Introduction to Graphical Models
lecture 2 - Bayesian Networks

Marc Toussaint[1]
TU Berlin

overview:
– multiplying probability tables – naive Bayes
– graphical models
– elimination algorithm

[1] slides based on Chris Williams' PMR lectures

cheat sheat

- a random variable X assignes probabilties $P(X=x) \in \mathbb{R}$ to values $x \in \text{dom}(x)$
- probabilty distribution $\leftrightarrow$ table (vector) of probabilties for each value (normalization: $\sum_X P(X) = 1$)
- joint distribution $P(X,Y)$ $\leftrightarrow$ table (matrix) of probabilties
- definition: marginal $P(X) = \sum_Y P(X,Y)$ (summing over columns/rows)
- definition: conditional $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
- implications:

$$P(X,Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

$$P(X_1,..,X_n) = \prod_{i=1}^n P(X_i|X_1,..,X_{i\text{-}1})$$

$$P(X|Y) = \frac{P(Y|X)}{P(Y)} P(X)\ , \quad \text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\ \text{prior}$$

- definition: *inference* is the problem to compute

$$P(Y_{1:k}\,|\,E_{1:m}) = \frac{P(Y_{1:k},E_{1:m})}{P(E_{1:m})} \propto \sum_{H_{1:n}} P(Y_{1:k},E_{1:m},H_{1:n})$$

Independence

- definition: X is independent of Y iff:

  $P(X|Y) = P(X)$

  for all possible values $x \in \text{dom}(X)$ and $y \in \text{dom}(Y)$
  (matrix thinkers: every column of $P(X|Y)$ is equal)
  (definition holds also for set so variables $X = (X_1,..,X_n)$, $Y = (Y_1,..,Y_m)$)

- in terms of the joint: X independent of Y iff:

  $P(X,Y) = P(X)\ P(Y)$

  (matrix thinkers: matrix $P(X,Y)$ is the outer product of $P(X)$ and $P(Y)$)

- X independent of Y $\iff$ Y independent of X

- a set of variables $X_1,..,X_n$ is independent iff

  $$P(X_1,..,X_n) = \prod_{i=1}^n P(X_i)$$

Independence

recall example:

	Toothache = true	Toothache = false
Cavity = true	0.04	0.06
Cavity = false	0.01	0.89

- is T independent from C ?

Conditional Independence

- definition: X is conditionally independent of Y given Z iff

$$P(X|Y,Z) = P(X|Z)$$

for all $x \in \text{dom}(X)$, $y \in \text{dom}(Y)$, $z \in \text{dom}(Z)$

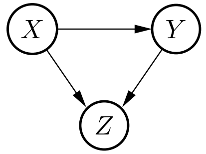
- in terms of the joint:

$$P(X,Y,Z) = P(X,Y|Z)\ P(Z) = P(X|Z)\ P(Y|Z)\ P(Z)$$

Bayesian Networks

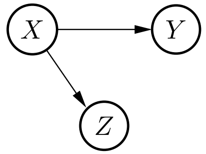
1st model:

$P(Z,Y,X) =$
$P(Z|Y,X)\ P(Y|X)\ P(X)$



2nd model:

$P(Z,Y,X) =$
$P(Z|Y)\ P(Y|X)\ P(X)$



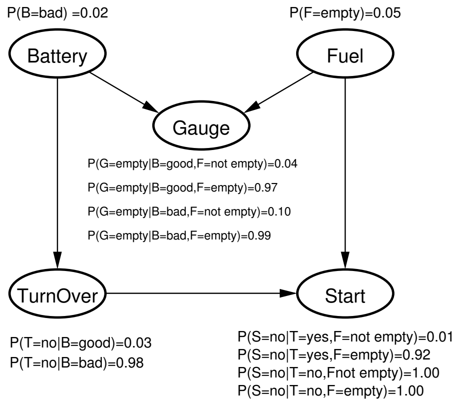
- recall the general chain rule:

$$P(X_1,..,X_n) = \prod_{i=1}^{n} P(X_i|X_1,..,X_{i\text{-}1})$$

- Bayesian network is a graphical notation of (in)dependence

Bayesian Network example

(Heckermann 1995)



$$\iff\ P(S,T,G,F,B) = P(S|T,F)\ P(T|B)\ P(G|F,B)\ P(F)\ P(B)$$

as compared to the general chain rule:
$P(S,T,G,F,B) = P(S|T,G,F,B)\ P(T|G,F,B)\ P(G|F,B)\ P(F|B)\ P(B)$

Bayesian Network example

$$P(S,T,G,F,B) = P(S|T,F)\ P(T|B)\ P(G|F,B)\ P(F)\ P(B)$$

- table sizes: LHS = 2^5 RHS = $s^3 + 2^2 + 2^3 + 2 + 2$

- what is the probability of:
  $P(B = good, T = no, G = empty, F = notempty, S = no)$?

Inference in the Bayes Net

recall: general def of inference:

$$P(Y_{1:k} \mid E_{1:m}) = \frac{P(Y_{1:k}, E_{1:m})}{P(E_{1:m})} \propto \sum_{H_{1:n}} P(Y_{1:k}, E_{1:m}, H_{1:n})$$

- in our example
  – compute $P(B|S = no)$ or $P(F|T = no)$
  – compute $P(B, F|T = no)$

- definition: *elimination* $\equiv$ "summing out variables"
  (eliminate Y from $P(X, Y|Z)$ means to compute $P(X|Z) = \sum_Y P(X, Y|Z)$)

common methods:

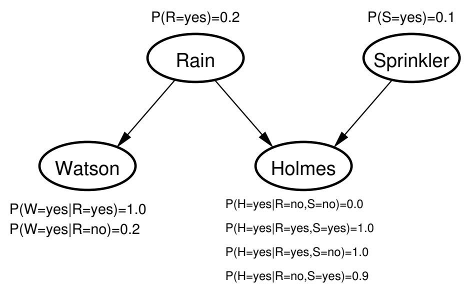
- to compute a *single* marginal (single inference query like $P(B|S = no)$):
  – Variable Elimination (see Jordan, ch 3)

- to compute *all* marginals (e.g., compute $P(B|S = no)$ and $P(F|S = no)$ and $P(G|S = no)$ and $P(T|S = no)$)
  - if model is a tree: inference in time linear in the number of nodes (Pearl, 1986); messages are passed up and down the tree; all the necessary computations can be carried out locally. HMMs (chains) are a special case of trees. Pearls method also applies to polytrees (DAGS with no undirected cycles)
  - if model is not a tree: clustering (grouping) of nodes to yield a tree of cliques (junction tree) (Lauritzen and Spiegelhalter, 1988)

- approximate methods in general graphs
  – sampling, loopy belief propagation, varational methods

- we will learn about these 'automatic' algorithms for inference next time

- here: some more examples...



$$\iff \quad P(H, W, S, R) = P(H|S, R)\ P(W|R)\ P(S)\ P(R)$$

- Mr. Holmes lives in Los Angeles. One morning when Holmes leaves his house, he realizes that his grass is wet. Is it due to rain, or has he forgotten to turn off his sprinkler?
- Calculate P(rjh), P(sjh) and compare these values to the prior probabilities
- Calculate P(r; sjh). r and s are marginally independent, but conditionally dependent
- Holmes checks Watsons grass, and finds it is also wet. Calculate P(rjh;w), P(sjh;w)
- This effect is called explaining away

Learning in Bayesian Networks

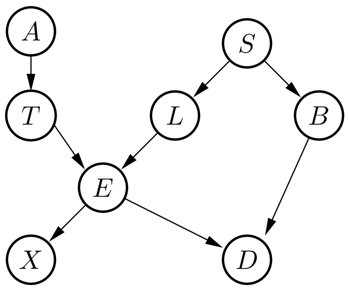
- General problem: learning probability models
  - learning CPTs; easier

    Especially easy if all variables are observed, otherwise can use EM
  - learning structure; harder

    Can try out a number of different structures, but there can be a huge number of structures to search through

- Say more about this later

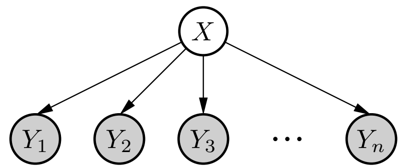
model for lung disease:



$\iff P(D, X, E, B, L, T, S, A) =$
$P(D|E, B)\ P(X|E)\ P(E|T, L)\ P(B|S)\ P(L|S)\ P(T|A)\ P(S)\ P(A)$

A=trip to asia
S=smoking
T=Tuberculosis
L=lung cancer
E=abnormality in chest
X=X-ray
D=Dyspnea
B=Bronchitis

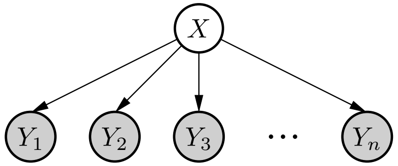
Naive Bayes



$\iff P(X, Y_{1:n}) = P(X)\ \prod_{i=1}^{n} P(Y_i|X)$

- one hidden variable, many *conditionally independent* evidences
- what is the posterior $P(X|y_{1:n})$?

Naive Bayes



$\iff P(X, Y_{1:n}) = P(X) \prod_{i=1}^n P(Y_i|X)$

- one hidden variable, many *conditionally independent* evidences
- what is the posterior $P(X|y_{1:n})$?
  $P(x|y_{1:n}) \propto P(x) \prod_{i=1}^n \mu_i(x)$ with $\mu_i(x) := P(Y_i = y_i \,|\, x)$

  – the posterior is a *product* of "messages" (prob. distributions $\mu_i(x)$)
  – each independent source of information contributes a "message"

- multiplying distributions is the core operation for fusing (independent!) information!