Lecture 15: Learning in Bayesian Networks

University of BRISTOL

EMAT31530/March 2018/Raul Santos-Rodriguez

Have a look at ...

... Russell and Norvig (Ch. 14 and Ch. 15)

... The introduction to the book Graphical Models; Foundations of Neural Computing (ed. M. Jordan)

... David barber's Bayesian reasoning and Machine Learning:
`http://www.cs.ucl.ac.uk/staff/d.barber/brml/`

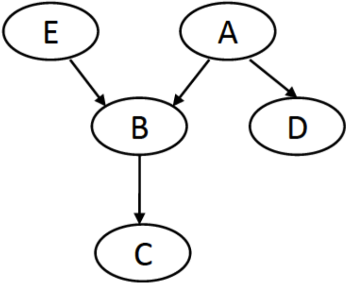
... Kevin Murphy's Toolbox:
`http://www.cs.ubc.ca/~murphyk/Software/`

Outline

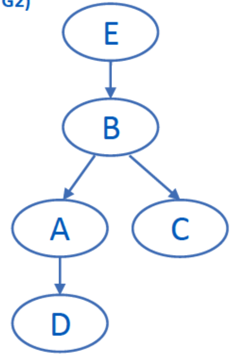
This lecture introduces the concept of learning in probabilistic graphical models. The objective is to discuss the following topics:

- Learning in Bayesian Networks
- Maximum likelihood
- Expectation-Maximization algorithm

G1)



G2)



Are these two bayesian networks equivalent?

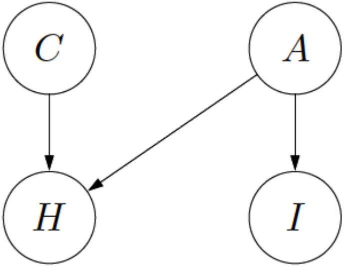
$$P(A,E,B,C,D) = P(A)P(E)P(B \mid A,E)P(C \mid B)P(D \mid A)$$

$$P(A,E,B,C,D) = P(E)P(B \mid E)P(A \mid B)P(C \mid B)P(D \mid A)$$

Review

So far, we have studied:

- Concept of Bayesian network
- Conditional independence
- Inference in Bayesian networks
- Dynamic Bayesian networks



Why learning?

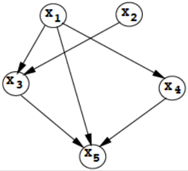
Knowledge acquisition bottleneck

- Knowledge acquisition is an expensive process
- Often we don't have an expert

Data is cheap

- Amount of available information growing rapidly
- Learning allows us to construct models from raw data

- Given:
  - A Bayesian network structure.



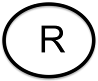
  - A data set

X_1	X_2	X_3	X_4	X_5
0	0	1	1	0
1	0	0	1	0
0	1	0	0	1
0	0	1	1	1
⋮	⋮	⋮	⋮	⋮

- Estimate conditional probabilities:

$$P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4|X_1), P(X_5|X_1, X_3, X_4)$$

Setting: Rating of a movie $\{1, 2, 3, 4, 5\}$



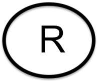
Parameters: $\theta = (P(1), P(2), P(3), P(4), P(5))$

Training data: $D_{train} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$

[Liang 2014]

We want to find θ using D_{train} but ...

... $P(R)$ is proportional to number of occurrences of R in D_{train}



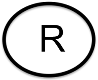
$$D_{train} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

θ :

R	P(R)
1	?
2	?
3	?
4	?
5	?

We want to find θ using D_{train} but ...

... $P(R)$ is proportional to number of occurrences of R in D_{train}



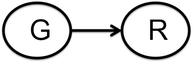
$$D_{train} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

θ :

R	P(R)
1	0.1
2	0
3	0.1
4	0.5
5	0.3

Example: two variables

Setting: Rating of a movie $\{1, 2, 3, 4, 5\}$ and Genre of a movie $\{drama, comedy\}$

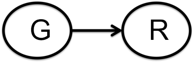


Parameters: $P(G, R) = P(G)P(R|G)$

Training data: $D_{train} = \{(d, 4), (d, 4), (d, 5), (c, 1), (c, 5)\}$

We want to find θ using D_{train} but ...

... $P(G)$, $P(R|G)$ are proportional to number of occurrences of R, G in D_{train}



$$D_{train} = \{(d,4),(d,4),(d,5),(c,1),(c,5)\}$$

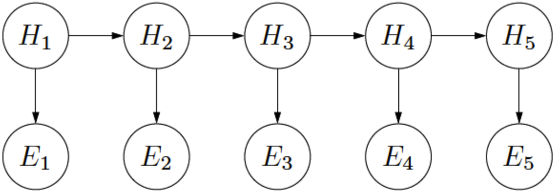
θ :

G	$P(G)$
d	3/5
c	2/5

G	R	$P(R\|G)$
d	4	2/3
d	5	1/3
c	1	1/2
c	5	1/2

Example: HMM

Setting: $H_1, \ldots, H_n$ are the Hidden variables and $E_1, \ldots, E_n$ are the observations



$$P(H, E) = \prod_{i=1}^{n} P_{trans}(H_i | H_{i-1}) P_{emi}(E_i | H_i)$$

Parameters: $\theta = \{P_{trans}, P_{emi}\}$

Maximum likelihood

Maximum likelihood objective:

$$\max_{\theta} \prod_{x_i \in D_{train}} P(X = x_i | \theta)$$

Example: $D_{train} = \{(d, 4), (d, 5), (c, 5)\}$

$$p(X = x|\theta) = P(G = d)P(R = 4|d)P(G = d)P(R = 5|d)P(G = c)P(R = 5|c)$$

Solution: take logs and solve for the best $\theta = \theta^*$

Question: what if we don't have data for all events?

Maximum likelihood vs Bayesian learning

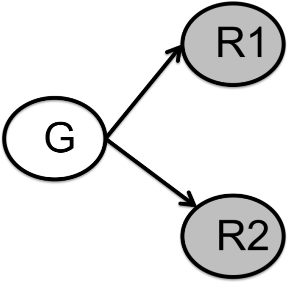
Maximum likelihood

- Assumes that θ is unknown but fixed parameter
- Finds θ^*, the value that maximizes the likelihood

Bayesian learning

- Treats θ as a random variable
- Assumes a prior probability of θ : $p(\theta)$
- Tries to compute the posterior probability of θ : $p(\theta|D)$

Expectation-Maximization (EM)



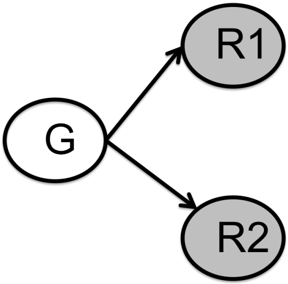
What if we don't observe some of the variables?

Example: $D_{train} = \{(?, 4, 5), (?, 4, 4), (?, 5, 3), (?, 1, 2), (?, 5, 4)\}$

Expectation-Maximization (EM)

Variables: H is hidden, E is observed (to be e)

Example: $H = \{G\}$, $E = \{R_1, R_2\}$



Maximum likelihood objective:

$$\max_{\theta} \prod_{e \in D_{train}} P(E = e | \theta)$$

$$= \max_{\theta} \prod_{e \in D_{train}} \sum_{h} P(E = e, H = h | \theta)$$

Expectation-Maximization (EM)

Algorithm: Expectation Maximization

E-step:

Compute $q(h) = P(H = h|E = e, \theta)$ for each h

Create weighted points: (h, e) with weight $q(h)$

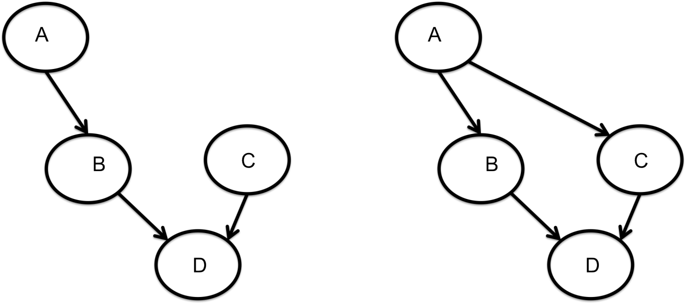
M-step:

Compute maximum likelihood (just count and normalise) to get θ

Repeat until convergence.

Model selection

Given a new dataset $\{A, B, C, D\}$, we can evaluate the probability of each model structure (using the parameters we learned by maximum likelihood) and pick the model with the highest $P(A, B, C, D|parameters)$.



For more information:
`http://research.microsoft.com/en-us/um/people/heckerman/tutorial.pdf`

Next lecture

In the next lecture, we will present a application area: text mining!