Machine learning: lecture 20

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics

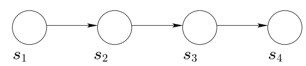
- Representation and graphical models
  - examples
- Bayesian networks
  - examples, specification
  - graphs and independence
  - associated distribution

What is a good representation?

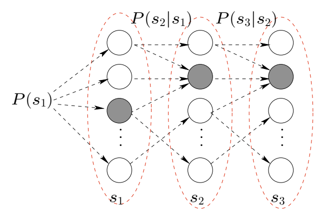
- Properties of good representations
  1. Explicit
  2. Modular
  3. Permits efficient computation
  4. etc.

Representation: explicit

- Representation in terms of variables and dependencies (a graphical model):



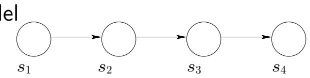
- Representation in terms of state transitions (transition diagram)



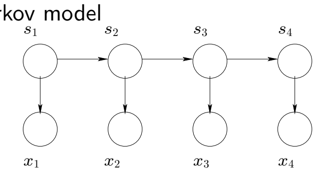
Representation: modular

- We can easily add/remove components of the model

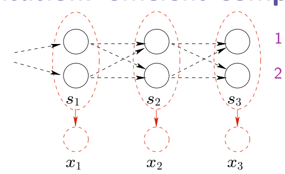
Markov model



Hidden Markov model



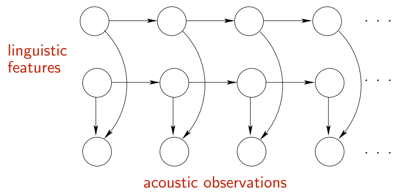
Representation: efficient computation



- Posterior marginals (forward-backward)
- Max-probabilities (viterbi)

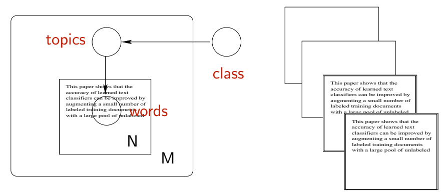
Graphical models: examples

- Factorial Hidden Markov model as a Bayesian network (directed graphical model)



Graphical models: examples

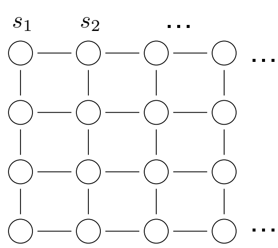
- Plates and repeated sampling



  – each document has N words, sampled from a distribution that depends on the choice of topics
  – the topics for each document are sampled from a class conditional distribution

Graphical models: examples

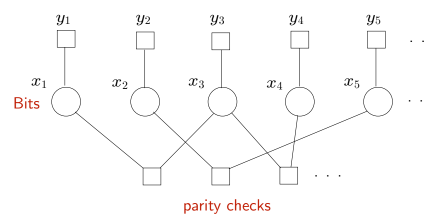
- Lattice models (e.g., Ising model) as a Markov random field



  – symmetric interactions (e.g., alignment of two nearby spins is energetically favorable)

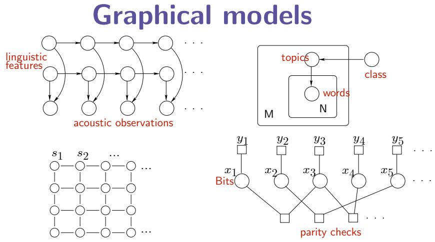
Graphical models: examples

- Factor graphs and codes (information theory)



  – circles denote variables while the squares are factors (functions) that constrain the values of the variables

Graphical models



- Graph semantics:

  graph ⇒ separation properties ⇒ independence

- Association with probability distributions:

  independence ⇒ family of distributions

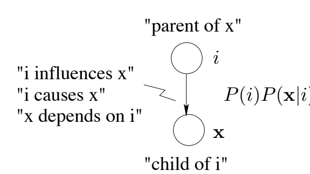
- Inference and estimation:

  graph structure ⇒ efficient computation

Bayesian networks

- Bayesian networks are directed acyclic graphs, where the nodes represent variables and directed edges capture dependencies

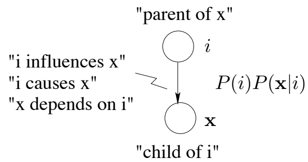
A mixture model as a Bayesian network



Bayesian networks

- Bayesian networks are directed acyclic graphs, where the nodes represent variables and directed edges capture dependencies

A mixture model as a Bayesian network



- Graph semantics:

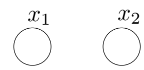
  graph $\Rightarrow$ separation properties $\Rightarrow$ independence

- Association with probability distributions:

  independence $\Rightarrow$ family of distributions

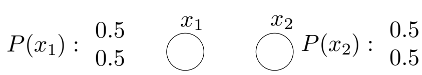
Example

- A simple Bayesian network: coin tosses



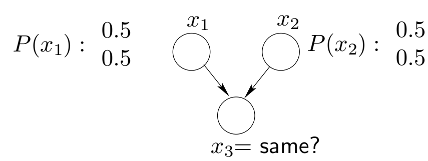
Example

- A simple Bayesian network: coin tosses



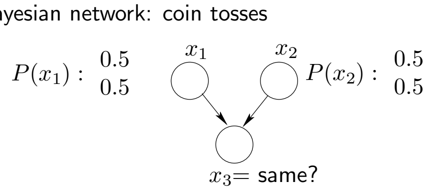
Example

- A simple Bayesian network: coin tosses



Example

- A simple Bayesian network: coin tosses

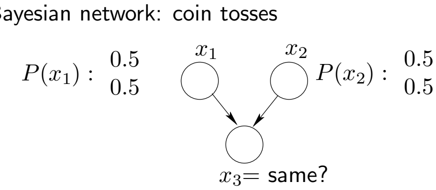


$P(x_3|x_1,x_2)$:

	hh	ht	th	tt
y	1.0	0.0	0.0	1.0
n	0.0	1.0	1.0	0.0

Example

- A simple Bayesian network: coin tosses



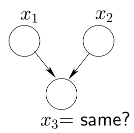
$P(x_3|x_1,x_2)$:

	hh	ht	th	tt
y	1.0	0.0	0.0	1.0
n	0.0	1.0	1.0	0.0

- Two levels of description
  1. graph structure (dependencies, independencies)
  2. associated probability distribution

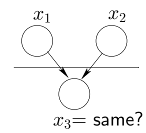
Example cont'd

- What can the graph alone tell us?



Example cont'd

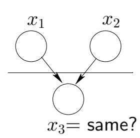
- What can the graph alone tell us?



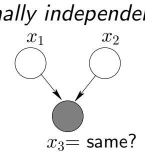
- x_1 and x_2 are *marginally independent*

Example cont'd

- What can the graph alone tell us?



- x_1 and x_2 are *marginally independent*

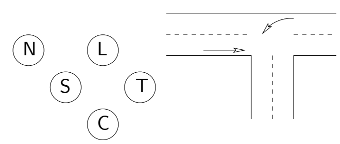


- x_1 and x_2 become *dependent* if we know x_3

  (the dependence concerns our beliefs about the outcomes)

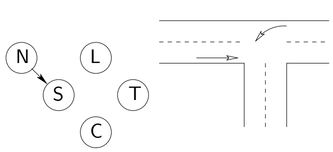
Traffic example

N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?



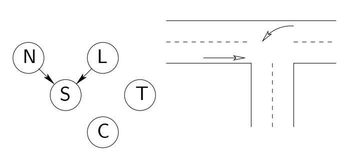
Traffic example

N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?



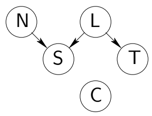
Traffic example

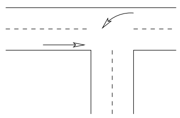
N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?



Traffic example

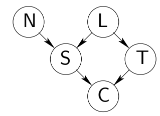
N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?

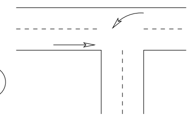




Traffic example

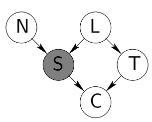
N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?

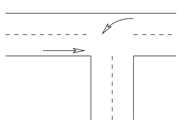




Traffic example

N = X is nice?
L = traffic light
S = X decides to stop?
T = the other car turns left?
C = crash?

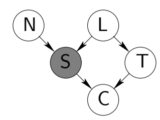




- If we only know that X decided to stop, can X's character (variable N) tell us anything about the other car turning (variable T)?

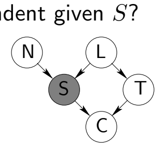
Graph, independence, d-separation

- Are N and T independent given S?



Graph, independence, d-separation

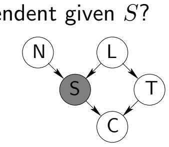
- Are N and T independent given S?



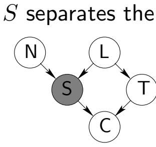
Definition: Variables N and T are D-separated given S if S separates them in the moralized ancestral graph

Graph, independence, d-separation

- Are N and T independent given S?



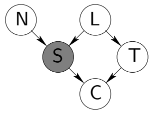
Definition: Variables N and T are D-separated given S if S separates them in the moralized ancestral graph



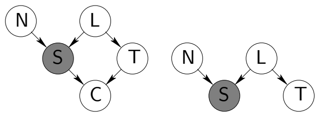
original

Graph, independence, d-separation

- Are N and T independent given S?



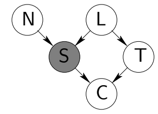
Definition: Variables N and T are D-separated given S if S separates them in the moralized ancestral graph



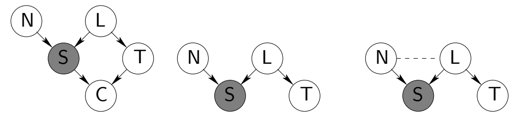
original ancestral

Graph, independence, d-separation

- Are N and T independent given S?



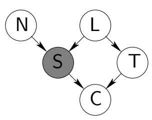
Definition: Variables N and T are D-separated given S if S separates them in the moralized ancestral graph



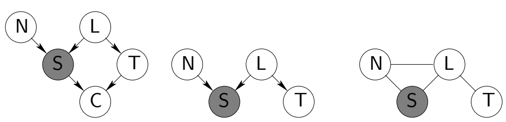
original ancestral moralized ancestral

Graph, independence, d-separation

- Are N and T independent given S?

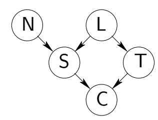


Definition: Variables N and T are D-separated given S if S separates them in the moralized ancestral graph



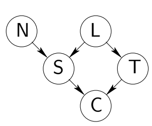
original ancestral moralized ancestral

Graphs and distributions



- A graph is a compact representation of a large collection of independence properties

Graphs and distributions



- A graph is a compact representation of a large collection of independence properties

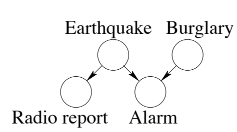
  Theorem: Any probability distribution that is consistent with a directed graph G has to factor according to "node given parents":

$$P(\mathbf{x}|G) = \prod_{i=1}^{d} P(x_i|\mathbf{x}_{pa_i})$$

  where $\mathbf{x}_{pa_i}$ are the *parents* of x_i and d is the number of nodes (variables) in the graph.

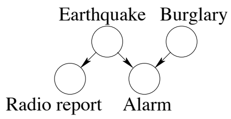
Explaining away phenomenon

- Model

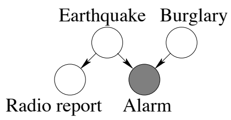


Explaining away phenomenon

- Model

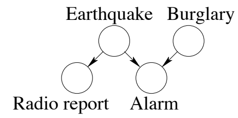


- Evidence, competing causes

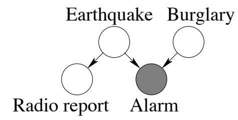


Explaining away phenomenon

- Model



- Evidence, competing causes



- Additional evidence and explaining away