Probabilistic Graphical Models

CVFX

2015.04.21

- 內容
  - representation
  - inference
  - learning
- 實例
  - 電腦視覺、影像處理

目標

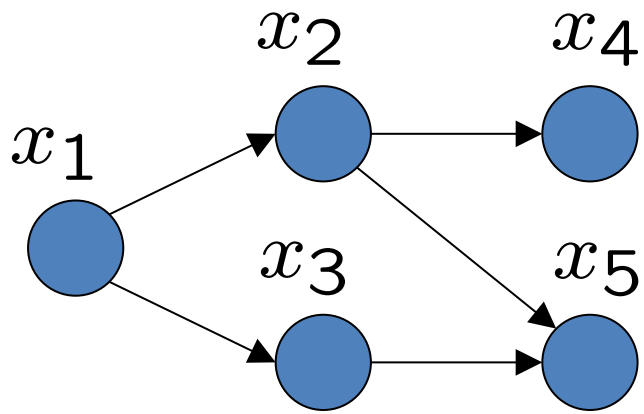
- 甚麼是 probabilistic graphical models?
- 可以用來解決甚麼問題? 怎麼用?

參考內容

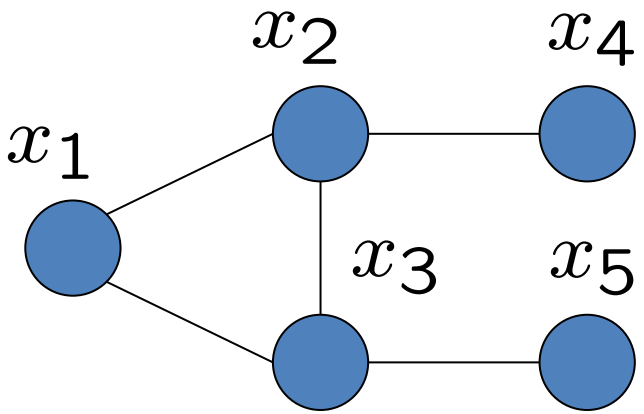
- “Probabilistic Graphical Models: Principles and Techniques”
  - Daphne Koller and Nir Friedman
  - http://pgm.stanford.edu/
  - MOOC course on *Coursera*
  - “Graphical Models in a Nutshell” http://ai.stanford.edu/~koller/Papers/Koller+al:SRL07.pdf

Graphs

nodes and links



directed



undirected

機率

random variables

joint probability

Independence

marginal probability

conditional probability

機率 + 圖形

a tool for modeling uncertainty

a general-purpose modeling language for exploiting the independence properties in the distribution

uncertainty:

probabilities

logical structure:

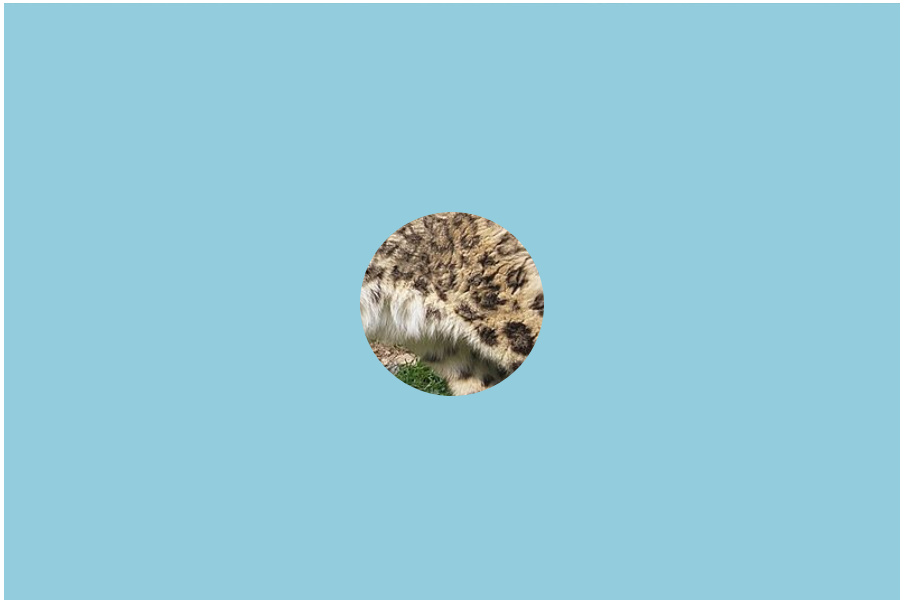
independence constraints

uncertainty:

1. observations are partial
2. observations are noisy
3. innate nondeterministic

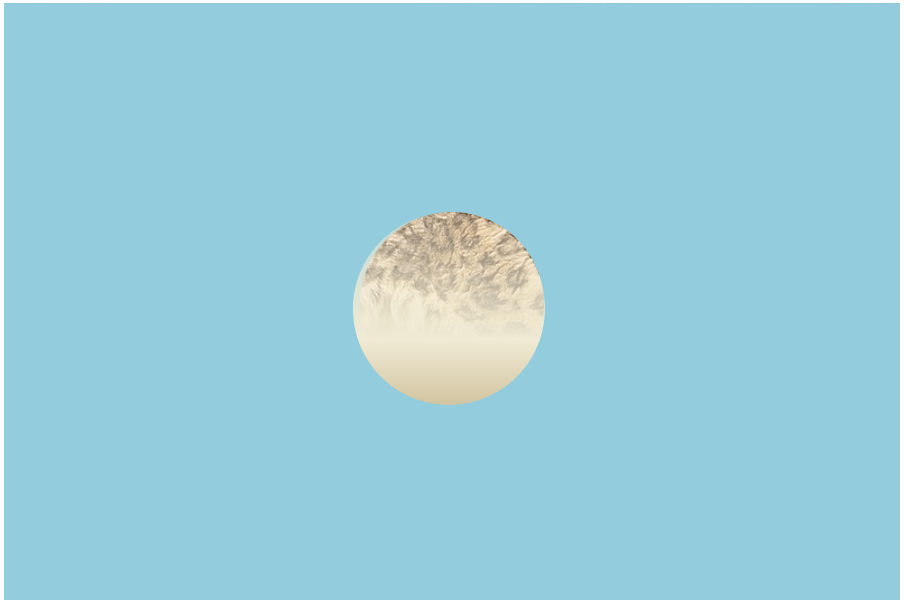
uncertainty:

1. observations are partial
2. observations are noisy
3. innate nondeterministic



uncertainty:

1. observations are partial
2. observations are noisy
3. innate nondeterministic

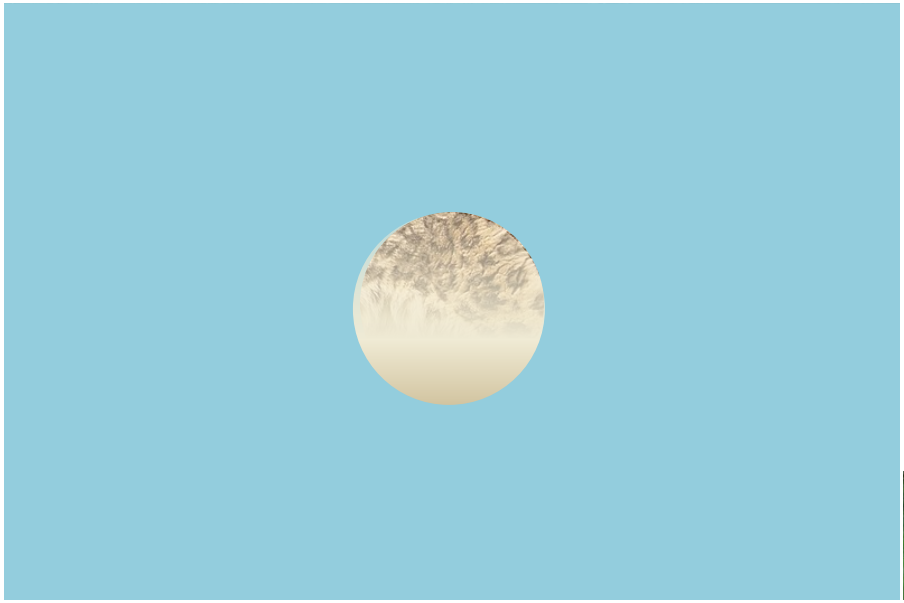


uncertainty:

1. observations are partial
2. observations are noisy
3. innate nondeterministic

uncertainty:

1. observations are partial
2. observations are noisy
3. innate nondeterministic



uncertainty:

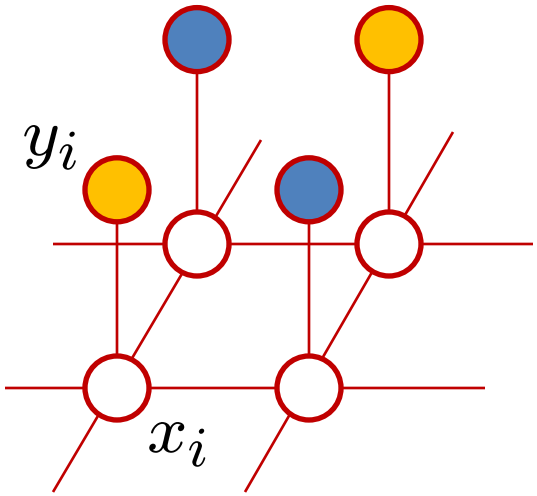
1. observations are partial
2. observations are noisy
3. innate nondeterministic

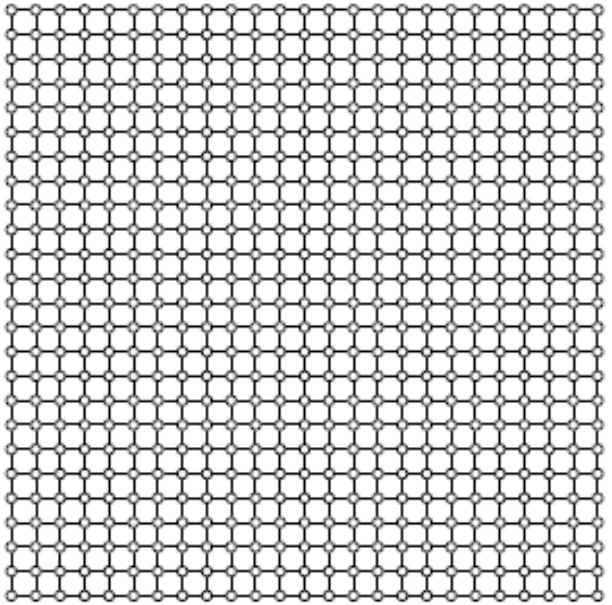
structure:

1. joint probability distribution P(A,B)
2. posterior distribution P(A|B = b)
3. conditional independence and factorization

應用範例: image de-noising

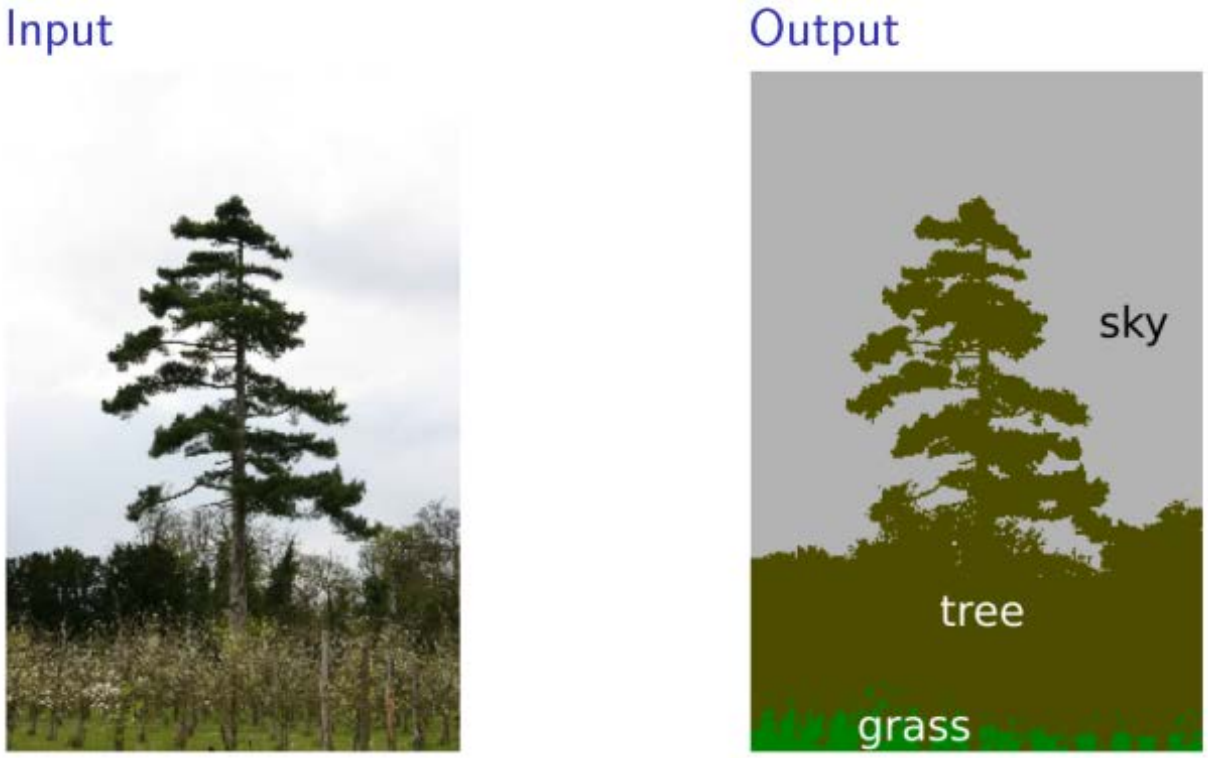
unknown noise-free binary values $x_i \in \{-1, +1\}$ $y_i \in \{-1, +1\}$ are binary pixel values of the observed noisy image





應用範例: image labeling

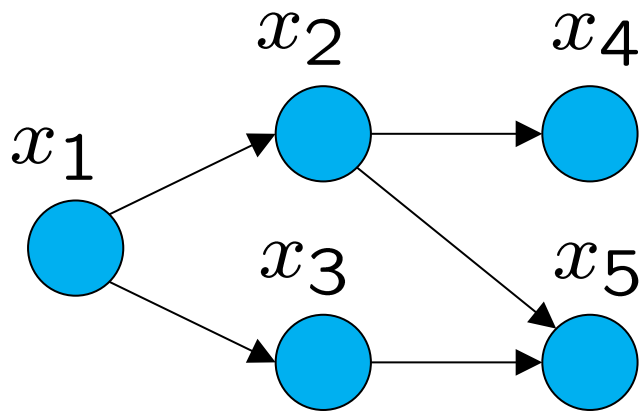
“Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials”, Krahenbuhl and Koltun



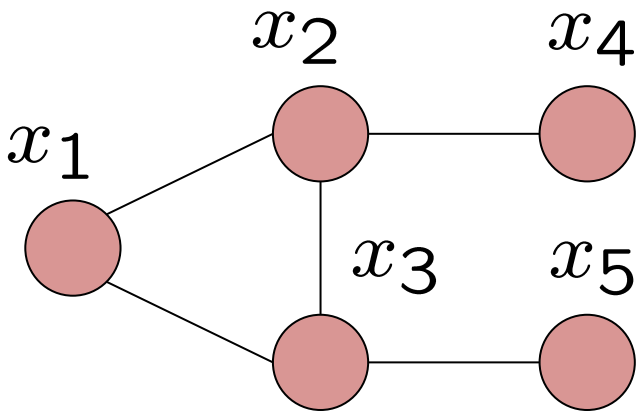
Graphical models

nodes: random variables

links: probabilistic constraints between variables



Bayesian network



Markov network

- Bayesian networks
  - directed acyclic graphs (DAGs)
  - conditional probability distributions (CPDs)
  - decompose the distribution as a product of CPDs

- Markov networks
  - undirected graphs
  - cliques (complete subgraphs) and factors
  - non-negativity: the only constraint on the parameters in the factor

Estimating joint distributions?

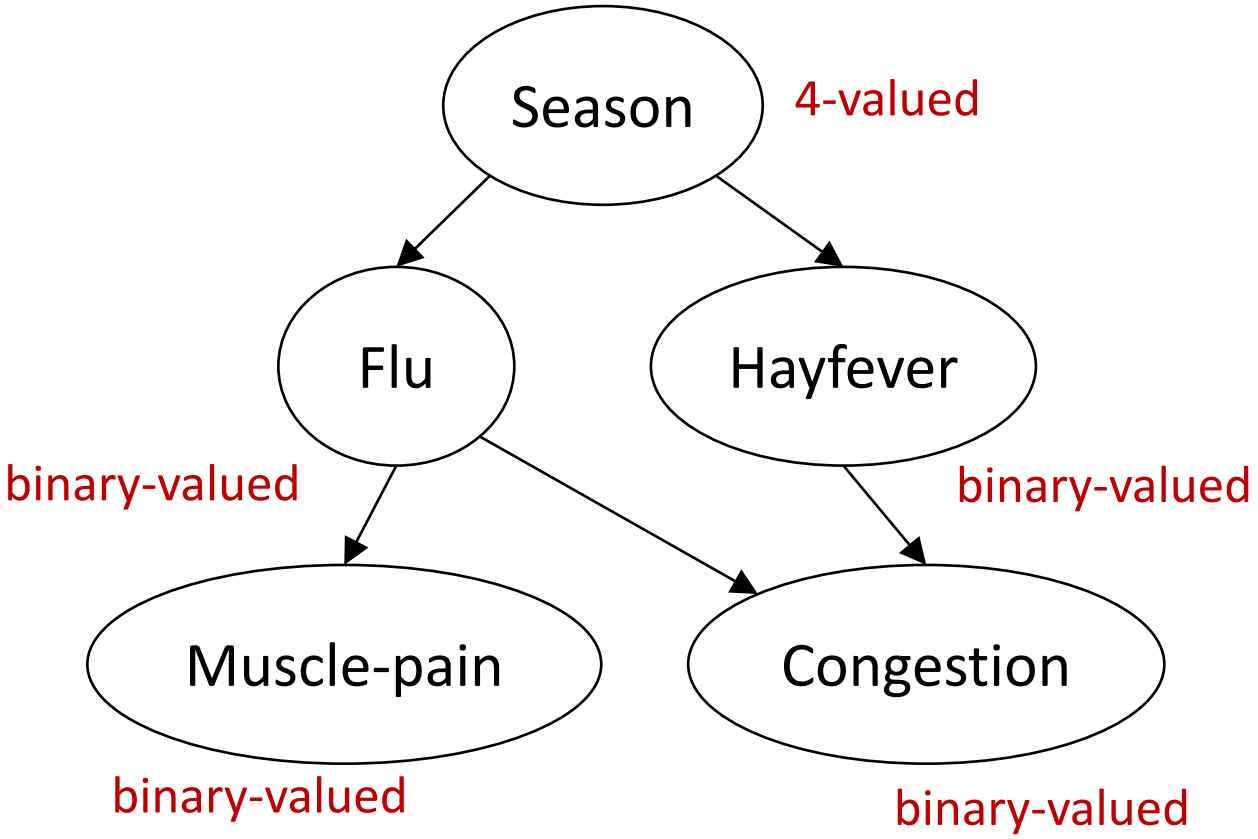
For me, estimating joint distributions is a bit like playing God.

You can't do everything!

Vladimir Vapnik

quote from "Graphical Models for Machine Learning and Digital Communication", Brendan J. Frey

Estimating joint distributions



modeling *P(S, F, H, C, M)*

4 × 2 × 2 × 2 × 2 = 64 configurations

稍微回憶一下機率

variables, states

joint probability $P(A, B)$

independence: $A \perp B|\emptyset$ if and only if $P(A, B) = P(A)P(B)$

marginal probabilities: $P(A) = \sum_B P(A, B)$, $P(B) = \sum_A P(A, B)$

table representations

AB	$P(A, B)$
$a^0 b^0$	0.08
$a^1 b^0$	0.20
$a^2 b^0$	0.12
$a^0 b^1$	0.12
$a^1 b^1$	0.30
$a^2 b^1$	0.18

joint

Table representations

joint

AB	$P(A, B)$
$a^0 b^0$	0.08
$a^1 b^0$	0.20
$a^2 b^0$	0.12
$a^0 b^1$	0.12
$a^1 b^1$	0.30
$a^2 b^1$	0.18

marginal

A	$P(A)$
a^0	0.2
a^1	0.5
a^2	0.3

B	$P(B)$
b^0	0.4
b^1	0.6

Conditional probabilities

conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A,B)}{\sum_A P(A,B)}$$

conditional independence

$$A \perp B \mid Z \qquad P(A|B,Z) = P(A|Z)$$

$$P(B,Z) > 0, \quad A \perp B \mid Z \qquad P(A,B|Z) = P(A|Z)P(B|Z)$$

table representation

	b^0	b^1
a^0	0.2	0.2
a^1	0.5	0.5
a^2	0.3	0.3

AB	$P(A,B)$
a^0b^0	0.08
a^1b^0	0.20
a^2b^0	0.12
a^0b^1	0.12
a^1b^1	0.30
a^2b^1	0.18

Evidence $P(A|B = b^0)$

conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A,B)}{\sum_A P(A,B)}$$

conditional independence

$$A \perp B \mid Z \qquad P(A|B,Z) = P(A|Z)$$

$$P(B,Z) > 0, \quad A \perp B \mid Z \qquad P(A,B|Z) = P(A|Z)P(B|Z)$$

table representation

	b^0	b^1
a^0	0.2	0.2
a^1	0.5	0.5
a^2	0.3	0.3

AB	$P(A,B)$
a^0b^0	0.08
a^1b^0	0.20
a^2b^0	0.12
a^0b^1	0.12
a^1b^1	0.30
a^2b^1	0.18

Factorization

chain rule

$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$

large factor → large table (不喜歡)

conditional independence 有助於簡化 factors

Factorization and graphs

directed graphs

- Bayesian networks
  - d-separation
  - parent-child
  - causality

undirected graphs

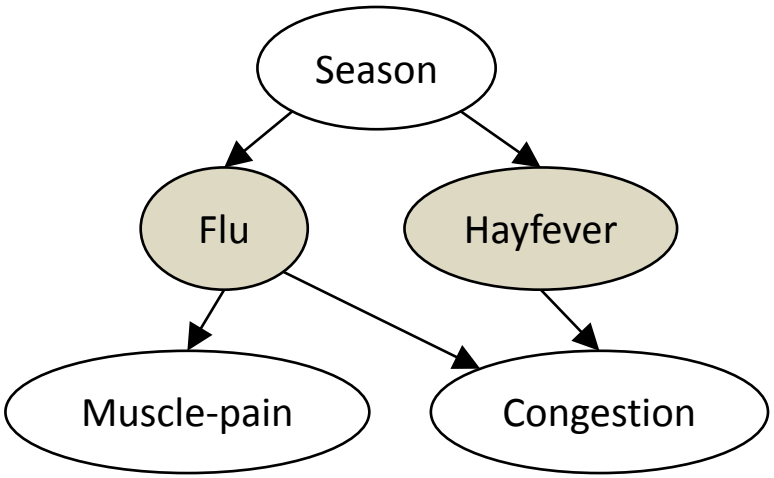
- Markov networks
  - blanket
  - neighbors
  - clique

Conditional independence

Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ be sets of random variables. $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in a distribution P if $P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z})$ is equal to $P(\mathbf{X} = \mathbf{x} | \mathbf{Z} = \mathbf{z}) P(\mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z})$ for all values $\mathbf{x} \in Val(\mathbf{X})$, $\mathbf{y} \in Val(\mathbf{Y})$, $\mathbf{z} \in Val(\mathbf{Z})$: the distribution P satisfies $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$.

independences: $(F \perp H | S)$, $(H \perp \{F, M\} | S)$, $(C \perp \{S, M\} | F, H)$, $(M \perp \{S, H, C\} | F)$

factorization:
$P(S, F, H, C, M) = P(S) P(F|S) P(H|S) P(C|F, H) P(M|F)$



Bayesian networks

- directed acyclic graphs (DAGs)
- joint distribution → factorization of conditional probability distributions (CPDs)

$P(X_i|\mathbf{Pa}_{X_i})$ where $\mathbf{Pa}_{X_i}$ are parents of X_i in the graph

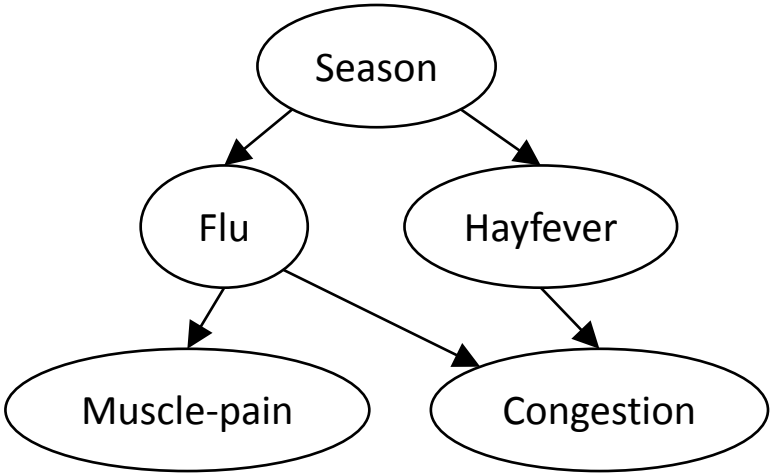
factorization:

$$P_{\mathcal{B}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}_{X_i})$$

Bayesian networks

A Bayesian network is a pair $(\mathcal{G}, \theta_{\mathcal{G}})$ where $P_{\mathcal{B}}$ factorizes over $\mathcal{G}$ and where $P_{\mathcal{B}}$ is specified as set of CPDs associated with $\mathcal{G}$'s nodes, denoted $\theta_{\mathcal{G}}$.

- example:
  $P(S, F, H, C, M) = P(S)P(F|S)P(H|S)P(C|F, H)P(M|F)$



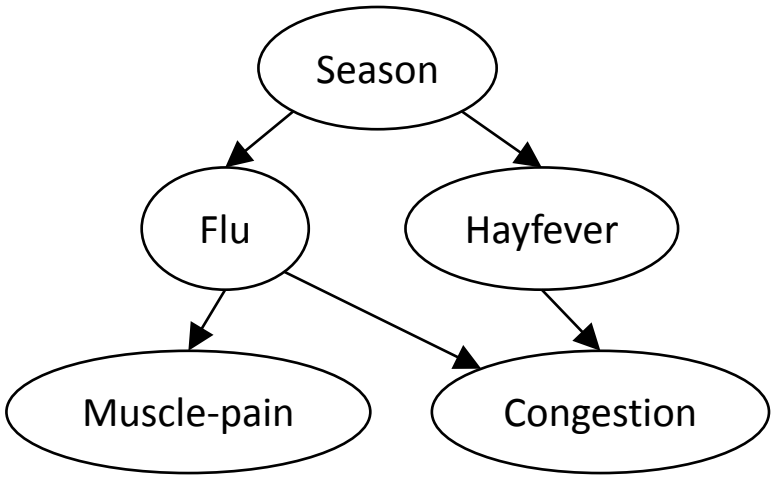
Conditional independence assumptions in Bayesian networks

For each variable X_i, we have that

$$(X_i \perp NonDescendants_{X_i} \mid \mathbf{Pa}_{X_i})$$

- example:

$$(F \perp H|S), (H \perp \{F, M\}|S), (C \perp \{S, M\}|F, H), (M \perp \{S, H, C\}|F)$$



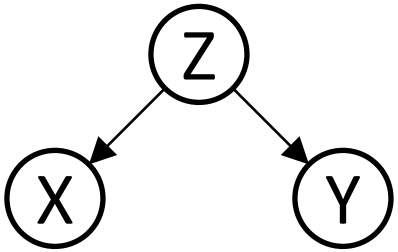
Flow of influence

Consider a simple three-node path X – Z – Y. If influence can flow from X to Y via Z, we say that the path X – Z – Y is active.

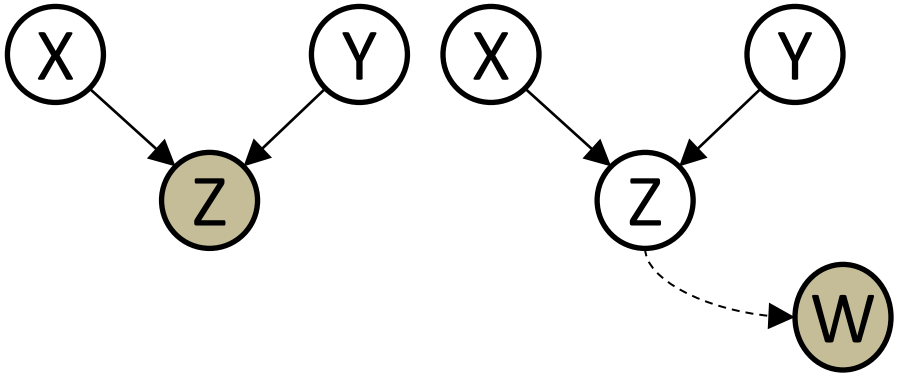
causal path

evidential path

common causal



common effect



Flow of influence

- causal path $X \rightarrow Z \rightarrow Y$: active if and only if Z is NOT observed
- evidential path $X \leftarrow Z \leftarrow Y$: active if and only if Z is NOT observed
- common cause $X \leftarrow Z \rightarrow Y$: active if and only if Z is NOT observed
- **common effect $X \rightarrow Z \leftarrow Y$: active if and only if either Z or one of Z's descendants is observed. (v-structure)**

Active paths

Let $\mathcal{G}$ be a BN structure, and $X_1 - \ldots - X_n$ be a path in $\mathcal{G}$. Let $\mathbf{E}$ be a subset of nodes of $\mathcal{G}$. The path $X_1 - \ldots - X_n$ is active given evidence $\mathbf{E}$ if

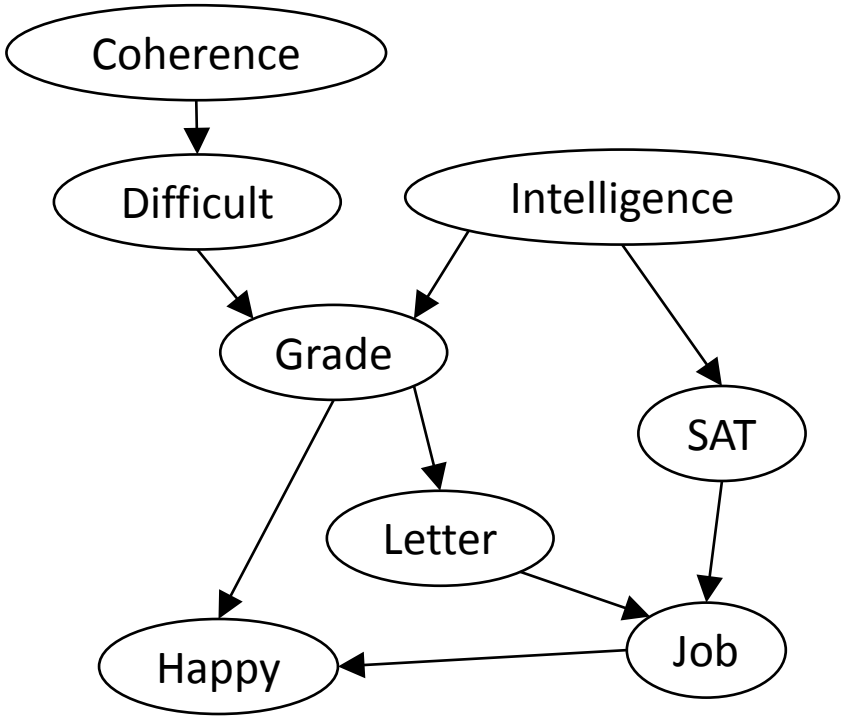
- whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants is in $\mathbf{E}$;
- no other node along the path is in $\mathbf{E}$.

Directed separation (d-separation)

Definition

Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ be three sets of nodes in $\mathcal{G}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$, denoted d-sep$_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$, if there is no active path between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

Independence and factorization in BN



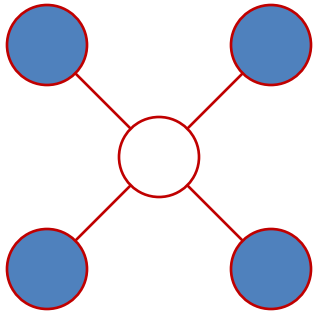
example from PGM, Koller and Friedman

It is true that d-sep$(D, J|L, I)$, but not d-sep$(D, I|L)$, d-sep$(D, J|L)$ and d-sep$(D, J|L, H, I)$.

factorization: $P(C, D, I, G, S, L, J, H) =$
$P(C)P(D|C)P(I)P(G|D, I)P(S|I)P(L|G)P(J|L, S)P(H|G, J)$

Markov networks

- Markov random fields (MRF)
- undirected graphs
  - Nodes: variables
  - Links: connect a pair of nodes



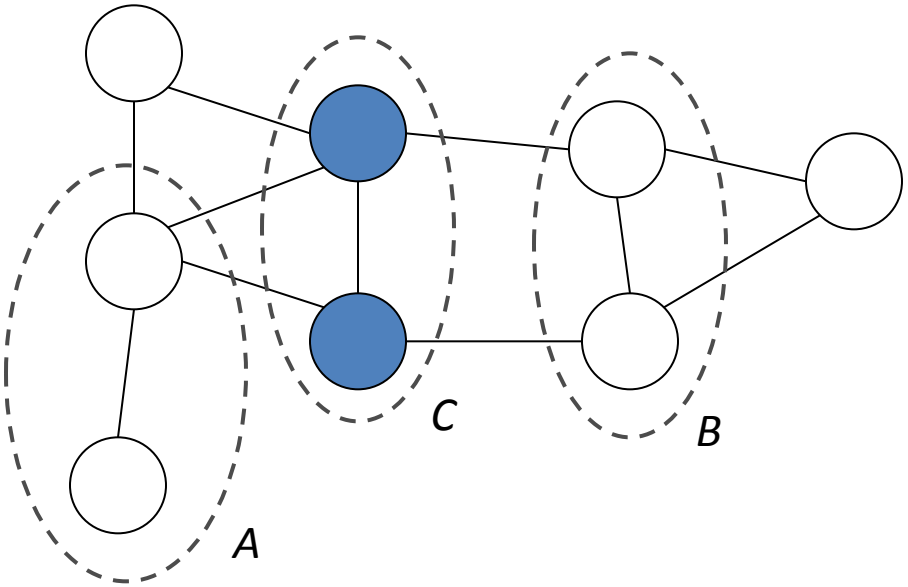
- specify a factorization and a set of conditional independence relations for the joint distribution of the random variables

Conditional independence properties in MRF

- consider all possible paths that connect nodes in set *A* to nodes in set *B*
  - if all such paths pass through one or more nodes in set *C*, then all such paths are 'blocked' and so the conditional independence property holds

$$A \perp B \mid C$$

$$P(A, B|\mathrm{C}) = P(A|\mathrm{C})P(B|\mathrm{C})$$



Factorization properties

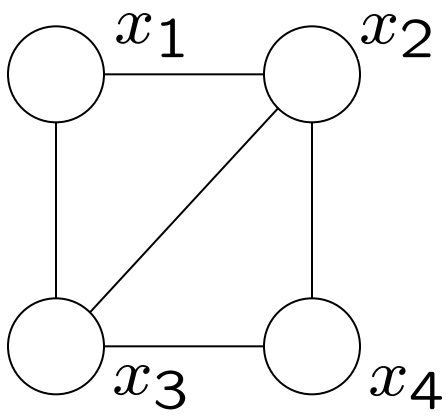
- expressing the joint distribution as a product of functions defined over sets of variables that are *local* to the graph

consider two nodes x_i and x_j that are not connected by a link

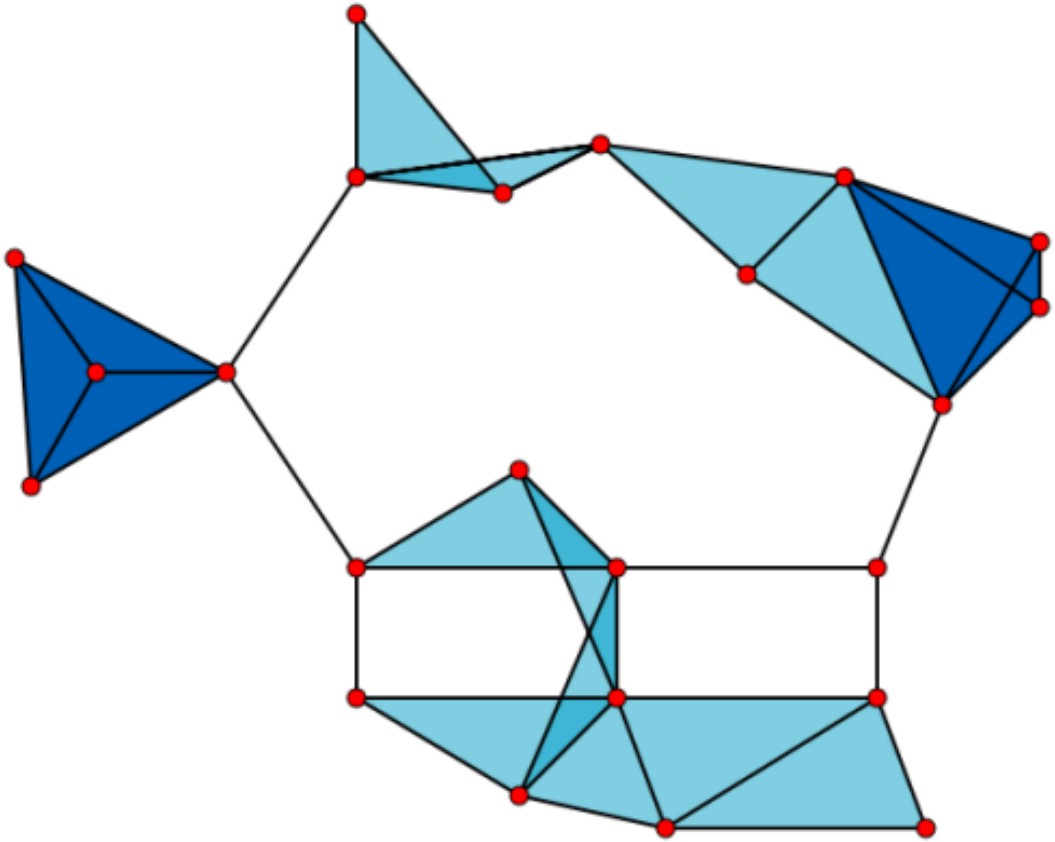
$$p(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,j\}})p(x_j|\mathbf{x}_{\setminus\{i,j\}})$$

Clique

- a complete subgraph
  - A subset of the nodes in a graph such that there exists a link between every pair of nodes in the subset
- a *maximal clique* is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique



Maximal cliques



[5] graph created by David Eppstein (Wikimedia Commons)

Potential functions

- the joint distribution can be written as a product of potential functions over the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$$

C : a clique

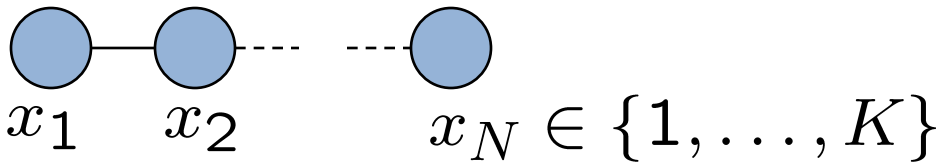
$\mathbf{x}_C$: the set of variables in clique C

$\psi_C(\mathbf{x}_C)$: a potential function over C

$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$:
the partition function for normalization

Computational limitation

Consider a model with N discrete nodes each having K states. Then the evaluation of the normalization term involves summing over K^N states and so is exponential in the size of the model.



$$Z = \sum_{\mathbf{x}} \prod_{C} \psi_C(\mathbf{x}_C)$$

Strictly positive potential functions

- express the potential functions as exponentials

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

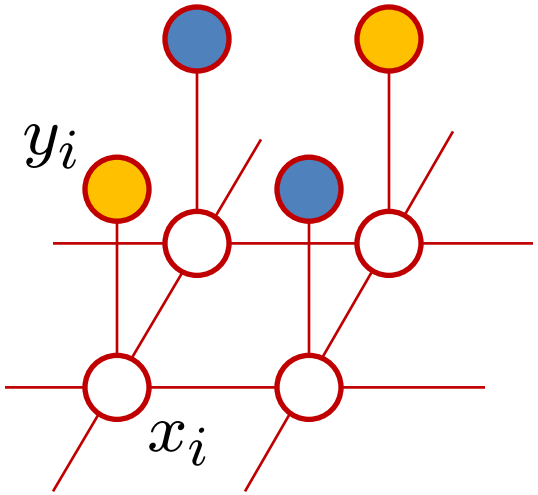
$E(\mathbf{x}_C)$ is called an *energy funciton*

- the joint distribution is defined as the product of potentials, and so the total energy is obtained by adding the energies of each of the maximal cliques

MRF modes as binary pixels

x_i is a binary variable denoting the state of pixels i in the unknown underlying image

y_i denotes the corresponding value of pixel i in the observed image



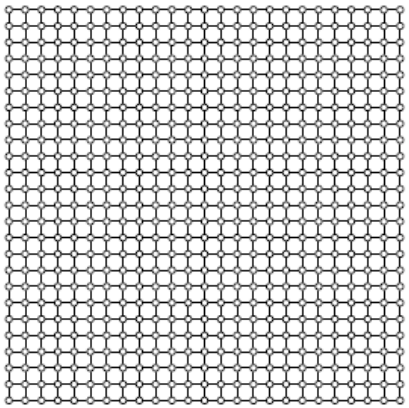
cliques?

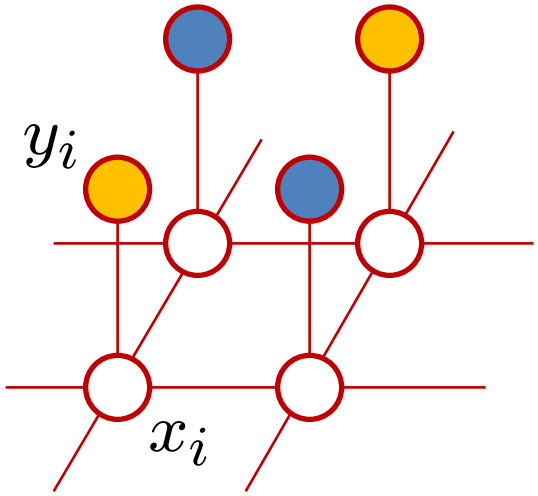
應用範例: image de-noising

unknown noise-free binary values $x_i \in \{-1, +1\}$

$y_i \in \{-1, +1\}$ are binary pixel values of the observed noisy image

- noise model
  - E.g., flipping the sign of the pixels with probability 10%





Joint probability

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(\mathbf{x}_C)$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \prod_{\{i,j\}} \psi(x_i, x_j) \prod_{i} \phi(x_i, y_i)$$

state

noisy image

state-state compatibility function

neighboring state nodes

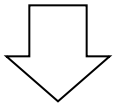
image-state compatibility function

local observations

Energy functions

- we need to choose energy functions for the cliques
  - a suitable energy function should express the relations among the nodes of a cliques
  - E.g.,

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$



$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

minimizing energy = maximizing probability

How to minimize the energy function?

Iterated conditional modes (ICM)

- Coordinate-wise gradient descent
- Not guaranteed to find the global minimum

- inference (下次上課)

Iterated conditional modes (ICM)

1. Initialize the variables $\{x_i\}$ by simply setting $x_i = y_i$ for all i;

2. Take one node x_j at a time and evaluate the total energy for the two possible states $x_j = +1$ and $x_j = -1$, keeping all other node variables fixed;

3. Set x_j to which ever state has the lower energy;

4. Repeat the update for another site, and so on, until some suitable stopping criterion is satisfied.

ICM example

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

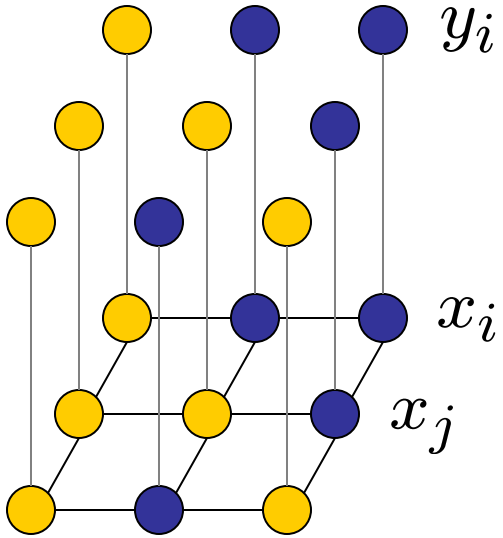
$\beta = 1.0$

$\eta = 2.1$

$h = 0$

-1

+1



Factorization and graphs

directed graphs

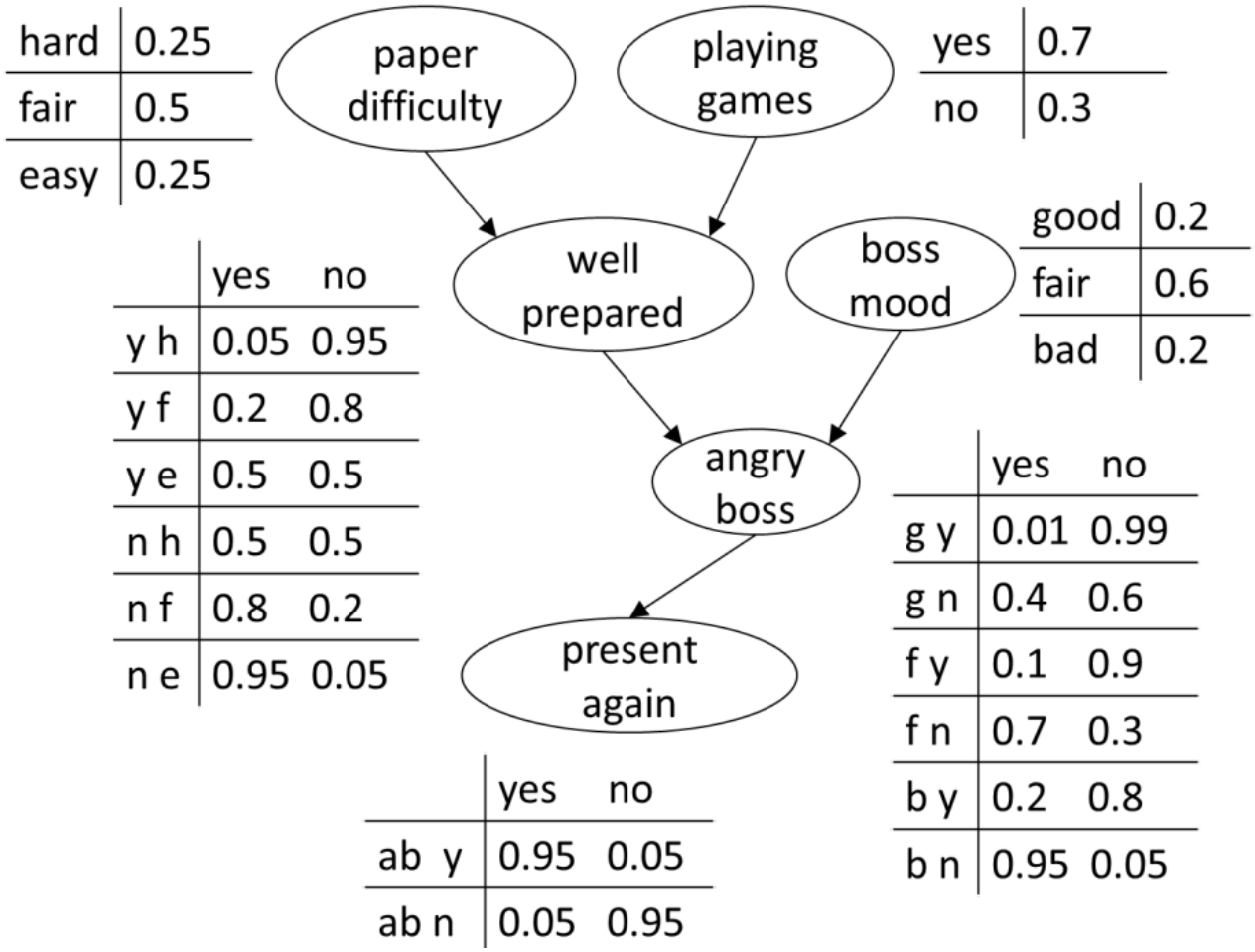
- Bayesian networks
  - d-separation
  - parent-child
  - causality

undirected graphs

- Markov networks
  - blanket
  - neighbors
  - clique

例子一: Bayesian network

- tool: `http://reasoning.cs.ucla.edu/samiam/` **Samlam**
- model:

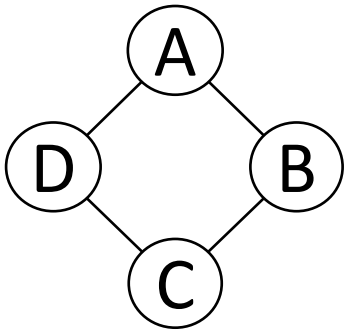


Summary

- graphical models 通常分成哪兩類?

- graphical models 好處?

Markov networks



- undirected graphs
- cliques (complete subgraphs)

factor: a function from $Val(\mathbf{D})$ to $\mathbb{R}^+$, where $\mathbf{D}$ is a set of random variables

$$P(A,B,C,D) = \frac{1}{Z}\tilde{P}(A,B,C,D) \text{ where}$$
$$\tilde{P}(A,B,C,D) = \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(A,D) \text{ and}$$
$$Z = \sum_{A,B,C,D} \tilde{P}(A,B,C,D)$$

Definition

Let $\mathcal{H}$ be a Markov network structure. A distribution $P_{\mathcal{H}}$ factorizes over $\mathcal{H}$ if it is associated with

- a set of subsets $\mathbf{D}_1, \ldots, \mathbf{D}_m$, where each $\mathbf{D}_i$ is a complete subgraph of $\mathcal{H}$;
- factors $\phi_1(\mathbf{D}_1), \ldots, \phi_m(\mathbf{D}_m)$,

such that

$$P_{\mathcal{H}}(X, \ldots, X_n) = \frac{1}{Z}\widetilde{P}_{\mathcal{H}}(X_1, ..., X_n),$$

where

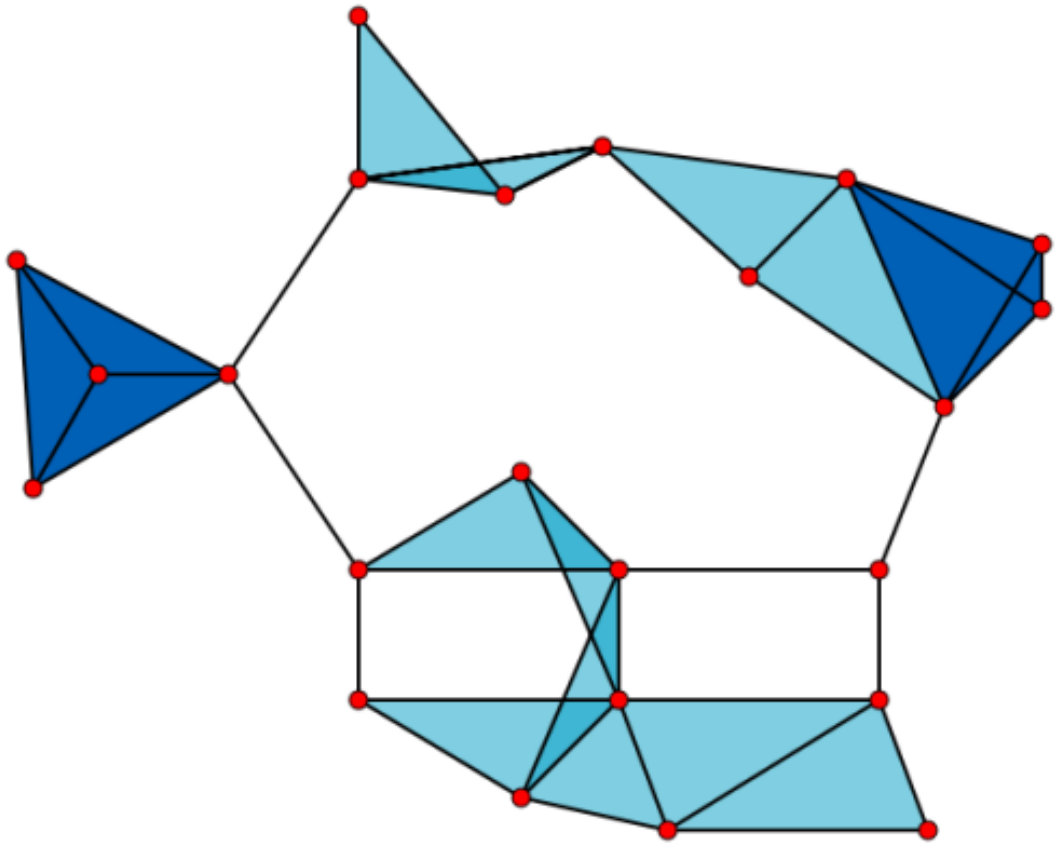
$$\widetilde{P}_{\mathcal{H}}(X_1, \ldots, X_n) = \phi_1(\mathbf{D}_1) \times \phi_2(\mathbf{D}_2) \times \ldots \times \phi_m(\mathbf{D}_m)$$

is an unnormalized measure and

$$Z = \sum_{X_1, \ldots, X_n} \widetilde{P}_{\mathcal{H}}(X_1, \ldots, X_n)$$

is a normalizing constant called the partition function. A distribution P that factorizes over $\mathcal{H}$ is also called a Gibbs distribution over $\mathcal{H}$.

Maximal cliques



[5] graph created by David Eppstein (Wikimedia Commons)

Factor and energy function

We can rewrite a factor $\phi(\mathbf{D})$ as

$$\phi(\mathbf{D}) = \exp(-\epsilon(\mathbf{D}))$$

where $\epsilon(\mathbf{D}) = -\ln \phi(\mathbf{D})$ is often called an energy function.

In this logarithmic representation, we have that

$$P_{\mathcal{H}}(X_1, \ldots, X_n) \propto \exp\left[-\sum_{i=1}^{m} \epsilon_i(\mathbf{D}_i)\right].$$

Independencies in Markov networks

Let $\mathcal{H}$ be an undirected graph. Then for each node $X \in \mathbf{X}$, the Markov blanket of X, denoted $\mathbf{N}_{\mathcal{H}}(X)$, is the set of neighbors of X in the graph. We define the local Markov independencies associated with $\mathcal{H}$ to be

$$\mathcal{I}(\mathcal{H}) = \{(X \perp \mathbf{X} - \{X\} - \mathbf{N}_{\mathcal{H}}(X) \mid \mathbf{N}_{\mathcal{H}}(X)) : X \in \mathbf{X}\}.$$

For example, $(A \perp C|\{B, D\})$, $(B \perp D|\{A, C\})$.

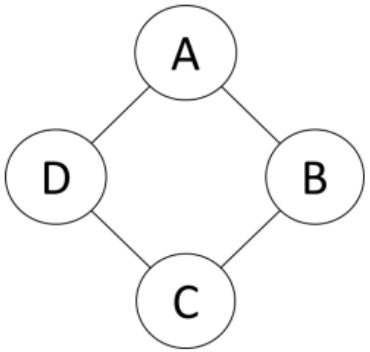
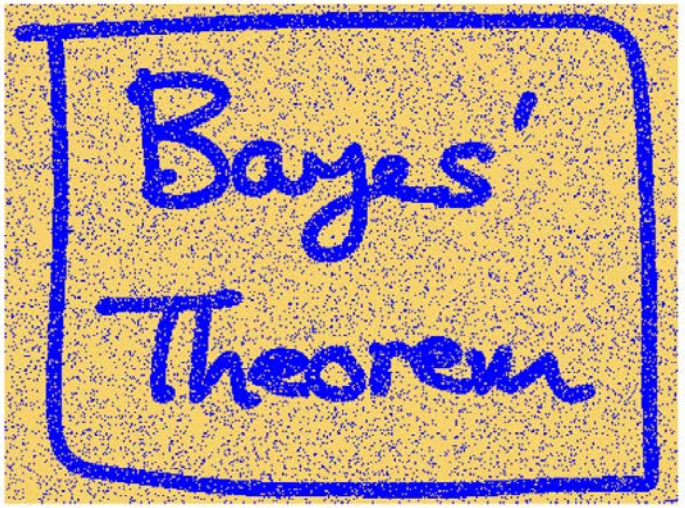
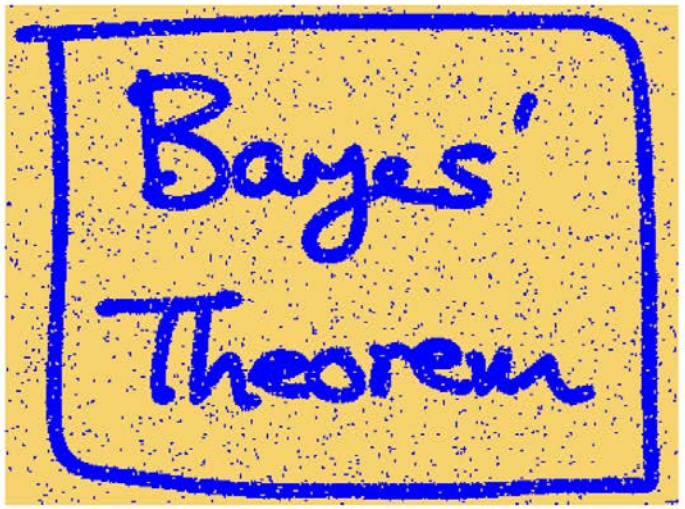
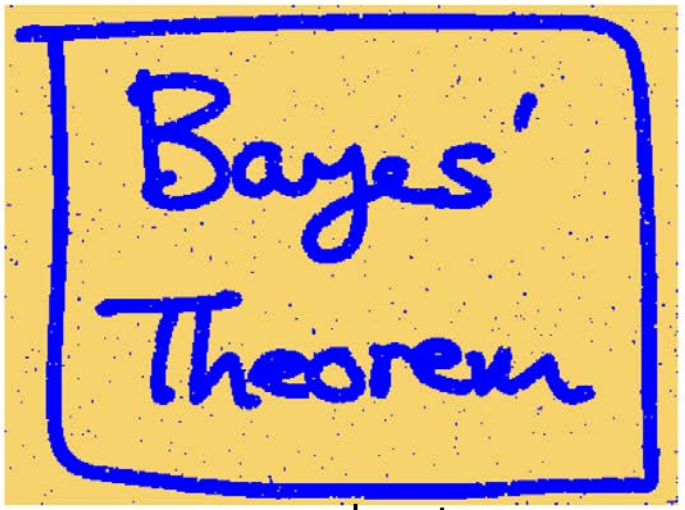


Image de-noising





iterated conditional modes



graph-cut