Full Bayesian inference using Naive Bayesian networks

Peter Antal
antal@mit.bme.hu

Overview

- Naive Bayesian networks
  - Definition
  - Inference
- Full Bayesian inference and learning
- Bayesian learning using conjugacy: Beta
  - Specification
  - Inference
  - Learning

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.

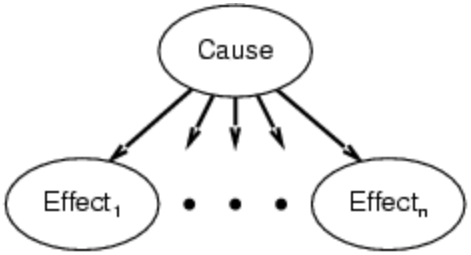
$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Simple probabilistic models

- Total independence
- **Naive Bayesian networks**
- Hidden Markov Models

Naive Bayesian network

Assumptions:



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

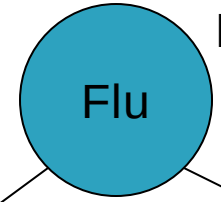
2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

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

Model

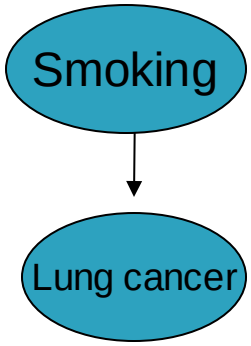
P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)



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

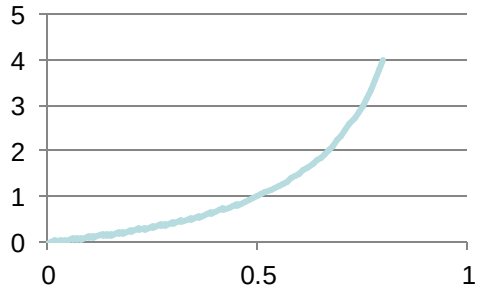
P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02

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)



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

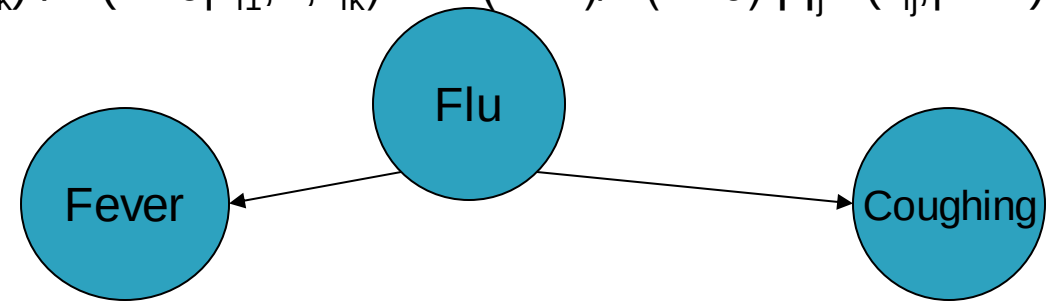
2n+1 parameteres!

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

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



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

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

Full Bayesian naive-BN

- Structure prior: p(G)
  - Specify priors for edges in G
  - Penalize deviation from a prior structure G_0
- Parameter prior: p(Θ|G)
  - θ denotes the complete parametrization for G
  - Specify p(Θ|G) independently for each variable?
  - Specify p(Θ|G) using a „convenient" (~conjugate) prior?
- Inference
  - ?

Full Bayesian inference by conjugacy

3. Definition. *A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.*

1. Example. *Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.*

$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{13}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \; where \; c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{14}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

In general a conjugate prior is updated to posterior using only an appropriate statistics of the observations to update its parametrization. It shows that the parameters frequently has an intuitive interpretation based on observations, that is in the prior specification the parameters corresponds to real or virtual past observations.

The Dirichlet distribution

3. Example. *Assume that the observed sequence* $D_n = \{X_i; i = 1, 2 \ldots, n\}$ *contains i.i.d. multinomial samples with* L *discrete values. The prior is a Dirichlet prior with hyperparameters* $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_L$ *and* $\alpha_\cdot = \sum_i \alpha_i$.

$$p(\theta) = Di(\boldsymbol{\alpha}) = c \prod_i \theta^{\alpha_i - 1} \; where \; c = \frac{\Gamma(\alpha_\cdot)}{\prod_i \Gamma(\alpha_i)} \tag{42}$$

Principles for induction

- Epicurus' (342? B.C. – 270 B.C.) principle of multiple explanations which states that one should *keep all hypotheses that are consistent with the data*.
- The principle of Occam's razor (1285 – 1349, sometimes spelt Ockham). Occam's razor states that when inferring causes *entities should not be multiplied beyond necessity*. This is widely understood to mean: Among all hypotheses consistent with the observations, choose the simplest. In terms of a prior distribution over hypotheses, this is the same as giving simpler hypotheses higher a priori probability, and more complex ones lower probability.

Bayesian inference with multiple models

Assume multiple models $M_i = (S_i, \theta_i)$ with prior $p(M_i)$ $i = 1, \ldots, M$.

The inference $p(Q = q|E = e)$ can be performed as follows:

$$p(q|e) = \Sigma_{i=1,\ldots,M} p(q, M_i|e) = \Sigma_{i=1,\ldots,M} p(q|M_i, e)p(M_i|e)$$

Note that $p(M_i|e)$ is a posterior over models with evidence e:

$$p(M_i|e) = \frac{p(e|M_i)p(M_i)}{p(e)} \propto p(e|M_i)p(M_i)$$

i.e., the evidence e reweight our beliefs in multiple models.

The inference is performed by **Bayesian Model Averaging** (BMA). Epicurus' (342(?) B.C. - 270 B.C.) **principle of multiple explanations** which states that one should keep all hypotheses that are consistent with the data.

Bayesian model averaging

Beside models, assume N multiple complete observations D_N.

The standard inference $p(Q = q|E = e, D_N)$ is defined as:

$$p(q|e, D_N) = \sum_{i=1,\ldots,M} p(q, M_i|e, D_N) = \sum_{i=1,\ldots,M} p(q|M_i, e, D_N)p(M_i|e, D_N)$$

Because $p(q|M_i, e, D_N) = p(q|M_i, e)$ and $p(M_i|e, D_N) \approx p(M_i|D_N)$:

$$p(q|e, D_N) \approx \sum_{i=1,\ldots,M} p(q|M_i, e)p(M_i|D_N)$$

where again $p(M_i|D_N)$ is a posterior after observations D_N:

$$p(M_i|D_N) = \frac{p(D_N|M_i)p(M_i)}{p(e)} \propto \underbrace{p(D_N|M_i)}_{likelihood} \underbrace{p(M_i)}_{prior}.$$

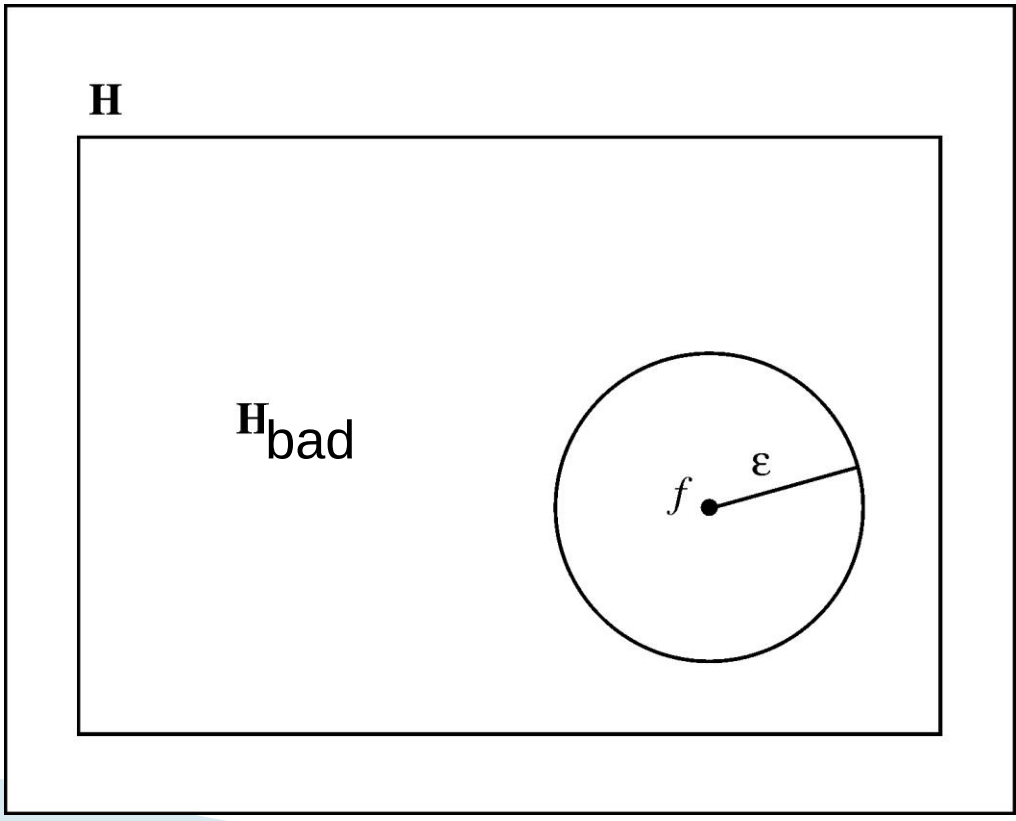
i.e., our rational foundation, probability theory, automatically includes and normatively defines learning from observations as standard Bayesian inference!

The Probably Approximately Correct PAC-learning

A single estimate of the expected error for a given hypothesis is convergent, but can we estimate the errors for all hypotheses uniformly well??

Example from concept learning

X: i.i.d. samples.
n: sample size
H: hypotheses
(hypothesis space)
|H|: cardinality
(number of hypotheses)



Full Bayesian inference with N-BNs using complete data

- Integration over parameters?
  - Analytical solution under parameter independence!
  - Hyperparameter update.
- Bayesian model averaging over exponential number of structures?
  - Analytical solution!
  - Existence of equivalent „super”-parametrization!!

Extensions of N-BNs

- Tree-augmented BNs
- BN-augmented BNs
- Hierarchical BNs
- Multiple parents
  - Explaining away
- „Context-sensitive" N-BNs

Inference and learning using incomplete data?

- Later (there is no analytic solutions)

References

- Domingos, Pedro, and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss." *Machine learning* 29.2-3 (1997): 103-130.
- Friedman, Jerome H. "On bias, variance, 0/1—loss, and the curse-of-dimensionality." *Data mining and knowledge discovery* 1.1 (1997): 55-77.
- Hand, David J., and Keming Yu. "Idiot's Bayes—not so stupid after all?." *International statistical review* 69.3 (2001): 385-398.
- Dash, Denver, and Gregory F. Cooper. "Exact model averaging with naive Bayesian classifiers." *ICML*. 2002.
- Langseth, Helge, and Thomas D. Nielsen. "Classification using hierarchical naive bayes models." *Machine learning* 63.2 (2006): 135-159.

Summary

- Naive Bayesian networks
  - Definition, Inference
  - Full Bayesian treatment: LATER