ST451 - Lent term
Bayesian Machine Learning

Kostas Kalogeropoulos

Graphical Models

Summary of last lecture

- Variational Inference: Approximate $\pi(\theta|y)$ with a family of distributions $q(\theta|y,\phi)$. Minimise the KL-divergence between π and q wrt ϕ by maximising ELBO.

- Mean Field Approximation: Assumes $q(\theta|y,\phi) = \prod_{i=1}^{p} q(\theta_i|y,\phi_i)$. Can optimise by setting each $q(\theta_i|y,\phi_i)$ to

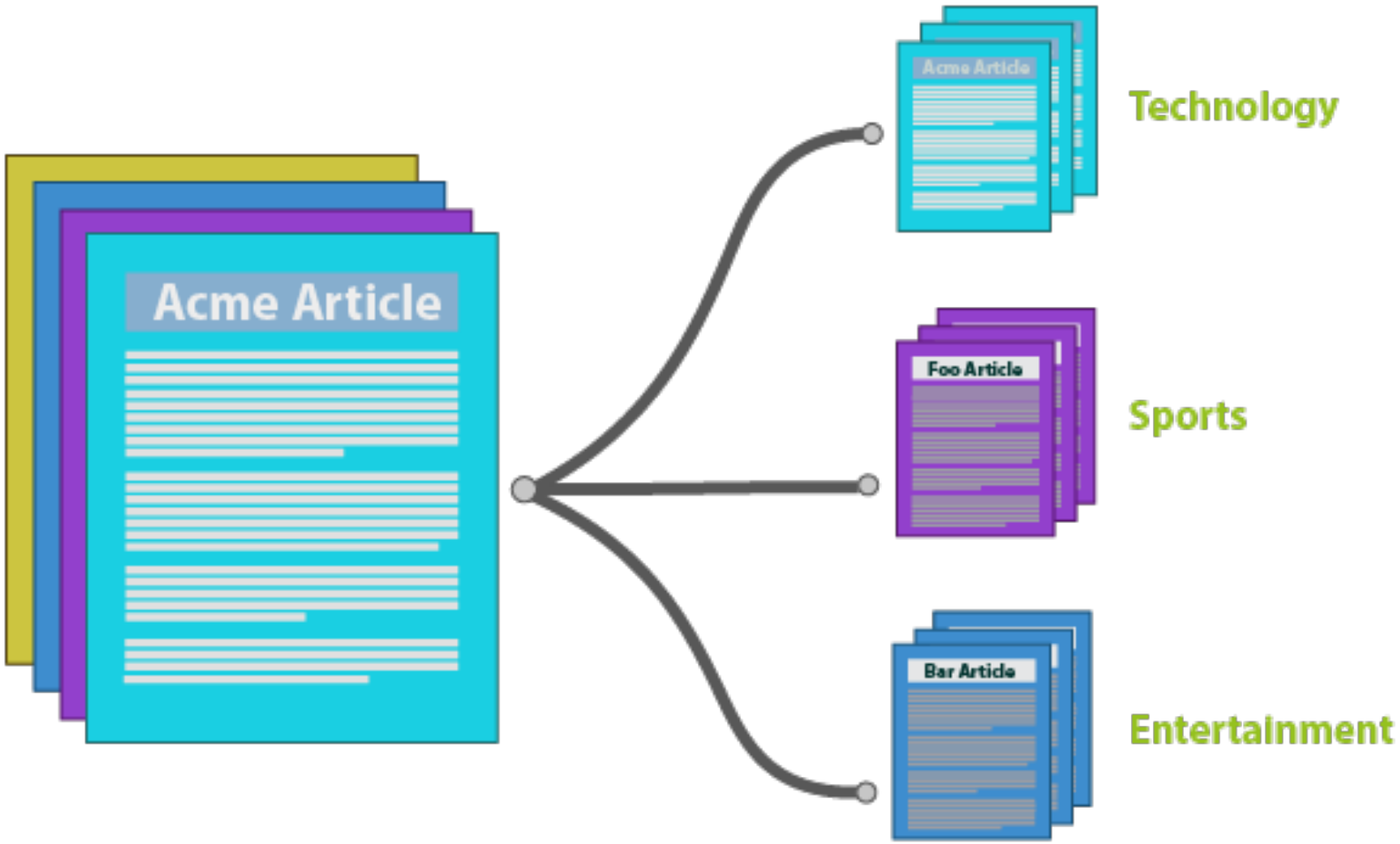
$$q(\theta_i|y,\phi_i) \propto \exp\left(\mathbb{E}_{q(\theta_{-i})}\left[\log \pi(y,\theta)\right]\right)$$

- Automatic Variational Inference: Transform θ to $\mathbb{R}^p$ and assign product of Gaussians. The perform gradient based optimisation (e.g. exact or stochastic gradient descent).

Outline

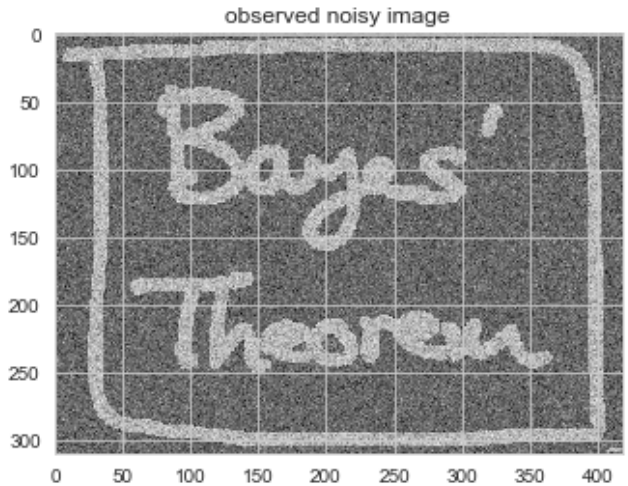
Outline

Example 1: Text Classification

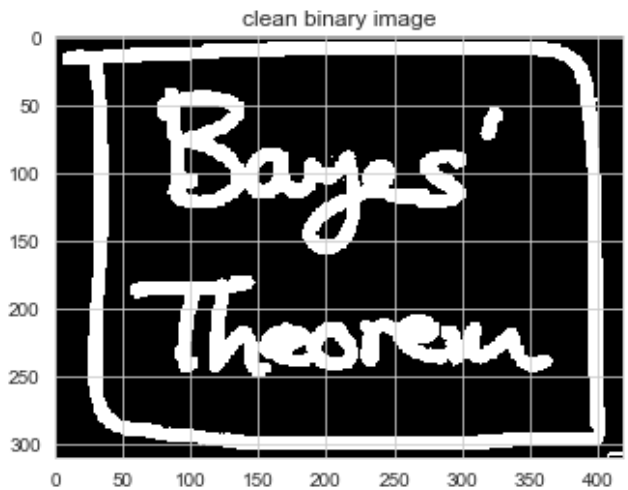


Example 2: Image Processing

Consider the following noisy image.



Can we clean it to get an image like this?



Applications of Graphical Models

Graphical models have been applied to

- Image Processing
- Speech Processing
- Natural Language Processing
- Document Processing
- Pattern Recognition
- Bioinformatics
- Computer Vision
- Economics
- Physics
- Social Sciences
- ...

Definition

Technically, Graphical Models are

- Multivariate probabilistic models
- with some structure
- in terms of conditional independence

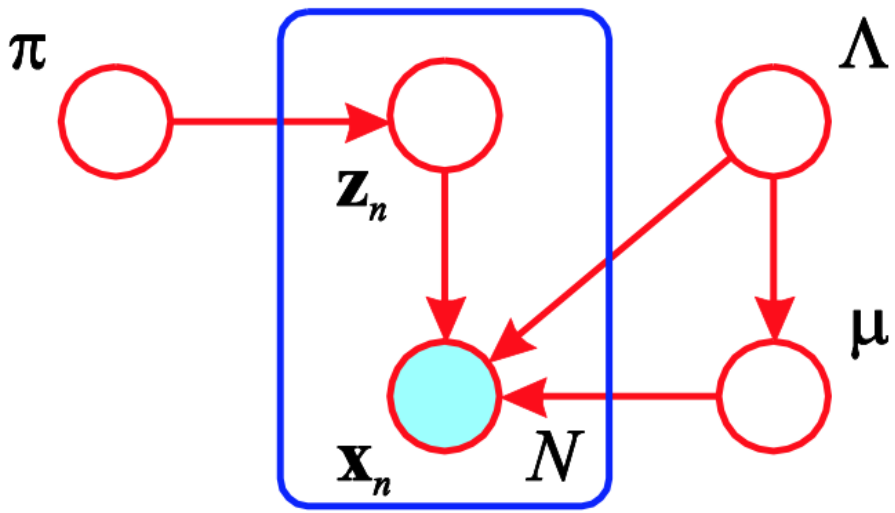
Informally

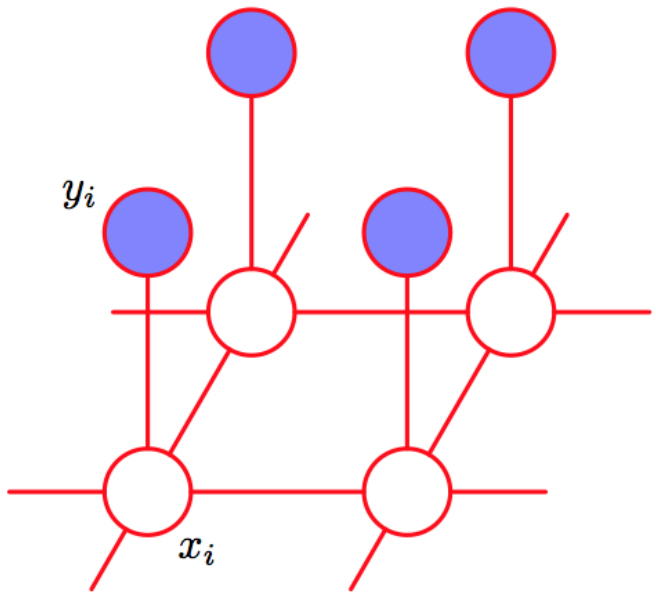
- Models that represent a system
- by its parts - nodes
- and the possible relations among them - edges

Types of Graphical Models

There are two types of Graphical Models

- Bayesian Networks - Directed Graphs
- Markov Random Fields - Undirected Graphs





Why use Graphical models?

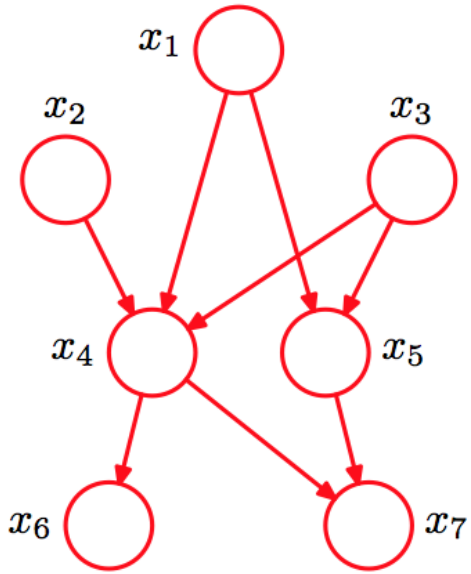
Graphical models are useful

- to understand the joint distribution of a probability model, get insights about conditional independence of different parts, etc.
- to define the joint distribution of a probability model taking into account characteristics of the real world phenomenon.
- to provide graph based algorithms for computation, inference and forecasting.

Outline

Directed graphs and joint probability distributions

Consider the following directed graph.



To which joint probability distribution it corresponds?

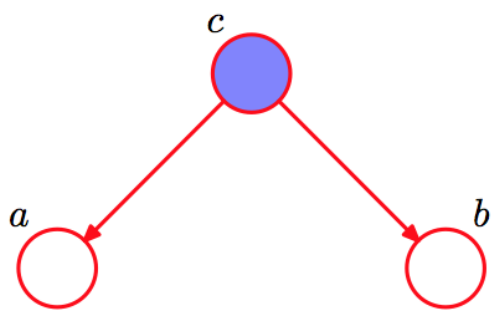
$$\pi(x_1)\pi(x_2)\pi(x_3)\pi(x_4|x_1,x_2,x_3)\pi(x_5|x_1,x_3)\pi(x_6|x_4)\pi(x_7|x_4,x_5)$$

Conditional independence

Nodes a and b are independent conditional on c when

$$\pi(a, b|c) = \pi(a|c)\pi(b|c)$$

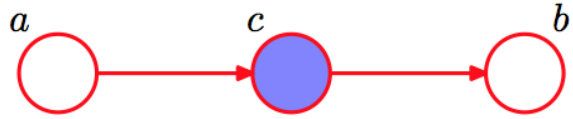
Tail to Tail: Are a and b independent in the graph below?



$$\pi(a, b|c) = \frac{\pi(a, b, c)}{\pi(c)} = \frac{\pi(c)\pi(a|c)\pi(b|c)}{\pi(c)} = \pi(a|c)\pi(b|c), \quad \text{yes.}$$

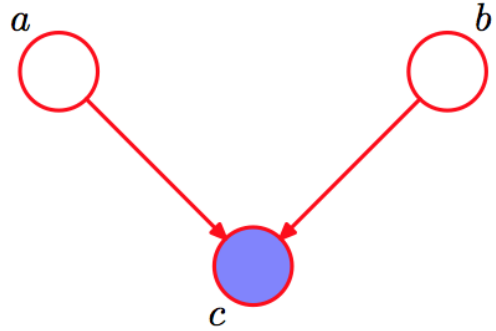
Conditional independence - more examples

Head to Tail: How about this graph?



$$\pi(a,b|c) = \frac{\pi(a,b,c)}{\pi(c)} = \frac{\pi(a)\pi(c|a)\pi(b|c)}{\pi(c)} \overset{\text{Bayes}}{=} \pi(a|c)\pi(b|c), \quad \text{yes.}$$

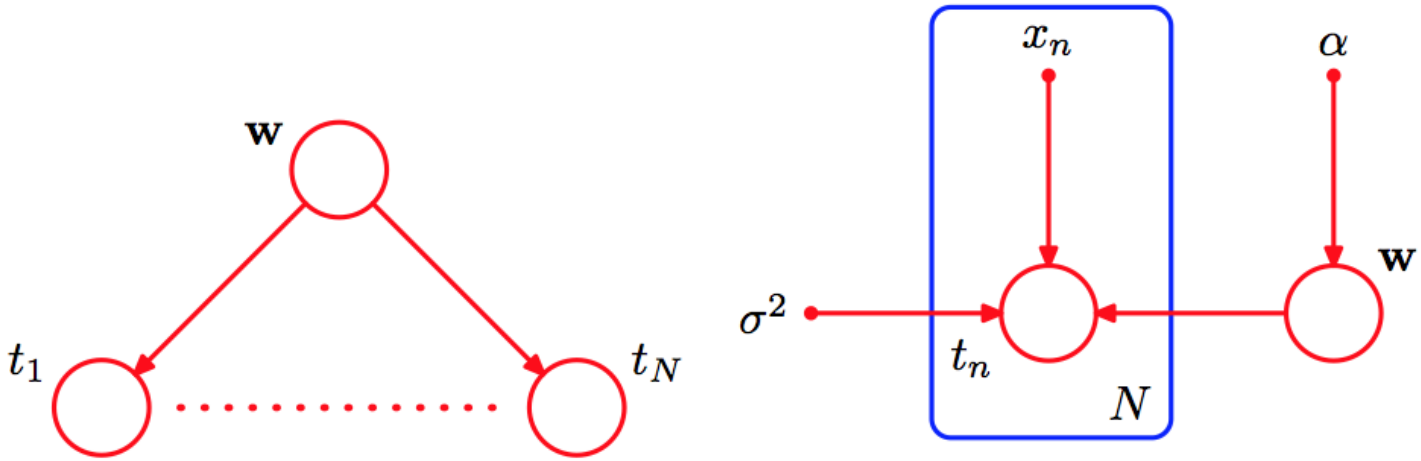
Head to Head: and how about this graph?



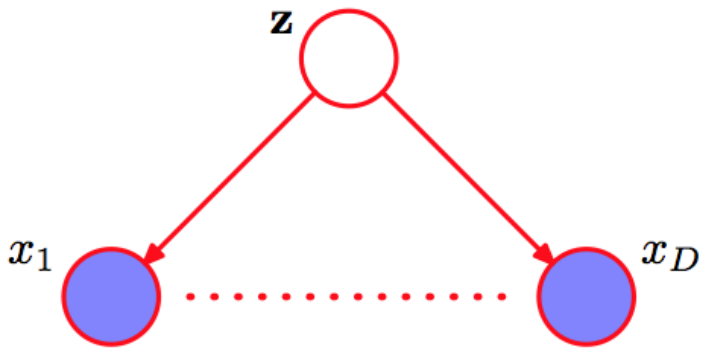
$$\pi(a,b|c) = \frac{\pi(a,b,c)}{\pi(c)} = \frac{\pi(a)\pi(b)\pi(c|a,b)}{\pi(c)}, \quad \text{no.}$$

Graphs for regression and naive bayes classifier

Regression:



Naive Bayes Classifier



Naive Bayes Classifier

Generative model with categorical response y and features $X = (X_1, \ldots, X_p)$.

Specifies $\pi(y, X|\theta) = \pi(y|\theta)\pi(X|y, \theta)$. Then we can write $\pi(y|X, \theta) \propto \pi(y|\theta)\pi(X|y, \theta)$.

The optimal decision rule is to choose $y = 1$ if $\pi(y = 1|X, \theta) > \pi(y = 0|X, \theta)$.

(Naive) Assumption: $\pi(X|y, \theta) = \prod_{j=1}^{p} \pi(X_j|y, \theta)$

Naive Bayes and Laplace smoothing

Given data $D_i = (y_i, X_{1_i}, \ldots, X_{ip})_{i=1}^n$ we can estimate θ using Maximum Likelihood. For binary y and X's we have independent Bernoulli's for each y_i, for each X_j given $y_i = 0$ and for each X_i given $y_i = 1$.

The MLE's can be derived to be the sample proportions of $y_i = 1$, $X_{ji} = 1$ among the points where $y_i = 0$ and among the points where $y_i = 1$ for all $j = 1, \ldots, p$. See exercise 1(a)

Problems occur when one of the X's has no occurrences when $y_i = 0$ or $y_i = 1$. The entire $\pi(X|y, \theta)$ is then 0 regardless of the other X's.

This can be fixed by assigning a prior on θ's e.g. Uniform$(0, 1)$ and using Bayesian Inference, aka Laplace smoothing. See exercise 1(b)

Example: Text classification with Naive Bayes

Consider the following text sentences from articles and their classification tag.

text	**tag**
'A great game'	Sports
'The election was over'	Not sports
'Very clean match'	Sports
'A clean but forgettable game'	Sports
'It was a close election'	Not sports

Now, which tag does the sentence A very close game' belong to?

Note that $\pi(y = 1) = 3/5$ and $\pi(y = 0) = 2/5$. Hence check if

$$\pi(\text{'A very close game'}|y = 1)\frac{3}{5} > \pi(\text{ 'A very close game'}|y = 0)\frac{2}{5}$$

Example: Text classification with Naive Bayes (cont'd)

Using the naive asssumption we can write
$\pi(\text{'a very close game'}|y) = \pi(\text{'a'}|y)\pi(\text{'very'}|y)\pi(\text{'close'}|y)\pi(\text{'game'}|y)$

Focusing on the texts of $y = 1$ we see that the word close does not appear in any sports text so the probability above is 0, even though the word game appears twice.

Laplace smoothing adds e.g. 1 to the numerator and 2 to the denominator to get the probabilities. So $\pi(\text{'close'}|y = 1) = \frac{0+1}{11+2} = \frac{1}{13}$.

A general Beta(α, α) prior would give the estimate

$$\pi(\text{'close'}|y = 1) = \frac{0+\alpha}{11+2\alpha},$$

where α is a hyper-parameter that needs to be specified.

Example: Text classification with Naive Bayes (cont'd)

With $\alpha = 1$ we get

$$\pi(\text{'A very close game'}|y = 1)\frac{3}{5} = ... = 0.000378$$

and

$$\pi(\text{'A very close game'}|y = 0)\frac{2}{5} = ... = 0.000109$$

So this text is classified as Sports.

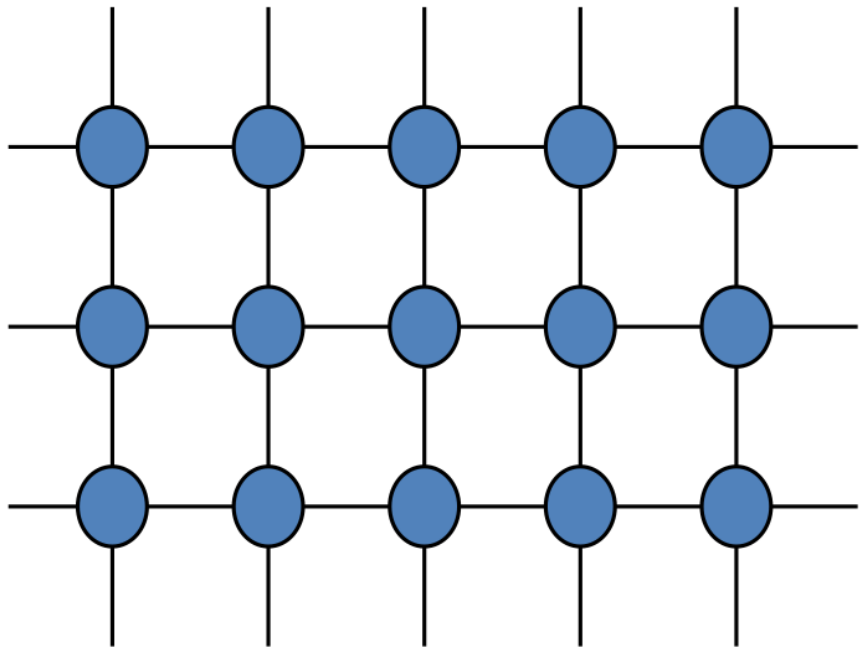
More on text classification

- Removing stop-words: Common words that don't really add anything to the classification, such as a, able, either, else, ever and so on. So for our purposes, 'the election was over' would be 'election over' and 'a very close game' would be 'very close game'.
- Lemmatizing words: This is grouping together different inflections of the same word. So election, elections, elected, and so on would be grouped together as more appearances of the same word.
- Using $n-$grams: Instead of counting single words like we did here, we could count sequences of words, like 'clean match'.
- Using TF-IDF: Instead of just counting frequency we could do something more advanced like also penalizing words that appear frequently in most of the texts.

Outline

Markov Random Fields

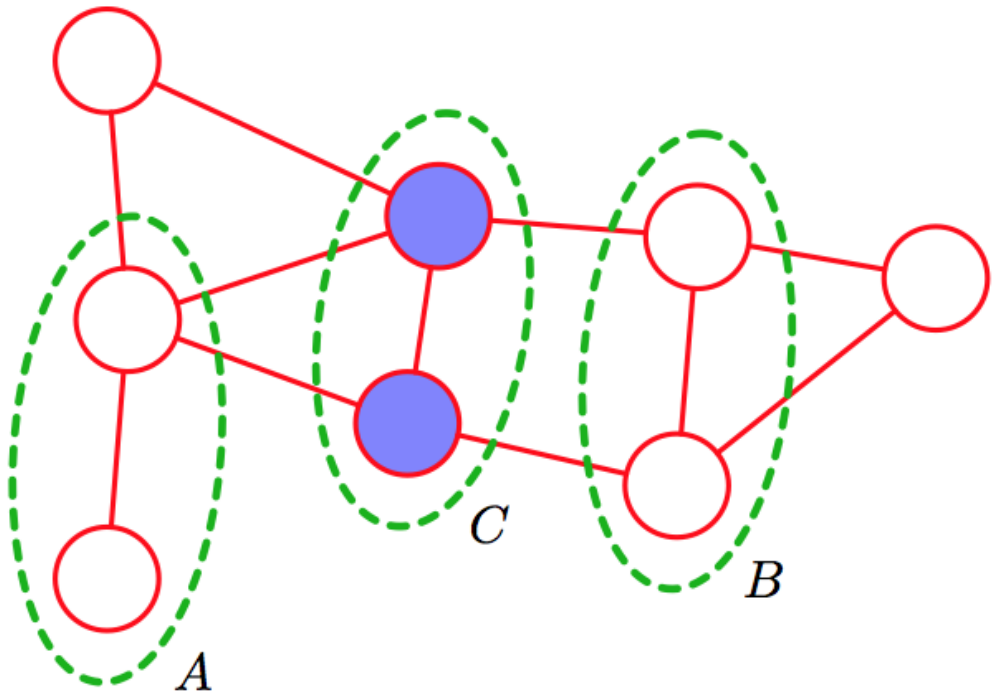
- In many real world phenomena, we can not determine exactly the directionality to the interaction between random variables.
- We can use Markov Networks instead of Bayesian Networks.



Conditional Independence on MRFs

The set of nodes *A* is independent of the set of nodes *B* conditional on the set of nodes *C* if

Every path from any node in set A to any node in set B passes through at least one node in set C.

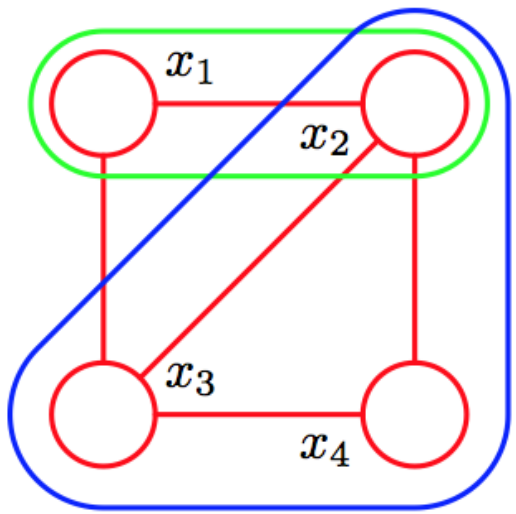


Cliques and Maximal Cliques

A clique is a subset of the nodes in a graph such that there exists a link between all pairs 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.

Below the green subset is a clique but not a maximal clique, whereas the blue subset is a maximal clique



Factorisation

The joint distribution can be written as a product of potential functions $\psi_C(x_C)$ corresponding to the maximal cliques C of the graph.

$$\pi(x) = \frac{1}{Z} \prod_C \psi_C(x_C),$$

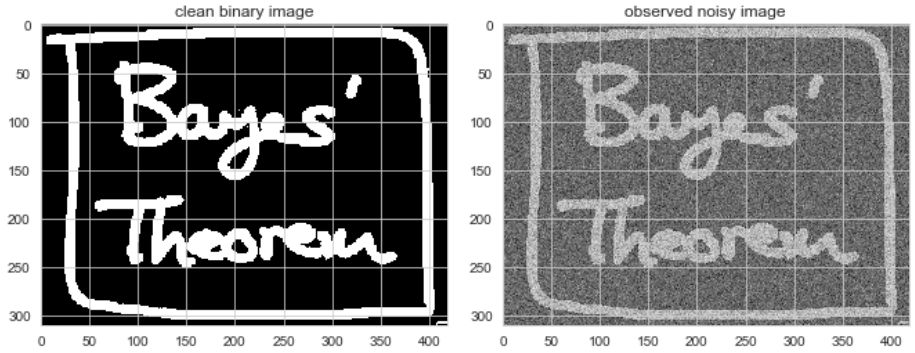
where Z is the normalising constant

So we have

$$\pi(x) = \pi(x_1)\pi(x_2|x_1)\ldots\pi(x_N|x_{N-1})$$

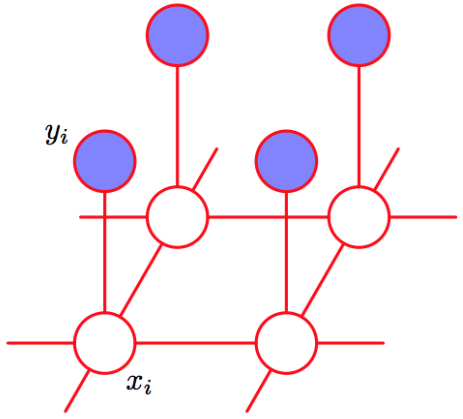
$$\pi(x) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\ldots\psi_{N-1,N}(x_{N-1}, x_N)$$

Example: Image processing



The above images are in bitmap format, consisting of an array of binary pixels y_i, e.g. of black ($y_i = 1$) or white ($y_i = -1$), or continuous y_i's on the grayscale form.

Assume that we observe y_i, a distorted version of the true (binary) x_i. The model can be represented by the MRF below



Ising Model

This is known as the Ising model, originating from Physics.

Under one of its simplified versions, we write

$$\pi(x)\pi(y|x) = \frac{1}{Z_0}\exp\left(\sum_{i=1}^{D}\sum_{j\in nbr}\tilde{W}_{ij}x_i x_j\right)\prod_{i=1}^{D}N(y_i; x_i, \sigma^2)$$

$$\pi(x|y) \propto \exp\left(\sum_{i=1}^{D}\sum_{j\in nbr}W_{ij}x_i x_j - \frac{1}{2\sigma^2}\sum_{i=1}^{D}(y_i - x_i)^2\right)$$

Note that it is more likely for each y_i to be equal with x_i. Also, for positive W's, each x_i is more likely to take the same value with its neighbours (nbr)

The aim is to find the mode of $\pi(x|y)$ (for each x_i) - not an easy task.

Variational Inference for Ising Model

A mean-field variational approximation takes $q(x) = \prod_{i=1}^{D} q(x_i|y, \mu)$

We can write

$$\log q(x_i|y, \mu_i) = \mathbb{E}_{x_{-i}} \left[x_i \sum_{j \in nbr} W_{ij} x_j - \frac{(y_i - x_i)^2}{2\sigma^2} + c \right]$$

$$q(x_i|y, \mu_i) \propto \exp\left(x_i \sum_{j \in nbr} W_{ij} \mu_j - \frac{(y_i - x_i)^2}{2\sigma^2} \right)$$

We would like to update μ_i^{n+1} based on μ_i^n by

$$\mu_i^{n+1} = \mathbb{E}_{q(x_i)} [x_i|y, \mu_i^n] = 1 \times q(x_i = 1|y, \mu_i^n) - 1 \times q(x_i = -1|y, \mu_i^n)$$

Variational Inference for Ising Model (cont'd)

Let $m_i = \sum_{j \in nbr} W_{ij}\mu_j$, $L_i(x_i) = -\frac{(y_i - x_i)^2}{2\sigma^2}$, $L_i^+ = L_i(1)$ and $L_i^- = L_i(-1)$.

We can then write

$$q(x_i = 1 | y, \mu_i^n) = \frac{e^{m_i + L_i^+}}{e^{m_i + L_i^+} + e^{-m_i + L_i^-}} = \frac{1}{1 + e^{-2m_i - (L_i^+ - L_i^-)}} = \sigma(2a_i).$$

where $a_i = m_i + 0.5(L_i^+ - L_i^-)$.

Similarly, $q(x_i = -1 | y, \mu_i^n) = \sigma(-2a_i)$, hence

$$\mu_i^{n+1} = \sigma(2a_i) - \sigma(-2a_i) = \tanh(a_i)$$

Variational Inference for Ising Model (cont'd)

Maybe better to use $\mu_i^{n+1} = \lambda \tanh(a_i) + (1-\lambda)\mu_i^n$ for some λ.

The ELBO in this case can be evaluated as

$$\mathbb{E}_{q(x)}\left[\sum_{i=1}^{D}\sum_{j\in nbr} W_{ij}x_i x_j - \frac{\sum_{i=1}^{D}(y_i - x_i)^2}{2\sigma^2}\right] - \sum_{i=1}^{D}\mathbb{E}_{q_i(x)}\left[\log q_i(x)\right].$$

Variational algorithm: Initiate μ_i's at -say y_i- and then keep updating each one of them as above until ELBO converges.

Outline

Inference and Learning of Graphs

- If all components of the graph are observed (e.g. x and y) the Maximum Likelihood or direct Bayesian methods can be used.

- If some of the components are unobserved (e.g. x), the computational cost increases substantially.
  - Sampling methods (e.g. MCMC) are appropriate but can be computationally infeasible.
  - Exact methods to find the mode of x given y exist but depend on the graph structure. The sum-product/belief propagation is suitable for trees and the junction tree algorithm for more general graphs.
  - Approximate Methods such as variational Bayes can be used.

- The above assumes that the graph structure is known. Learning the graph structure can be seen as model choice but the number of graphs increases exponentially with nodes.

Today's lecture - Reading

Bishop: 8.1 8.2 8.3 and optionally 8.4.

Murphy: 10.1 10.2 10.5 19.1 19.2 19.4.1 21.3.2 and optionally 10.3 10.4 19.3 19.5.