

LTAT.02.004 MACHINE LEARNING II

Multivariate normal distribution

Direct applications

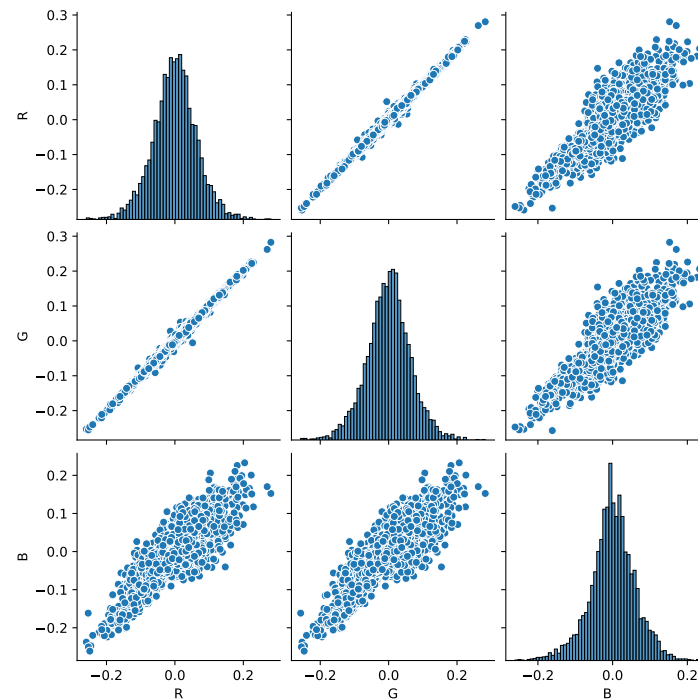
Sven Laur
University of Tartu

Motivating examples

Filtering and smoothing

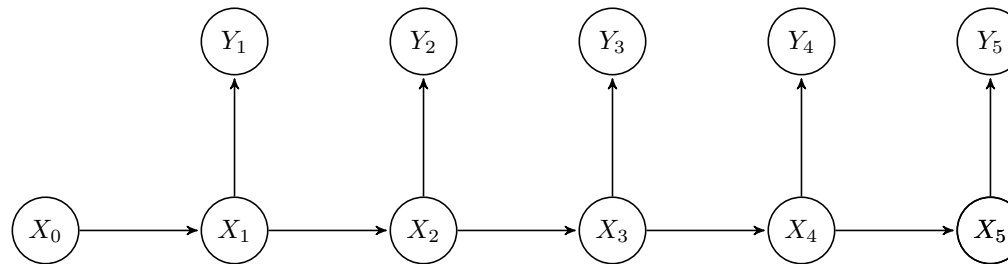
Prediction of vector values

Prediction errors of different vector components can be correlated.



As a result combined model can outperform coordinatewise predictions.

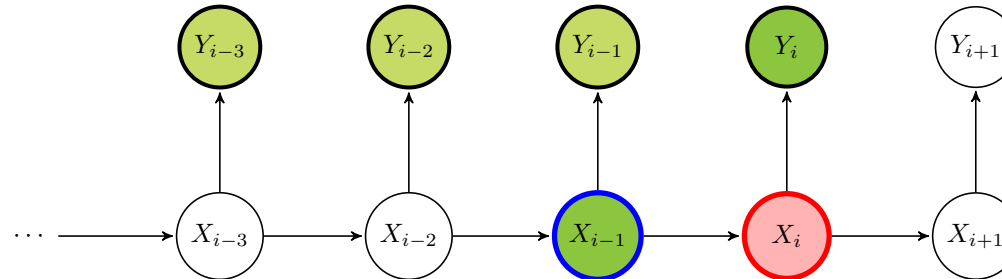
Sensor fusion with Hidden Markov Models



A standard problem in robotics or machine perception is following.

- ▷ Several sensors measure a physical system
- ▷ Measurements are observable as $\mathbf{y} \in \mathbb{R}^p$.
- ▷ Physical system has an hidden state $\mathbf{x} \in \mathbb{R}^n$.
- ▷ Physical system evolves linearly $\mathbf{x}_{i+1} = A\mathbf{x}_i + \mathbf{w}_i$.
- ▷ Measurements are linear from the state $\mathbf{y}_i = C\mathbf{x}_i + \mathbf{v}_i$.
- ▷ Distribution of error terms \mathbf{v}_i and \mathbf{w}_i is known.
- ▷ Error terms \mathbf{v}_i and \mathbf{w}_i are independently drawn.

Kalman filter



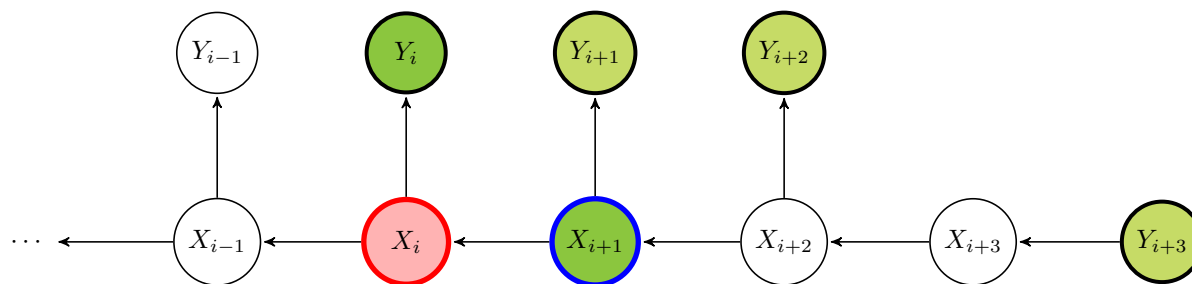
As before we can consider the prior and filter densities

$$\pi[\mathbf{x}_i] = p[\mathbf{x}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}]$$

$$f[\mathbf{x}_i] = p[\mathbf{x}_i | \mathbf{y}_1, \dots, \mathbf{y}_i] \propto \pi[\mathbf{x}_i] \cdot p[\mathbf{y}_i | \mathbf{x}_i]$$

A similar update logic assures that both distributions are normal distributions and that we can only compute the parameters of these normal distributions.

Smoothing and reverse Hidden Markov Model

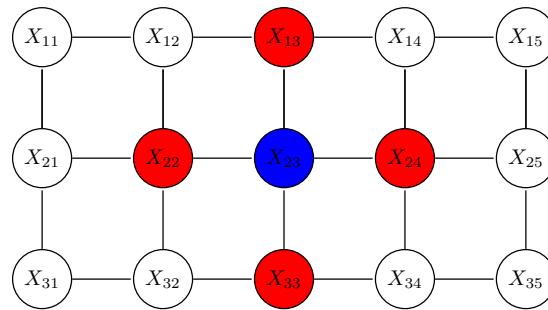


- ▷ We need likelihoods $\lambda[\mathbf{x}_i] = p[\mathbf{y}_{i+1}, \dots, \mathbf{y}_n | \mathbf{x}_i]$ for the smoothing.
- ▷ Likelihood propagation formula is analogous to the prior propagation.
- ▷ We can define a reverse HMM such that the prior $\pi^*[\mathbf{x}_i] \propto \lambda[\mathbf{x}_i]$.
- ▷ The resulting HMM has reversed dynamics.
- ▷ It turns out that all likelihoods $\lambda[\mathbf{x}_i]$ are normal distributions.
- ▷ The posterior as product $\pi[\mathbf{x}_i] \cdot \lambda[\mathbf{x}_i] \cdot p[\mathbf{y}_i | \mathbf{x}_i]$ is also a normal distribution.

Motivating examples

Markov fields

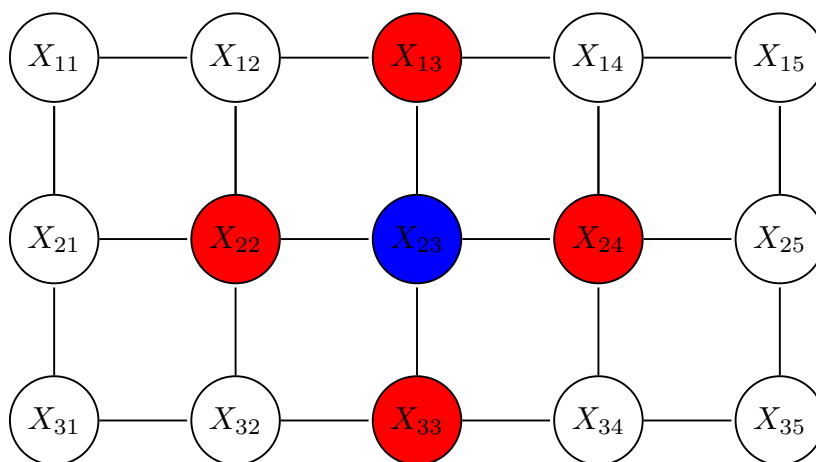
Background model for digital images



In most images intensity of pixel is influenced only by its neighbours:

- ▷ For simple textures the neighbourhood consist of four adjacent pixels.
- ▷ For complex textures the the neighbourhood contains much more pixels.
- ▷ For homogenous textures the conditional probabilities are universal.
 - ◇ Generative repetitive patterns for textile and grass
- ▷ For complex patterns conditional probabilities can be location dependent.
 - ◇ Generative patterns for human faces and fashion accessories

Random Markov Fields



Definition. Markov random field is specified by undirected graph connecting random variables X_1, X_2, \dots such that for any node X_i

$$\Pr [x_i | (x_j)_{j \neq i}] = \Pr [x_i | (x_j)_{j \in \mathcal{N}(X_i)}]$$

where the set of neighbours $\mathcal{N}(X_i)$ is also known as *Markov blanket* for X_i .

Hammersley-Clifford theorem

The probability of an observation $\mathbf{x} = (x_1, x_2, \dots)$ generated by a Markov random field can be expressed in the form

$$\Pr[\mathbf{x}] = \frac{1}{Z(\omega)} \cdot \exp \left(- \sum_{c \in \text{MaxClique}} \Psi_c(\mathbf{x}_c, \omega) \right)$$

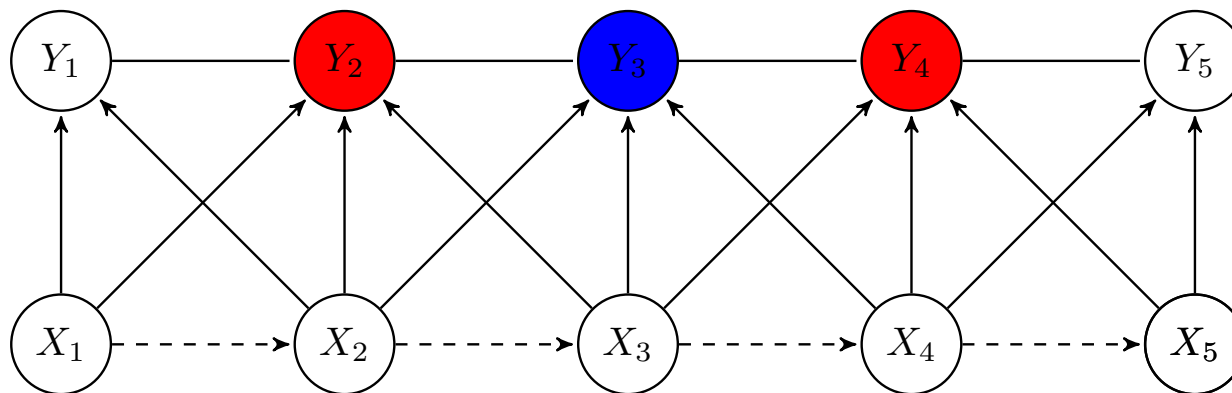
where

- ▷ $Z(\omega)$ is a normalising constant
- ▷ MaxClique is the set of maximal cliques in the Markov random field
- ▷ Ψ_c is defined on the variables in the clique c

The formula implies that the distribution belongs to the exponential family.

- ▷ Multivariate normal distribution belongs to the exponential family

Conditional Random Fields

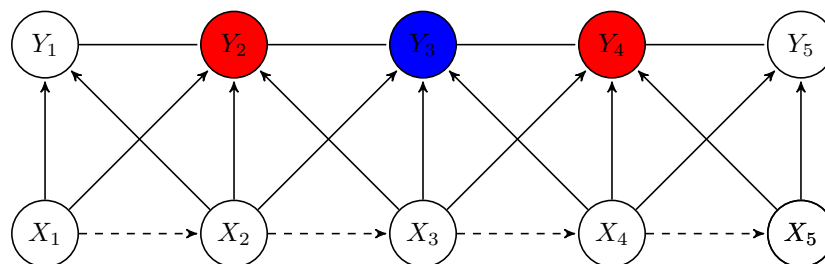


Definition. Let X_1, X_2, \dots and Y_1, Y_2, \dots be random variables. The entire process is conditional random field if random variables Y_1, Y_2, \dots conditioned for any sequence of observations x_1, x_2, \dots form a Markov random field

$$\Pr [y_i | (x_k)_{k=1}^{\infty}, (y_j)_{j \neq i}] = \Pr [y_i | (x_k)_{k=1}^{\infty}, (y_j)_{j \in \mathcal{N}(Y_i)}]$$

where the set of neighbours $\mathcal{N}(Y_i)$ is a *conditional Markov blanket* for Y_i .

Image segmentation and sequence labelling



- ▷ The input x is used to predict labels y_1, y_2, \dots
- ▷ A correct label sequence must satisfy possibly unknown restrictions.
- ▷ These restrictions are captured by conditional random random field.

Consequences of Hammersley-Clifford theorem

- ▷ Clique features Ψ_c can depend on $(y_i)_{i \in c}, (x_i)_{i=1}^{\infty}$
- ▷ Features can be defined as linear combination of vertex and edge features.
- ▷ A vertex feature looks only variable y_i associated with the vertex.
- ▷ An edge feature looks only variables y_i, y_j associated with the edge.

Markov fields
with
multivariate normal distributions

General form of the likelihood function

The celebrated Hammersley-Clifford theorem fixes the format in which the corresponding probability distribution must be sought:

$$p[\mathbf{x}|\omega] = \frac{1}{Z(\omega)} \cdot \exp \left(- \sum_{c \in \text{MaxClique}} \Psi_c(\mathbf{x}_c, \omega) \right)$$

where

- ▷ ω is a set of model parameters
- ▷ $Z(\omega)$ is a normalising constant
- ▷ MaxClique is the set of maximal cliques in the Markov random field
- ▷ Ψ_c is defined on the variables x_i in the clique c .

Multivariate normal distribution as likelihood

If individual sub-potentials $\Psi_c(\mathbf{x}_c, \omega)$ are quadratic forms then the energy

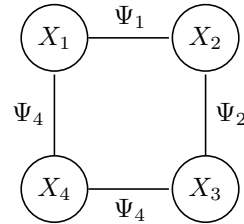
$$\Psi(\mathbf{x}) = \sum_{c \in \text{MaxClique}} \Psi_c(\mathbf{x}_c, \omega)$$

is also a quadratic form and thus $p[\mathbf{x}|\omega]$ is a multivariate normal distribution.

Sub-potentials are often fixed directly based on smoothness constraints

- ▷ Intensities have bounded variance: $\Psi_e = \delta^2 x_{ij}^2$.
- ▷ Intensity changes smoothly vertically: $\Psi_e = \beta(x_{i,j} - x_{i+1,j})^2$.
- ▷ Intensity changes smoothly horizontally: $\Psi_e = \alpha(x_{i,j} - x_{i,j+1})^2$.

Toy example



Sub-potentials corresponding four edges are:

$$\Psi_1(x_1, x_2) = \alpha_1(x_1 - x_2)^2 = \alpha_1 x_1^2 - 2\alpha_1 x_1 x_2 + \alpha_1 x_2^2$$

$$\Psi_2(x_2, x_3) = \alpha_2(x_2 - x_3)^2 = \alpha_2 x_2^2 - 2\alpha_2 x_2 x_3 + \alpha_2 x_3^2$$

$$\Psi_3(x_3, x_4) = \alpha_3(x_3 - x_4)^2 = \alpha_3 x_3^2 - 2\alpha_3 x_3 x_4 + \alpha_3 x_4^2$$

$$\Psi_4(x_4, x_1) = \alpha_4(x_4 - x_1)^2 = \alpha_4 x_4^2 - 2\alpha_4 x_4 x_1 + \alpha_4 x_1^2$$

Sub-potentials corresponding to four vertices are $\Psi_i^*(x_i) = \delta_i^2 x_i^2$

Resulting potential function

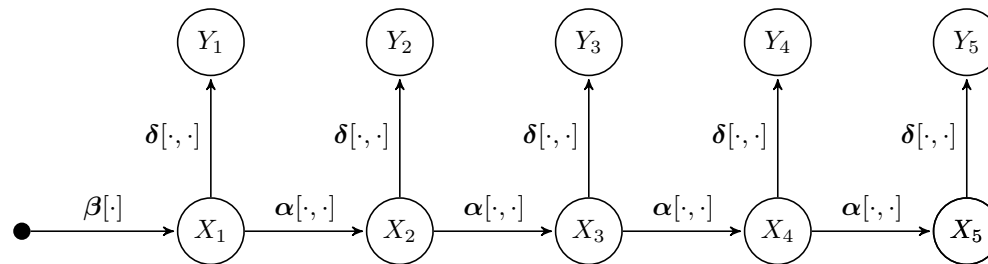
$$\Psi(\mathbf{x}) = \mathbf{x}^T \begin{pmatrix} \alpha_1 + \alpha_4 + \delta_1^2 & -\alpha_1 & 0 & -\alpha_4 \\ -\alpha_1 & \alpha_1 + \alpha_2 + \delta_2^2 & -\alpha_2 & 0 \\ 0 & -\alpha_2 & \alpha_2 + \alpha_3 + \delta_3^2 & -\alpha_3 \\ -\alpha_4 & 0 & -\alpha_3 & \alpha_3 + \alpha_4 + \delta_4^2 \end{pmatrix} \mathbf{x}$$

and thus the covariance matrix Σ and mean $\boldsymbol{\mu}$ can be computed by matching the shape of the multivariate normal density

$$p[\mathbf{x}|\boldsymbol{\mu}, \Sigma] \propto \frac{1}{\sqrt{\det \Sigma}} \cdot \exp \left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Motivating examples

Hidden Markov Model



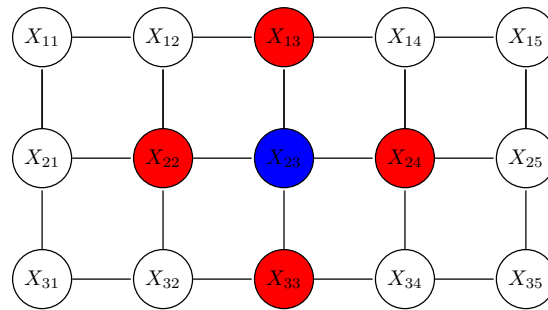
- ▷ Inference of Hidden Markov models requires a lot of data.
- ▷ Continuous distributions are rarely compatible with belief propagation.

$$\pi_{X_i}(\mathbf{x}_i) \propto \int_{\mathbf{x}_{i-1}} \alpha[\mathbf{x}_{i-1}, \mathbf{x}_i] \cdot \lambda_{i-1}^*(\mathbf{x}_{i-1}) \cdot \pi_{X_{i-1}}(\mathbf{x}_{i-1}) d\mathbf{x}_{i-1}$$

$$\lambda_{X_i}(x_i) \propto \int_{\mathbf{x}_{i+1}} \alpha[\mathbf{x}_i, \mathbf{x}_{i+1}] \cdot \lambda_i^*(\mathbf{x}_i) \cdot \lambda_{X_{i+1}}(\mathbf{x}_{i+1}) d\mathbf{x}_{i+1}$$

- ▷ Family of normal distributions is compatible with belief propagation.

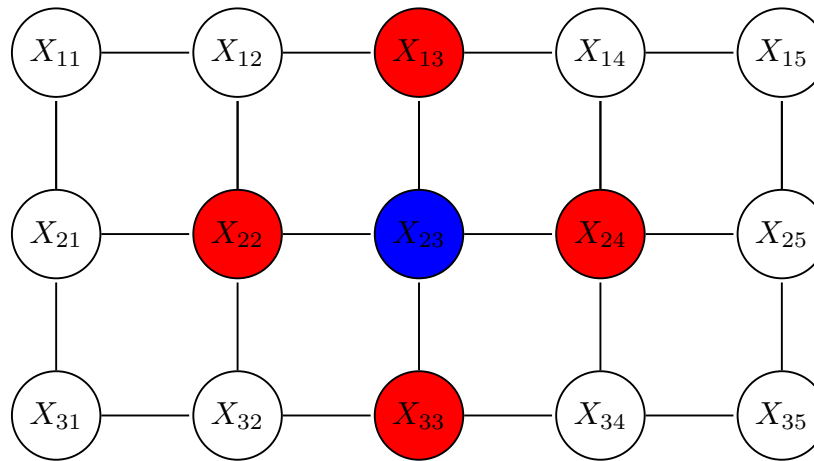
Background model for digital images



In most images intensity of pixel is influenced only by its neighbours:

- ▷ For simple textures the neighbourhood consist of four adjacent pixels.
- ▷ For complex textures the the neighbourhood contains much more pixels.
- ▷ For homogenous textures the conditional probabilities are universal.
 - ◇ Generative repetitive patterns for textile and grass
- ▷ For complex patterns conditional probabilities can be location dependent.
 - ◇ Generative patterns for human faces and fashion accessories

Random Markov Fields



Definition. Markov random field is specified by undirected graph connecting random variables X_1, X_2, \dots such that for any node X_i

$$\Pr [x_i | (x_j)_{j \neq i}] = \Pr [x_i | (x_j)_{j \in \mathcal{N}(X_i)}]$$

where the set of neighbours $\mathcal{N}(X_i)$ is also known as *Markov blanket* for X_i .

Hammersley-Clifford theorem

The probability of an observation $\mathbf{x} = (x_1, x_2, \dots)$ generated by a Markov random field can be expressed in the form

$$\Pr[\mathbf{x}] = \frac{1}{Z(\omega)} \cdot \exp \left(- \sum_{c \in \text{MaxClique}} \Psi_c(\mathbf{x}_c, \omega) \right)$$

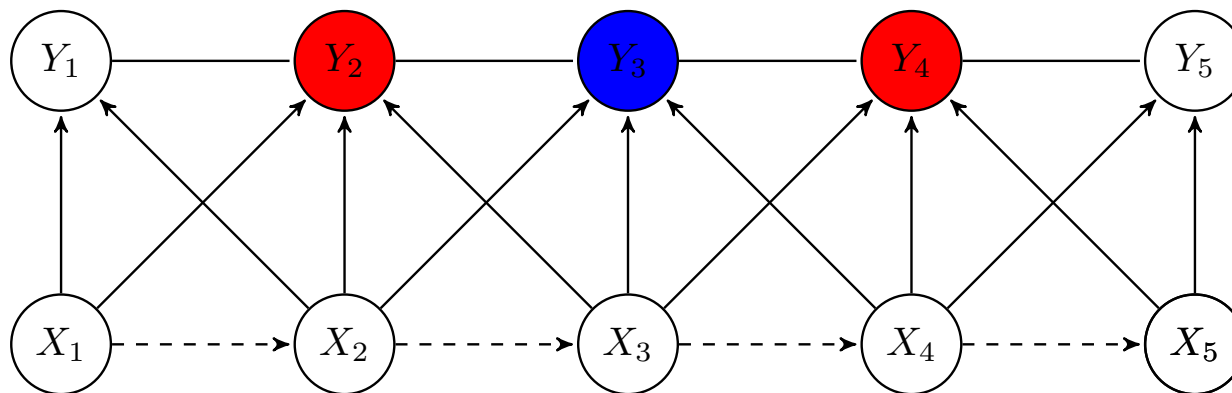
where

- ▷ $Z(\omega)$ is a normalising constant
- ▷ MaxClique is the set of maximal cliques in the Markov random field
- ▷ Ψ_c is defined on the variables in the clique c

The formula implies that the distribution belongs to the exponential family.

- ▷ Multivariate normal distribution belongs to the exponential family

Conditional Random Fields

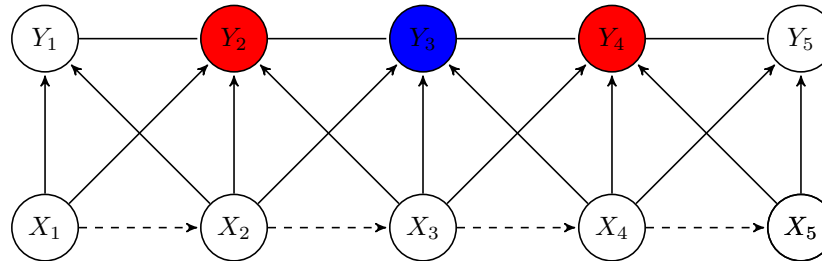


Definition. Let X_1, X_2, \dots and Y_1, Y_2, \dots be random variables. The entire process is conditional random field if random variables Y_1, Y_2, \dots conditioned for any sequence of observations x_1, x_2, \dots form a Markov random field

$$\Pr [y_i | (x_k)_{k=1}^{\infty}, (y_j)_{j \neq i}] = \Pr [y_i | (x_k)_{k=1}^{\infty}, (y_j)_{j \in \mathcal{N}(Y_i)}]$$

where the set of neighbours $\mathcal{N}(Y_i)$ is a *conditional Markov blanket* for Y_i .

Image segmentation and sequence labelling

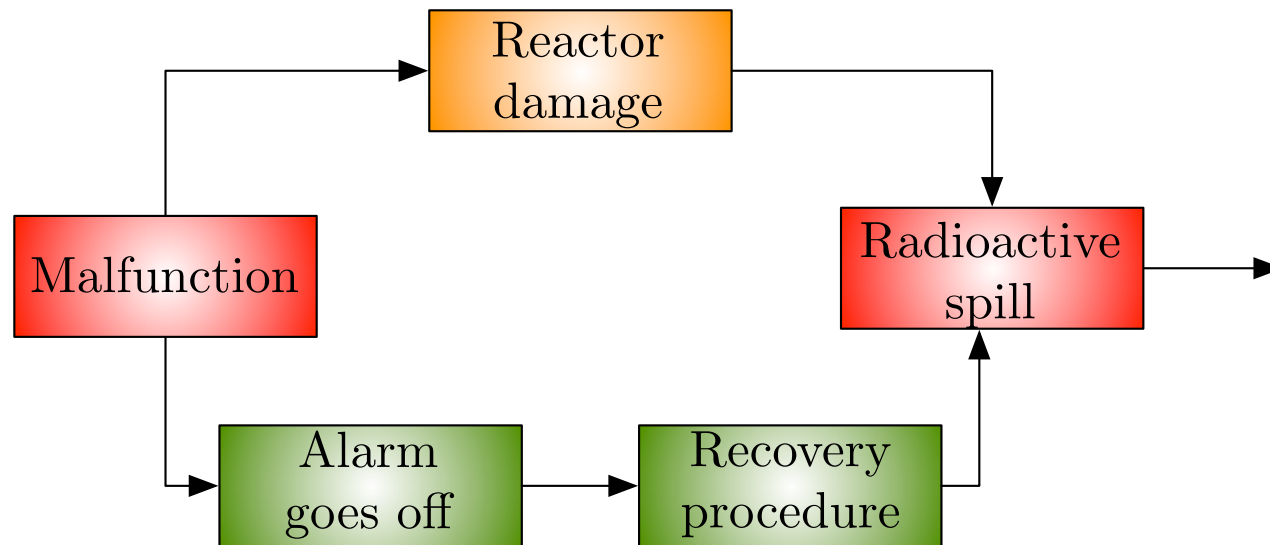


- ▷ The input x is used to predict labels y_1, y_2, \dots
- ▷ A correct label sequence must satisfy possibly unknown restrictions.
- ▷ These restrictions are captured by conditional random random field.

Consequences of Hammersley-Clifford theorem

- ▷ Clique features Ψ_c can depend on $(y_i)_{i \in c}, (x_i)_{i=1}^{\infty}$
- ▷ Features can be defined as linear combination of vertex and edge features.
- ▷ A vertex feature looks only variable y_i associated with the vertex.
- ▷ An edge feature looks only variables y_i, y_j associated with the edge.

Going beyond naive Bayesian models



Complex causal models are often defined through Bayesian networks

- ▷ A complex processes is first split into sub-events
- ▷ Direct causal dependencies between sub-events are detected
- ▷ Causation mechanisms are characterised with probability tables

Strength and weaknesses of Bayesian networks

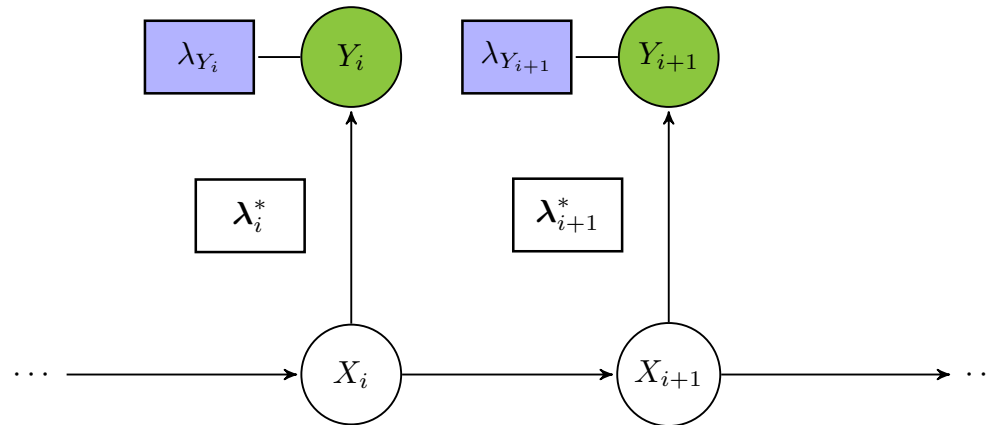
Strengths

- ▷ Bayesian networks are easy to interpret
- ▷ Bayesian networks are good for formalising fuzzy background knowledge
- ▷ Estimation of individual probability tables is tractable
- ▷ There are tools for doing inference with Bayesian networks

Weaknesses

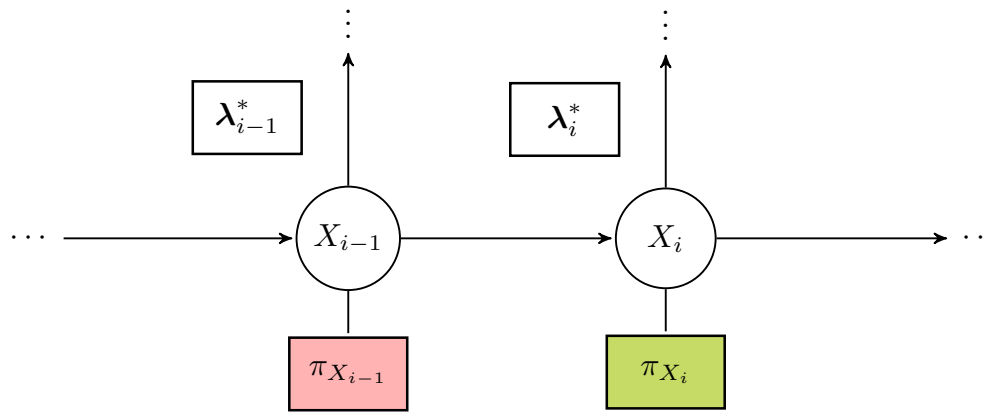
- ▷ You must know the causal structure of sub-events
- ▷ Identification of causal structure from data alone is very difficult
- ▷ It is notoriously difficult to model non-trivial causal dependencies
- ▷ Standard inference procedures often do not have closed solutions

Belief propagation. Initialisation



- ▷ We have a direct evidence $Y_i = y_i$ for each node Y_i .
- ▷ The likelihood vector is infinite and captured by $\lambda_{Y_i} = \delta_{y_i}$.
- ▷ The local likelihood $\lambda_i^*(x_i) = p[Y_i = y_i | x_i]$ is an infinite vector.
- ▷ The form $\mathbf{y}_i = C\mathbf{x}_i + \mathbf{v}_i$ assures that $\mathbf{y}_i | \mathbf{x}_i$ is normal distribution.
- ▷ The local likelihood λ_i^* has a finite description.

Prior propagation. Filtering

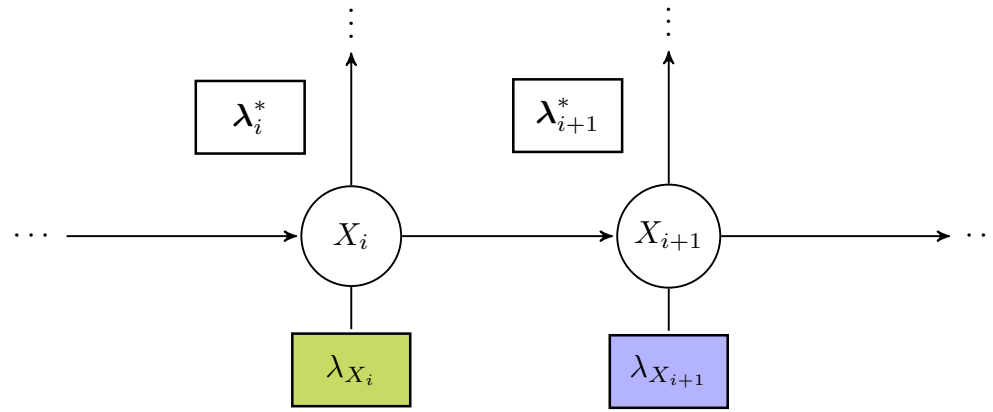


Prior propagation rule

$$\pi_{X_i}(\mathbf{x}_i) \propto \int_{\mathbf{x}_{i-1}} \alpha[\mathbf{x}_{i-1}, \mathbf{x}_i] \cdot \lambda_{i-1}^*(\mathbf{x}_{i-1}) \cdot \pi_{X_{i-1}}(\mathbf{x}_{i-1}) d\mathbf{x}_{i-1}$$

leads to a finite description because on the right is a normal distribution.

Likelihood propagation. Smoothing



Likelihood propagation rule

$$\lambda_{X_i}(x_i) \propto \int_{\mathbf{x}_{i+1}} \alpha[\mathbf{x}_i, \mathbf{x}_{i+1}] \cdot \lambda_{X_{i+1}}(\mathbf{x}_{i+1}) \cdot \lambda_i^*(\mathbf{x}_i) d\mathbf{x}_{i+1}$$

leads to a finite description because on the right is a normal distribution.