Adapted from AIMA slides

Simple probabilistic models

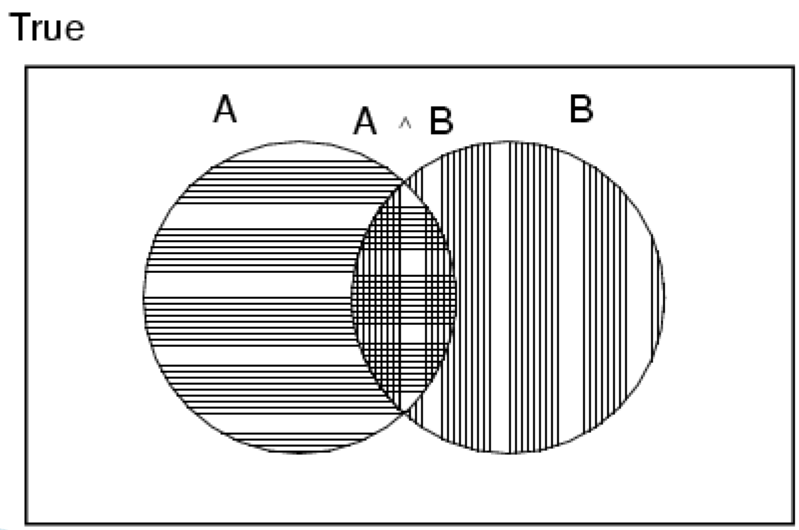
Peter Antal
antal@mit.bme.hu

Outline

- Basics of probability theory
- Relation of two-valued vs probabilistic logic
  - Truth vs belief
  - Proofs vs inference
- Naïve Bayesian networks
- SPAM filter
- Special local models
  - Noisy-OR
  - Decision tree CPDs
  - Decision graph CPDs

Axioms of probability

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



Probability theory: concepts for the course

- Joint distribution ("omic-ness")
  - ("omic-ness": "comprehensiveness" + "query-free")
- Conditional probability ("simple inference")
- Chain rule ("factorization")
- Bayes' rule ("inversion")
- Marginalization/expansion ("complex inference")
- [Conditional] independence ("simplification")

Bayes' Rule

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- $\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) \,/\, P(b)$
- or in distribution form
- $$P(Y|X) = P(X|Y)\, P(Y) \,/\, P(X) = \alpha P(X|Y)\, P(Y)$$
- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis, *S* be stiff neck:
  - $P(m|s) = P(s|m)\, P(m) \,/\, P(s) = 0.8 \times 0.0001 \,/\, 0.1 = 0.0008$
  - Note: posterior probability of meningitis still very small!

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.

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

Inference by enumeration

Every question about a 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?

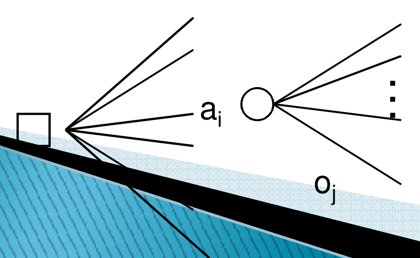
Decision theory= probability theory+utility theory

- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
    - QALY, micromort
  - Maximum Expected Utility Principle (MEU)
    - Best action is the one with maximum expected utility

$$EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$$

$$a^* = \arg\max_i EU(a_i)$$

Actions a_i (which experiment)	Outcomes (e.g. dataset)	Probabilities	Utilities, costs	Expected utilities
		$P(o_j \mid a_i)$	$U(o_j)$, $C(a_i)$	$EU(a_i) = \sum P(o_j \mid a_i) U(o_j \mid a_i)$



About the event space

- Atomic events are mutually exclusive and exhaustive.
- The single variable case.
  - *Weather* is one of <*sunny,rainy,cloudy,snow*>
  - $P((Weather = sunny) \lor (Weather = rainy))$
- Challenges in the multivariate case.
  - *Weather* is one of <*sunny,rainy,cloudy,snow*>
  - *TemperatureofRain* is one of <*icy,cold,warm*>
    - NONE?

Classical vs probabilistic logic: truth and beliefs

P_1	…	P_3	KB	s	pKB	P(query|evidence)
F	F	F	F	T	.01	.1
F	F	T	T	F	.12	.2
F	T	F	F	T	.35	.3
F	T	T	F	F	..	..
T	F	F	F	T	..	..
T	F	T	T	T	…	…
T	T	F	F	T	..	..
T	T	T	F	T	..	..

Classical vs probabilistic logic: provability and inference

$P(KB \vdash_i \alpha$ = sentence α can be derived from *KB* by procedure $i | pKB)$

"Belief propagation in networks"

Conditional independence

„Probability theory=measure theory+independence”

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z:
- $P(X;Y|z)=P(Y|z)\,P(X|z)$ for all z with $P(z)>0$.

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

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

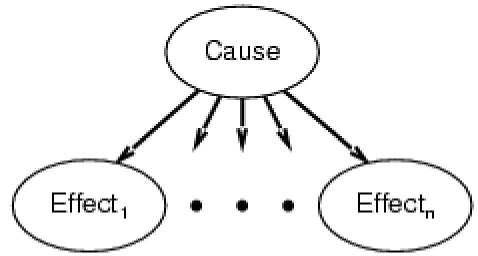
Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Contextual independence: for not all z.

Homeworks:
- Intransitivity: show that it is possible that D(X;Y), D(Y;Z), but I(X;Z).
- order : show that it is possible that I(X;Z), I(Y;Z), but D(X,Y;Z).

Naive Bayesian network

Assumptions:



1, Two types of nodes: a cause and effects.

2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

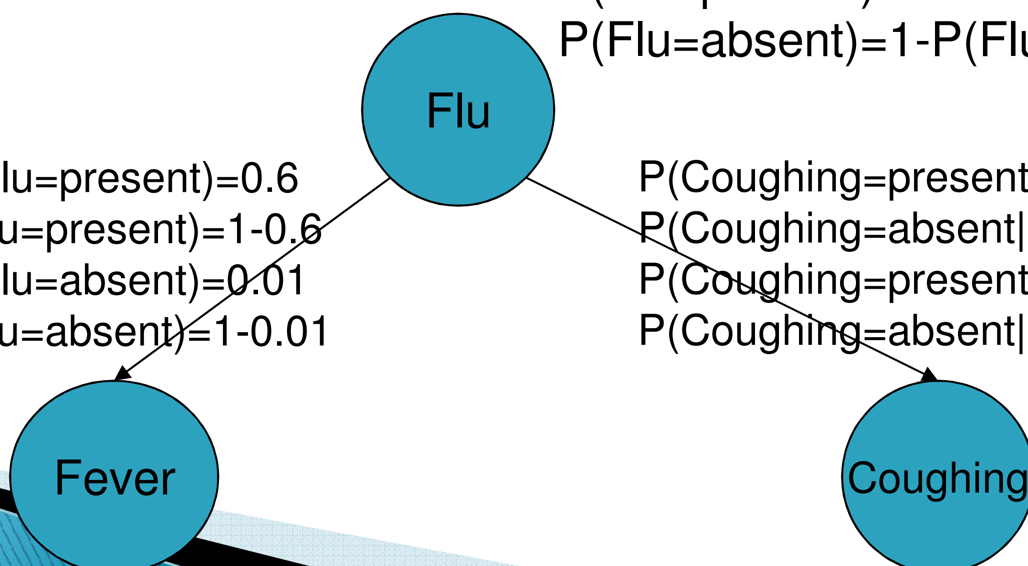
Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model

P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)

P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1-0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1-0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02



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

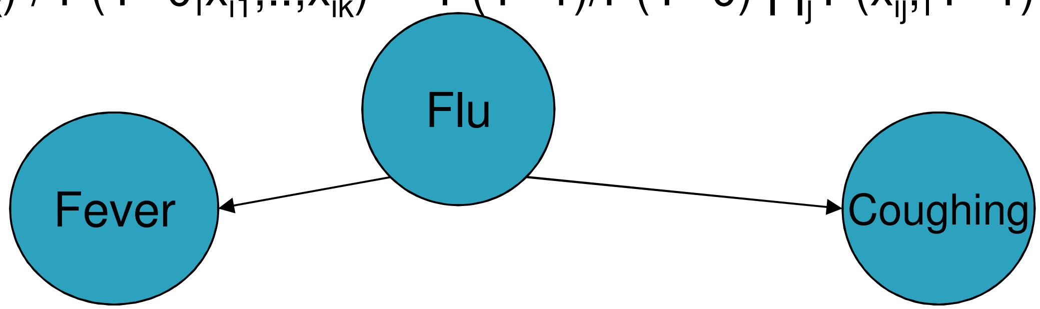
2n+1 parameteres!

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

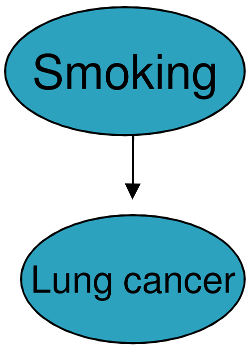
$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



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

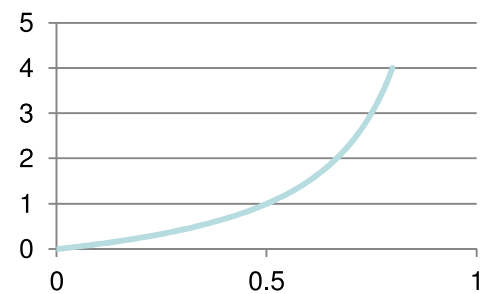
$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Conditional probabilities, odds, odds ratios

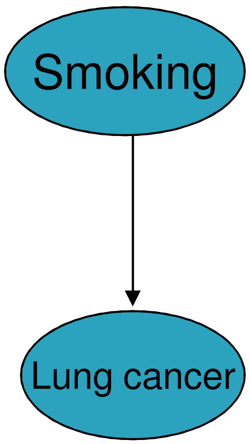


	¬S	S	
¬LC	P(¬S, ¬LC)	P(S, ¬LC)	P(¬LC)
LC	P(¬S, LC)	P(S, LC)	P(LC)
	P(¬S)	P(S)	

Probability:
P(LC)
Conditional probabilities (e.g., probability of LC given S)**:**
P(LC| ¬S)= ??? P(LC| S)= ??? P(LC)
Odds:
[0,1] →[0,∞]: Odds(p)=p/(1-p)
O(LC| ¬S)= ??? O(LC| S)
Odds Ratio (OR) Independent of prevalence!
OR(LC,S)=O(LC| S)/O(LC| ¬S)



Probabilities, odds, odds ratios



?

Independence:
null modell (H_0)

	¬S	S	
¬LC	8	7	15
LC	1	4	5
	9	11	20

Contingency table with marginals

	¬S	S	
¬LC	.4	.35	.75
LC	.05	.2	.25
	.45	.55	

Conditional probabilities:

P(LC| ¬S)=.11 ??? P(LC| S)=.36 ??? P(LC)=.25

Odds:

$[0,1] \rightarrow [0,\infty]$: Odds(p)=p/(1-p)

O(LC| ¬S)=.12 ??? O(LC| S)=.56

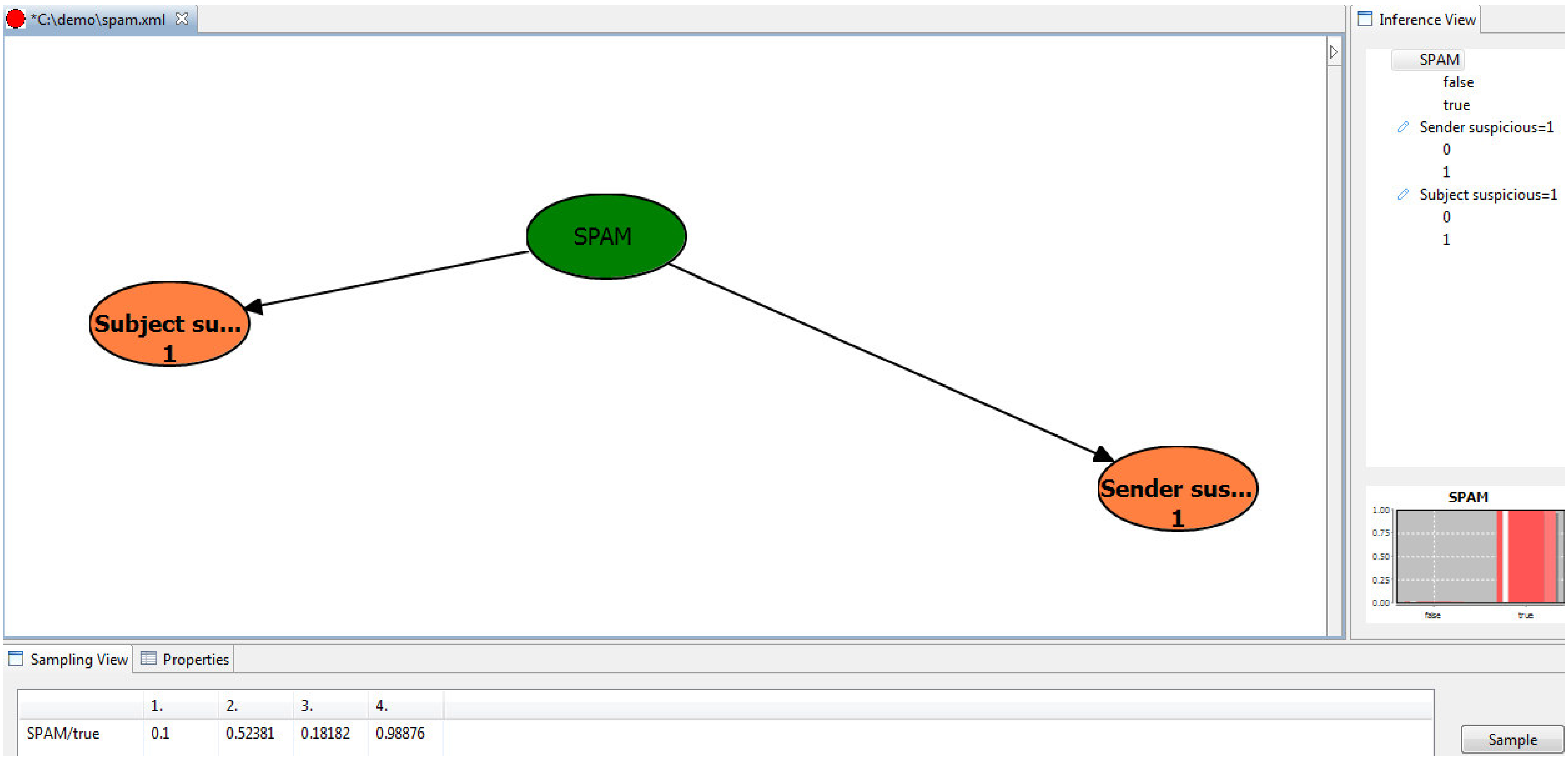
Odds Ratio (OR):

OR(LC,S)=O(LC| S)/O(LC| ¬S)=4.6

Bayes Cube (~Bayes Eye)

http://redmine.genagrid.eu/projects/bayescubedownload/wiki/Wiki

Example: Construct a spam filter



Summary

- Naïve Bayesian networks (N-BNs) demonstrate the use of independencies to cope with
  - model complexity (~space complexity, number of parameters)
  - inferential complexity (~time complexity).
- The assumption of conditional independence of the effects given their common cause allows
  - the efficient representation of the joint distribution
    - (in the discrete, multinomial case: linear number of parameters instead of exponential),
  - the efficient computation of the diagnostic posterior p(Y|X)
    - (linear number of steps instead of exponential),
- Odds, log odds are popular transformations of probabilities.
- N-BNs are robust knowledge engineering and data analysis tools.
-
- **Suggested reading:**
  - Druzdzell: Building Probabilistic Networks: Where Do the Numbers Come From?, IEEE Transactions on Knowledge and data engineering, 2000