www.mit.bme.hu/system/files/oktatas/targyak/9383/Antal_IDA.pdf
- Naive Bayesian networks allow
  - linear number of free parameteres,
  - inference in linear number of steps.