Basic Generative and Graphical Models

Generative and Graphical Models AI60201, Module 1

Adway Mitra

Indian Institute of Technology Kharagpur

8 August 2022

Contents

Background

Introduction

- We have data observations $\{x_1, x_2, \ldots, x_N\}$
- Imagine these to be realizations of **Random Variables** $X_1, X_2, \ldots, X_N$
- Simple assumption: independent and identically distributed (IID)
- $X_i \sim f(\alpha)$
- f: any suitable distribution, α: parameters
- If we know f and α, we can generate new data by sampling
- But how to find f and α?

Choose the distribution

- Distribution should have same support as the observed data
  - Binary data: Bernoulli Distribution, Discrete data: Categorical
  - Count data (integers): Binomial, Poisson, Multinomial, Geometric
  - Real-valued data: Gaussian (all real numbers), Gamma (positive real), Beta ($(0, 1)$)
- Vector data: multivariate Bernoulli/Gaussian etc (dimensions may or may not be independent)
- Distribution's PMF/PDF should match histogram of data

Estimate the parameters

- Likelihood function of parameters: joint distribution of the data (given the parameter values)
- $\mathcal{L}(\alpha) = f(\{X_1, X_2, \ldots, X_N\}|\alpha)$
- Assuming data is IID, $\mathcal{L}(\alpha) = \prod_{i=1}^{N} f(X_i|\alpha)$
- Option 1: maximum likelihood estimate
  - $\alpha_{MLE} = \mathit{argmax}_{\alpha}\mathcal{L}(\alpha)$
  - Not great idea if N is small!
- Option 2: imagine α to be random variables!
  - $g(\alpha)$: *prior distribution* on α
  - $p(\alpha|\{X_1, \ldots, X_N\})$: *posterior distribution* on α

Bayesian Parameter Estimate

- Bayes Theorem: $p(\alpha|\{X_1,\ldots,X_N\}) \propto f(\{X_1,X_2,\ldots,X_N\}|\alpha) * g(\alpha)$
- Aim: Posterior p and prior g should belong to the same family of distributions
- If so, then g is called *conjugate prior* of f
  - $g : Beta$, $f : Bernoulli \rightarrow p : Beta$ (Beta-Bernoulli conjugacy)
  - $g : Gamma$, $f : Poisson \rightarrow p : Gamma$ (Gamma-Poisson conjugacy)
  - $g : Gaussian$, $f : Gaussian \rightarrow p : Gaussian$ (Gaussian mean parameter only)
- Paramater estimate: Maximum a-Posteriori (MAP): mode of posterior!
- If data comes sequentially, posterior of one round becomes prior for next round!

Conjugate Prior

1. Data generation model: $X_i \sim Bernoulli(p)$
   - Prior distribution: $p \sim Beta(a, b)$
   - Posterior distribution on $p \propto \prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i} \times p^{a-1}(1-p)^{b-1}$ i.e. $p^{n_1+a-1}(1-p)^{n_0+b-1}$ where $n_1 = \sum_{i=1}^{N} X_i$ and $n_0 = N - n_1$
   - This is the PDF of $Beta(n_1 + a, n_0 + b)$, i.e.
   - Posterior $p|X \sim Beta(n_1 + a, n_0 + b)$!
2. Data generation model: $X_i \sim Poisson(\lambda)$
   - Prior distribution $\lambda \sim Gamma(k, \theta)$
   - Posterior $\propto \lambda^{\sum_{i=1}^{N} X_i} e^{-\lambda} \times \lambda^{k-1} e^{-\frac{\lambda}{\theta}} = \lambda^{\sum_{i=1}^{N} X_i + k - 1} e^{-\lambda(1+\frac{1}{\theta})}$
   - Posterior $\lambda|X \sim Gamma(\sum_{i=1}^{N} X_i + k, \frac{1}{1+\frac{1}{\theta}})$

Models for Discrete and Continuous Data

Binary Data Generation

- Choose a coin bias, then toss it!
- Coin bias $c \in (0, 1)$: $c \sim Beta(a, b)$
- Coin Toss $X_i \in \{H, T\}$: $X_i \sim Ber(c)$
- Posterior after N tosses: $c|X \sim Beta(n_H + a, n_T + b)$
- Parameter estimation
- Multivariate Bernoulli data: bias $c_1, c_2, \dots, c_D$
- Simplifying assumption: all biases independent
- If not true: model becomes complex (how to encode dependence)?

Categorical Data Generation

- Choose a multi-face dice with weights, then roll it!
- D-faced dice with weights $c_1, c_2, \ldots, c_D$
- $0 \leq c_k \leq 1$, $\sum_{k=1}^{D} c_k = 1$
- $C = \{c_1, \ldots, c_D\} \sim Dirichlet(\alpha_1, \ldots, \alpha_D)$
- Data generation $X_i \sim Categorical(C)$
- Posterior $C|X \sim Dirichlet(\alpha_1 + n_1, \ldots, \alpha_D + n_D)$
- n_d: number of times value d is obtained

Language Model

- Aim: generate a text document by sequentially generating words
- Text document consists of sequence of word tokens: $W_1, W_2, \ldots, W_{d-1}, W_d$
- Word-token W_{d+1} depends on n previous words $W_{d-n+1}, \ldots, W_d$
- Each word-token follows a categorical distribution over the vocabulary
- N-gram model: $W_{d+1} \sim Categorical(\theta)$ where $\theta = \{\theta_1, \ldots, \theta_V\} = f(W_{d-n+1}, \ldots, W_d)$
- High complexity!
- Solution: Bag-of-Words (use frequencies of words instead of sequence of tokens)
- $W_{d+1} \sim Categorical(c_1, \ldots, c_V)$, where $(c_1, \ldots, c_V)$ are counts of each word in $W_{d-n+1}, \ldots, W_d$

Latent Variable Models

Latent Variables

- Supervised learning: $\{X_i, Y_i\}_{i=1}^N$ where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$
- Unsupervised learning: $\{X_i\}_{i=1}^N$
- Generative model: first generate label, then features!
  - $Y_i \sim g(\pi)$ (prior distribution)
  - $X_i \sim f(\theta_k)$ where $k = Y_i$ (class-conditional)
  - Parameters of f specific to Y_i
- Classification/regression: Find $p(Y_i|X_i)$ (posterior)
- Clustering: Y_i is simply the cluster number
- Essentially an inference problem now!

Gaussian Mixture Model

- There are K clusters, with membership proportion π
- In cluster k, the features follow a Gaussian distribution with parameters (μ_k, σ_k)
- $Z_i \sim Categorical(\pi)$ where $Z_i \in \{1, K\}$
- $X_i \sim \mathcal{N}(\mu_k, \sigma_k)$ where $k = Z_i$
- Observed data: $\{X_i\}_{i=1}^N$, K known (or assumed)
- Parameter estimation problem: estimate $\theta = \{\pi, \mu, \sigma\}_{k=1}^K$
- Inference problem: $prob(Z|X, \theta)$

Parameter Estimation of GMM

- Likelihood function
  $\mathcal{L}(\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{\mathbb{I}(Z_i=k)} exp(-\frac{1}{2\sigma_k^2}(X_i - \mu_k)^2 \mathbb{I}(Z_i = k))$
- or log-likelihood $\ell(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}(Z_i = k)(log(\pi_k) - \frac{1}{2\sigma_k^2}(X_i - \mu_k)^2)$
- Cannot evaluate likelihood as it contains unknown Z
- Solution: evaluate *expected likelihood* by replacing $\mathbb{I}(Z_i = k)$ with $E(\mathbb{I}(Z_i = k))$ where the expectation is over $p(Z_i|X_i, \theta)$
- Make initial estimate of θ, use it to calculate the expected likelihood (E-step)
- Update the parameter values to maximize the expected likelihood (M-step)
- Keep repeating the above till convergence (E-M algorithm)
- Result: final estimates of θ^{EM}, posterior $p(Z_i|X_i, \theta^{EM})$ (soft clustering)

Continuous Latent Variables (Probabilistic PCA)

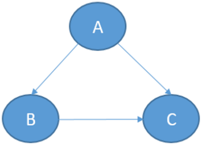
- Data $X_i \in \mathcal{R}^{\mathcal{D}}$
- Data Generation Model: $X_i \sim \mathcal{N}(WZ_i + \mu, \sigma^2 I)$, $Z_i \sim \mathcal{N}(0, I)$ (where I is the $D \times D$ identity matrix)
- Latent variable $Z_i \in \mathcal{R}^d$ where $d \leq D$ (low-dimensional representation)
- We can get the posterior distribution as $p(Z_i|X_i) = p(X_i|Z_i)p(Z_i)$
- Turns out: $Z_i|X_i \sim \mathcal{N}(M^{-1}W^T(X_i - \mu), \sigma^2 M^{-1})$ where $M = W^T W + \sigma^2 I$
- It can also be shown that $X_i \sim \mathcal{N}(\mu, WW^T + \sigma^2 I)$
- Parameter estimate $W_{ML} = \mathit{argmax}_W p(X) = U_d(\Lambda_d - \sigma^2 I)R$ where U_d is the first d eigenvectors of sample covariance matrix C, Λ_d contains the corresponding eigenvalues and R is $d \times d$ orthogonal rotation matrix
- Low-dimensional estimate $Z_i = M^{-1}W_{ML}^T(X_i - \mu)$

Directed Graphical Models

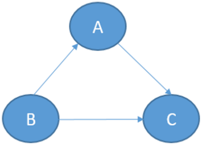
Definition

- Represent each random variable as a node/vertex
- Vertices V are connected by directed edges, each vertex $v \in V$ has *parents* denoted by $pa(v)$
- At each vertex v, a conditional probability distribution $p(v|pa(v))$ is specified
- Whole graph represents joint distribution over the concerned random variables, like $p(V) = \prod_{i=1}^{|V|} p(V_i|pa(V_i))$
- Structure of tree represents a particular *factorization* of the distribution
- $p(A, B, C) = p(A)p(B|A)p(C|A, B)$
- If $A \perp B$, $p(A, B, C) = p(A)p(B)p(C|A, B)$
- If $A \perp C$, $p(A, B, C) = p(A)p(B|A)p(C|B)$
- These factorizations can be represented by different graph structures!

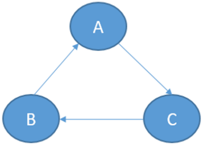
Examples



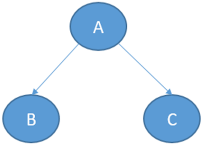
P(A,B,C) = p(A)*p(B|A)*p(C|A,B)



P(A,B,C) = p(B)*p(A|B)*p(C|A,B)



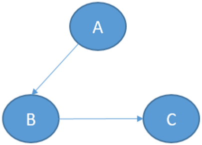
CYCLE ALERT!!



P(A,B,C) = p(A)*p(B|A)*p(C|A)

Implication: B⊥C|A (different from B⊥C)

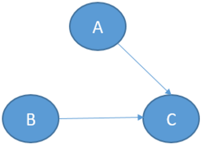
A is the common cause of B and C



P(A,B,C) = p(A)*p(B|A)*p(C|B)

Implication: A⊥C|B

Chain structure



P(A,B,C) = p(A)*p(B)*p(C|A,B)

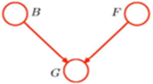
Implication: B⊥A (but not B⊥A|C)!

Collider/V-structure

Examples

Fuel System: Given Probability Values

- We are given prior probabilities and one set of conditional probabilities



Battery Prior Probabilities $p(B)$

B	$p(B)$
1	0.9
0	0.1

Fuel Prior Probabilities $p(F)$

F	$p(F)$
1	0.9
0	0.1

Conditional probabilities of Guage $p(G|B,F)$

B	F	$p(G=1)$
1	1	0.8
1	0	0.2
0	1	0.2
0	0	0.1

Fuel System Joint Probability

- Given prior probabilities and one set of conditional probabilities

B	$p(B)$
1	0.9
0	0.1

F	$p(F)$
1	0.9
0	0.1

Conditional probs

B	F	$p(G=1)$
1	1	0.8
1	0	0.2
0	1	0.2
0	0	0.1

Probabilistic structure is completely specified since

$$p(G,B,F)=p(B,F|G)p(G)$$
$$=p(G|B,F)p(B,F)$$
$$=p(G|B,F)p(B)p(F)$$

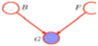
e.g., $p(G)=\Sigma_{B,F}p(G|B,F)p(B)p(F)$

$p(G|F)=\Sigma_B p(G,B|F)$ by sum rule

$= \Sigma_B p(G|B,F)p(B)$ prod rule

$$p(F|G,B)=\frac{p(G,B|F)pF)}{p(G,B)}=\frac{p(G\mid B,F)p(B\mid F)p(F)}{p(G\mid B)p(B)}=\frac{p(G\mid B,F)p(F)}{\sum_F p(G\mid B,F)p(F)}$$

Fuel System Example



- Suppose guage reads empty $(G=0)$
- We can use Bayes theorem to evaluate fuel tank being empty $(F=0)$

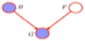
$$\boxed{p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}}$$

- Where $p(G=0)=\sum_{B=0,1}\sum_{F=0,1}p(G=0\mid B,F)p(B)p(F)=0.315$

$$p(G=0\mid F=0)=\sum_{B=0,1}p(G=0\mid B,F=0)p(B)=0.81$$

- Therefore $p(F=0|G=0)=0.257>p(F=0)=0.1$ 21

Clamping additional node



- Observing both fuel guage and battery
- Suppose Guage reads empty $(G=0)$ *and* Battery is dead $(B=0)$
- Probability that Fuel tank is empty

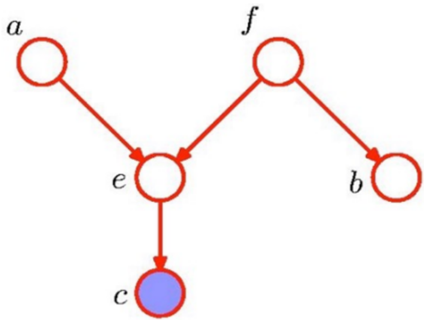
$$\boxed{p(F=0|G=0,B=0)=\frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F=0,1}p(G=0\mid B=0,F)p(F)}=0.111}$$

- Probability has decreased from 0.257 to 0.111

Conditional Independence

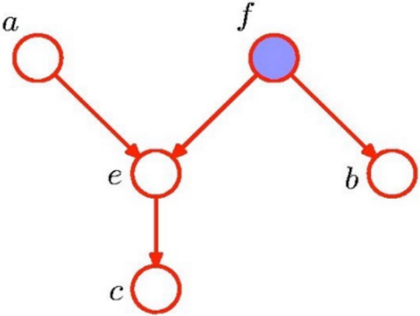
- A probability distribution $f(X_1, \ldots, X_N)$ may entail various conditional independence relations
- Suppose Bayesian Network G represents a valid factorization of f. Can G represent its independence relations too?
- Bayesian Network employs d-separation between pairs of vertices (say A and B), conditioned on a set of vertices (say C)
- A and B are d-separated iff every (undirected) path between them is *blocked*,i.e.
  - If for any $v \in C$ on the path, the configuration is either head-to-tail or tail-to-tail, **OR**
  - If for every v on the path with head-to-head configuration, neither v nor its descendants are in C
- d-separations entailed by a Bayesian Network may be used to compare it to the distribution!

Examples



$\neg\mathrm{dsep}(a, b|c)$

We condition on a descendant of e, i.e. it does not block the path from a to b.



$\mathrm{dsep}(a, b|f)$

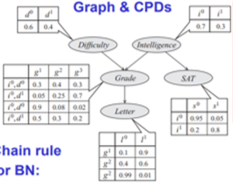
We condition on a tail-to-tail node on the only path from a to b, i.e f blocks the path.

Conditional Independence

- If $X_i \perp X_j | X_C$ with respect to f implies V_i and V_j are d-separated in G with respect to V_C, then the Bayesian Network G is an D-MAP for f
- If V_i and V_j are d-separated in G with respect to V_C implies $X_i \perp X_j | X_C$ with respect to f, then the Bayesian Network G is an I-MAP for f
- If G is both D-MAP and I-MAP of f, then it is called a *Perfect Map*
- To remember: a factorization of f can be easily represented by a Bayesian Network and vice-versa. But conditional independence relations may not match!
- Theorem: If A and B are any two random variables such that B is not an ancestor of A w.r.t. the perfect map G, then $A \perp B | C$ where C is the set of ancestors of A w.r.t. G

Examples

Graph & CPDs



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$Val(I)=\{i^0=$low intelligence, $i^1=$ high intelligence$\}$
$Val(D)=\{d^0=$easy, $d^1=$hard$\}$
$Val(G)=\{g^1=A,\ g^2=B,\ g^3=C\}$
$Val(S)=\{s^0=$low, $s^1=$high$\}$
$Val(L)=\{l^0=$weak, $l^1=$strong$\}$

Chain rule for BN:

$$P(D,I,G,S,L)=P(D)P(I)P(G|D,I)P(S|I)P(L|G)$$
$$P(i^1,d^0,g^2,s^1,l^0)=P(i^1)P(d^0)P(g^2|i^1,d^0)P(s^1|i^1)P(l^0|g^2)$$
$$=0.3\cdot 0.6\cdot 0.08\cdot 0.8\cdot 0.4=0.004608$$

→ P(high intelligence, easy course, grade=B, high SAT, weak letter)= very low

Machine Learning Srihari

Causal Reasoning

1. How likely *George* will get a strong *Letter* (No evidence)?
   $P(l^1)=0.502$ $\quad P(l^1)=\sum_{D,I,G,S} P(D,I,G,S,L=l^1)=\sum_{D,I,G,S} P(D)P(I)P(G|D,I)P(S|I)P(l^1|G)$

- Obtained by summing-out other variables in joint distribution

2. Knowing *George* is not so *Intelligent* (i^0)
   - $P(l^1|i^0)=0.389$ $\quad P(l^1|i^0)=\frac{P(l^1,i^0)}{P(i^0)}=\frac{\sum P(D)P(i^0)P(G|D,i^0)P(S|i^0)P(l^1|G)}{\sum P(D)P(i^0)P(G|D,i^0)P(S|i^0)P(L|G)}$

3. Knowing ECON101 is not *Difficult* (d^0)
   - $P(l^1|i^0,\ d^0)=0.513$

Machine Learning Srihari

Evidential Reasoning

- Recruiter wants to hire *Intelligent student*
- A priori *George* is 30% likely to be *Intelligent*
- $P(i^1)=0.3$
- Finds that *George* received *Grade* C (g^3) in *ECON*101
- $P(i^1|g^3)=0.079$
- Similarly probability of *Difficult* goes up from 0.4 to
- $P(d^1|g^3)=0.629$
- If recruiter has lost *Grade* but has *Letter*
- $P(i^1|l^0)=0.14$

- Recruiter has both *Grade* and *Letter*
- $P(i^1|l^0,g^3)=0.079$
  - Same as if he had only *Grade*
  - *Letter* is immaterial
- Reasoning from effects to causes is called evidential reasoning

8

Courtesy:Srihari Sargur, SUNY Buffalo

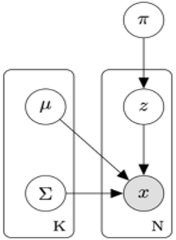
Latent Dirichlet Allocation

- A generative model for text documents based on topics
- Topic: distribution over the vocabulary V of words
- Prior on topic k: $\phi_k \sim Dirichlet(\beta 1^{|V|})$
- For document d: a prior distribution over the K topics: $\theta_k \sim Dirichlet(\alpha 1^K)$
- For word-token i in document d: choose topic $Z_{di} \sim Categorical(\theta_d)$
- Now, generate the word for that token $X_{di} \sim Categorical(\phi_k)$ where $k = Z_{di}$
- Observed variables $\{X_{di}\}$, latent variables $\{\theta\}, \{\phi\}, \{Z_{di}\}$, hyperparameters α, β
- Joint distribution $L = \prod_k \phi_k \prod_d \theta_d \prod_i p(Z_{di}|\theta_d)p(X_{di}|Z_{di}, \phi)$
- Target: topics ϕ, topic assignments $p(Z_{di}|X)$

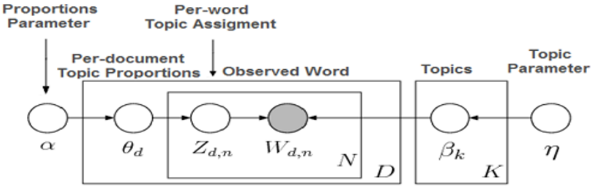
Hidden Markov Model

- First state distribution $pi \sim Dirichlet(\alpha 1^K)$
- State transition distribution from state k: $A_k \sim Dirichlet(\beta 1^K)$
- Emission distribution from state k: $B_k \sim Dirichlet(\gamma 1^V)$ where V is the output space size
- Initial state of sequence s: $Z_{s1} \sim \pi$
- State at time t: $Z_{st} \sim Categorical(A_k)$ where $k = Z_{s,t-1}$
- Output at time t: $X_{st} \sim Categorical(B_l)$ where $l = Z_{st}$
- Joint distribution
  $L = p(\pi) \prod_k p(A_k)p(B_k) \prod_s p(Z_{s1}) \prod_{t=2}^T p(Z_{st}|Z_{s,t-1}) \prod_{t=1}^T p(X_{st}|Z_{st})$

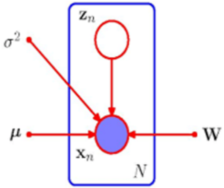
Plate Notation



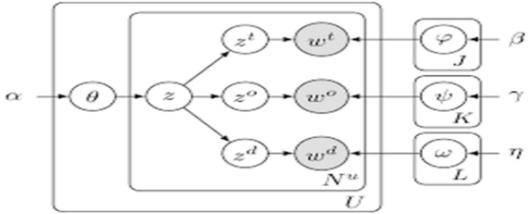
GAUSSIAN MIXTURE MODEL



LATENT DIRICHLET ALLOCATION



PROBABILISTIC PCA



Undirected Graphical Models

Ising Model

- In some situations, joint distributions are more expressive than conditional distributions
- In a ferromagnetic substance, each dipole's spin should be aligned with neighbors
- State where all dipoles have aligned spin is most desirable
- State where maximum neighbors are in contrasting orientation is the least desirable
- Each pixel in a binary image: should follow its neighbors
- Represent each dipole or pixel as random variable
- Define a distribution over state configurations: joint distribution over these random variables
- Every adjacent pair of pixels should contribute as a factor!

Gibbs Random Field

- Undirected Graphical Model to represent the factorization of a joint distribution
- In a graph, $\mathcal{C}$ denotes the set of maximal cliques (complete subgraph)
- In Bayesian Network (directed graphical model), the factors are conditional distributions of individual variables conditioned on their parents
- In Gibbs Random Field (undirected graphical model), the factors are defined as potential functions over these cliques
- $p(X) = \prod_c \psi_c(X_c)$ where $c \in \mathcal{C}$ is a clique in the MRF, the associated random variables are X_c and ψ_C is the corresponding *potential function*
- For many graphs, the maximal cliques are just the edges, in which case we have *edge potential functions*

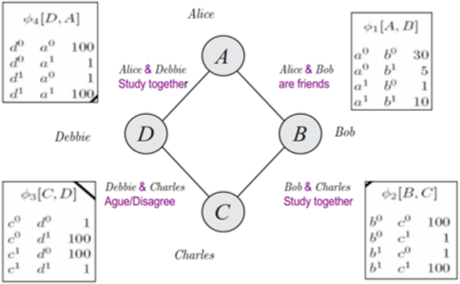
GRF for Ising Model

- $X_{i,j}$ and $X_{i+a,j+b}$ are neighboring if $a = \{-1, 0, 1\}$ and $b = \{-1, 0, 1\}$
- Define $\psi_e(X_{i,j}, X_{i+a,j+b}) = exp(\mathbb{I}(X_{i,j} = X_{i+a,j+b}))$ where e is the edge between (i, j) and $(i + a, j + b)$
- Idea: edge potential function takes high value if end-nodes have equal value
- Joint distribution $p(X) = \frac{1}{Z} \prod_e \psi_e(X_{i,j}, X_{i+a,j+b})$ where Z is called the partition function for normalization
- Mode of the distribution: all X equal!
- Coherent structures of X have high probability under distribution p

Markov Random Field

- A Markov Random Field is a Gibbs Random Field with special conditional independence properties
- In an MRF, the RV represented by node X is independent of the other RVs conditioned on variables represented by the neighbors of X
- $A \perp B|C$ if $V_C = \mathcal{N}(V_A)$ and $V_B \notin \mathcal{N}(V_A)$
- If all paths connecting V_A and V_B pass through V_C, then $A \perp B|C$ (d-separation in Undirected Graphical Models)
- *Hammersley-Clifford Theorem*: Any Gibbs Random Field based on a *positive* distribution is also a Markov Random Field
- Some Bayesian Networks (but not all) can be represented as a Markov Random Field and vice versa
- Note: Ising Model follows MRF due to the edge potential function!

Examples



Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

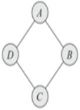
ameters

- We can obtain any desired probability from the joint distribution as usual

$P(b^0)=0.268$: *Bob* is 26% likely to have a misconception

$P(b^1|c^0)=0.06$: if *Charles* does not have the misconception, *Bob* is only 6% likely to have misconception.

- Most probable joint probability (from table): $P(a^0,b^1,c^1,d^0)=0.69$
  - *Alice,Debby* have no misconception, *Bob,Charles* have misconception



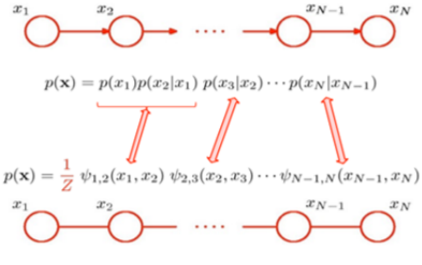
Z=7,201,840

Courtesy:Srihari Sargur, SUNY Buffalo

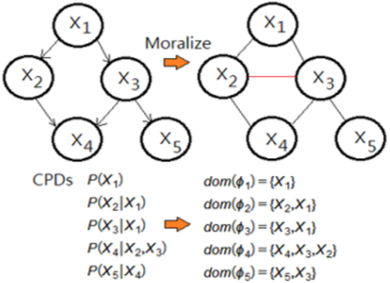
Directed to Undirected Model

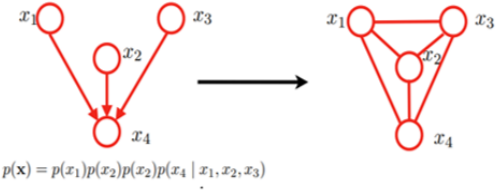
- Not all Bayesian Networks have an equivalent Markov Network (i.e. any given probability distribution may not have a BN as perfect map and also a MRF as perfect map) - except *chordal graphs*
- No MRF is equivalent to the *V-structure* of a BN
- No BN is equivalent to the square MRF
- But it is possible to create an MRF that will be an I-MAP of the BN, i.e. every conditional independence entailed by the MRF's perfect distribution will also hold for the BN's perfect distribution
- Steps:
  1. **Moralization**: remove all V-structures by adding undirected edges among the parents (results in loss of some conditional independences!)
  2. remove directionality from all edges
- Output: **moral graph** of the given BN
- Resultant MRF will be perfect map for BN iff there are no v-structure in original BN

Examples



In this case: Z=1





Problem: This process can remove conditional independence relations (inefficient)

Generally: There is no one-to-one mapping between the distributions represented by directed and by undirected graphs.

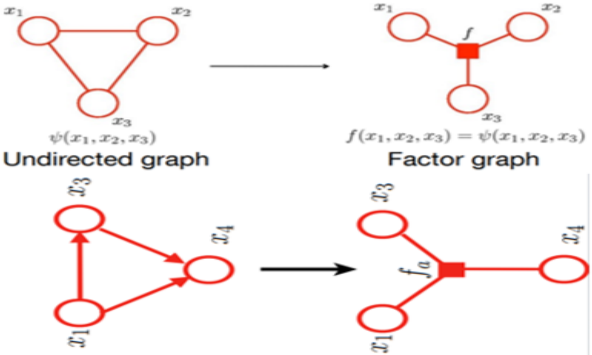
I-Map of a given distribution

- *Markov Blanket* of a variable X: minimal set of variables $MB(X)$ such that $X \perp Y | MB(X)$ where $Y \notin MB(X)$
- Approach to create MRF G that is an I-MAP of distribution P:
    - Create a node for each variable
    - for every variable X, identify its Markov Blanket from P
    - connect the node representing X to all nodes representing $MB(X)$
- Approach to create BN G that is an I-MAP of P:
    - We need an *ordering* over the variables
    - Factorize the joint distribution according to this ordering, using the Markov Blanket of each variable
    - Add directed edges according to this factorization

Hybrid Graphical Models

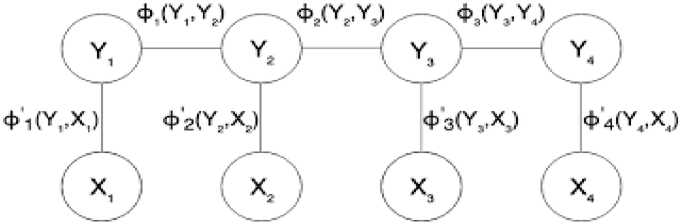
Factor Graph

- A generalization of directed and undirected graphical models
- A bipartite undirected graph: one set of vertices for each variable, one set for each *factor*
- Each factor vertex connected to corresponding variables by edges
- Useful representation for inference (later)



Conditional Random Field

- Suitable for situations where each observation i has *latent features* X_i
- Further, the labels Y of different datapoints are inter-related
- Conditional Random Fields to model $p(Y|X)$ where $Y = \{Y_1, \ldots, Y_T\}$ and $X = \{X_1, \ldots, X_T\}$ as product of factors
- $p(Y|X) = \prod_{t=1}^{T-1} \phi_Y(Y_t, Y_{t+1}) \times \prod_{t=1}^{T} \phi_{XY}(X_t, Y_t)$ where ϕ_Y maintains relation between adjacent observations and ϕ_{XY} maintains relation between observation and latent feature
- In some cases, $Y(t) - X(t)$ edges represented as directed



Sample Questions

- Given a set of observations, how well does a given generative model fit it?
- Write the likelihood function, i.e. joint distribution of observed and latent variables using the parameters. Since latent variables are not known, marginalize them by summing or integrating over all possible values
- Given a set of observations and parameter values for a model, what are the likely values of latent variables of that model?
- Write the likelihood function, find the values of latent variables that maximize this function
- Comparison of two models to fit a given dataset
- Calculate the likelihood functions for both models (after marginalizing over latent variables if needed), and choose the model with higher likelihood

Sample Questions

- Is a given graphical model an I-MAP/D-MAP/Perfect-MAP of a given distribution?
- Find all the D-separations entailed by the model (for various conditioning sets). Check if the corresponding independence relations hold with respect to the given distribution, by marginalizing over the remaining variables in each case
- Given a probability distribution, create its I-MAP
- Identify conditional independence relations entailed by distribution. Identify Markov Blanket of each variable. Order the variables if Bayesian Network needed. Set one node for each variable. Add edges to each node according to Markov Blanket property.
- Given a graphical model, find the most likely value of a variable, given the observations of some other variables and values of all the parameters
- Write down the full joint distribution using the parameters. Marginalize the variables other than the target and the observed ones. Maximize this marginalized joint distribution with respect to the target variable

Thank you!