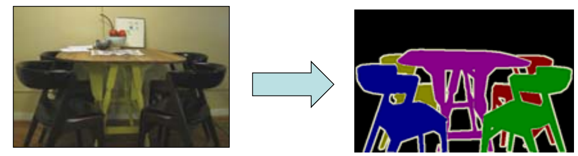
Undirected Graphical Models 1

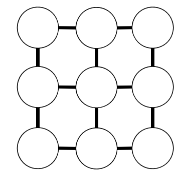
1

Symmetric interactions (Examples)

Image Segmentation (From PASCAL VOC 2011 data)



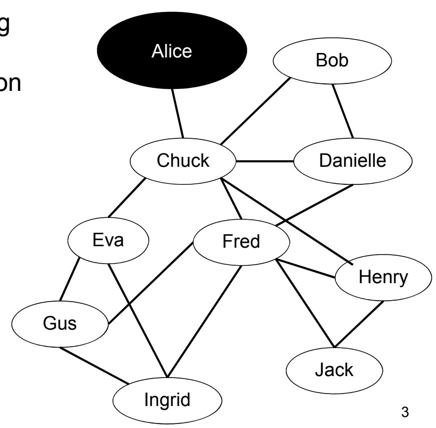
Each node in this undirected graphical model is a pixel / region



2

Symmetric interactions (Examples)

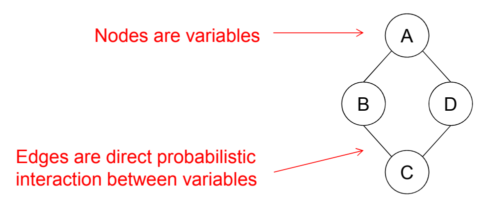
Social network modeling
- Marketing
- Insider threat detection
- Fraud detection



3

Introduction

Markov network



What about the parameters?
- Standard CPD doesn't work – no notion of a "parent"
- Need a more symmetric parameterization

4

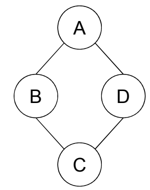
Introduction

Let $\boldsymbol{D}$ be a set of random variables. We define a factor ϕ to be a function from $Val(\boldsymbol{D}) \rightarrow \Re$. A factor is nonnegative if all its entries are nonnegative.

The set of variables $\boldsymbol{D}$ is called the scope of the factor and denoted $Scope[\phi]$

Unless stated otherwise, we restrict attention to nonnegative factors

Introduction



A	B	$\phi_1(A,B)$
0	0	30
0	1	5
1	0	1
1	1	10

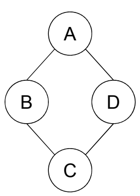
B	C	$\phi_2(B,C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\phi_3(C,D)$
0	0	1
0	1	100
1	0	100
1	1	1

D	A	$\phi_4(D,A)$
0	0	100
0	1	1
1	0	1
1	1	100

Introduction

Think of $\boldsymbol{\phi_1(A,B)}$ like an unnormalized joint distribution between A and B. This column doesn't have to sum to 1



A	B	$\phi_1(A,B)$
0	0	30
0	1	5
1	0	1
1	1	10

The bigger the value, the more likely the configuration eg. A = 0, B = 0 is the most likely

I can increase this value to make A=1 and B=1 more likely but it is not clear how this affects the full joint distribution between A, B, C, and D

Introduction

Because the factors are not normalized, need to normalize everything at the end to produce a probability distribution.

$$P(a,b,c,d) = \frac{1}{Z}\emptyset_1(a,b)\ \emptyset_2(b,c)\emptyset_3(c,d)\emptyset_4(d,a)$$

$$Z = \sum_{a,b,c,d} \emptyset_1(a,b)\ \emptyset_2(b,c)\emptyset_3(c,d)\emptyset_4(d,a)$$

Normalizing constant (also called the partition function). Can be difficult to compute!

Introduction

Connections between factorization and independence properties

- Structure of the factors allows us to decompose the distribution
- $P \models (\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z})$ iff $P(\mathcal{X}) = \phi_1(\boldsymbol{X}, \boldsymbol{Z})\phi_2(\boldsymbol{Y}, \boldsymbol{Z})$
  Independence properties of the distribution *P* correspond to separation properties of the graph *G* over which *P* factorizes

9

Parameterizations

10

Parameterization

- Factors subsume (generalize) the notion of a joint distribution:
  - A joint distribution over $\boldsymbol{D}$ is a factor over $\boldsymbol{D}$
- Factors subsume a conditional probability distribution (CPD)
  - A CPD $P(X|\boldsymbol{U})$ is a factor over $\{X\} \cup \boldsymbol{U}$.
  - A CPD is a special case of a factor that is normalized

11

Parameterization

Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be three disjoint sets of variables, and let $\phi_1(\boldsymbol{X}, \boldsymbol{Y})$ and $\phi_2(\boldsymbol{Y}, \boldsymbol{Z})$ be two factors. We define the factor product $\phi_1 \times \phi_2$ to be a factor Ψ: $Val(X, Y, Z) \rightarrow \Re$ as follows:

$$\Psi(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \phi_1(\boldsymbol{X}, \boldsymbol{Y})\phi_2(\boldsymbol{Y}, \boldsymbol{Z})$$

12

Parameterization

Example of a factor product:

A	B	$\phi_1(A,B)$
0	0	0.5
0	1	0.8
1	0	0.1
1	1	0
2	0	0.3
2	1	0.9

B	C	$\phi_2(B,C)$
0	0	0.5
0	1	0.7
1	0	0.1
1	1	0.2

A	B	C	$\Psi(X,Y,Z)$
0	0	0	(0.5)(0.5)=0.25
0	0	1	(0.5)(0.7)=0.35
0	1	0	(0.8)(0.1)=0.08
0	1	1	(0.8)(0.2)=0.16
1	0	0	(0.1)(0.5)=0.05
1	0	1	(0.1)(0.7)=0.07
1	1	0	(0)(0.1)=0
1	1	1	(0)(0.2)=0
2	0	0	(0.3)(0.5)=0.15
2	0	1	(0.3)(0.7)=0.21
2	1	0	(0.9)(0.1)=0.09
2	1	1	(0.9)(0.2)=0.18

13

Parameterizations

For Bayesian Networks;

- Since CPDs and joint distributions are factors
- Chain rule for BNs can be thought of as the product of CPD factors
- Letting $\phi_{X_i}(X_i, Parents(X_i))$ =$P(X_i|Parents(X_i))$

$$P(X_1, \ldots, X_N) = \prod_i \phi_{X_i}(X_i, Parents(X_i))$$

14

Parameterizations

A distribution P_Φ is a Gibbs distribution parameterized by a set of factors $\Phi = \{\phi_1(\boldsymbol{D_1}), \ldots, \phi_K(\boldsymbol{D_K})\}$ if it is defined as follows:

$$P_\Phi(X_1, \ldots, X_n) = \frac{1}{Z}\tilde{P}_\Phi(X_1, \ldots, X_n)$$

where

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \phi_1(\boldsymbol{D_1}) \times \phi_2(\boldsymbol{D_2}) \times \ldots \times \phi_K(\boldsymbol{D_K})$$

is an unnormalized measure and

$$Z = \sum_{X_1, \ldots, X_n} \tilde{P}_\Phi(X_1, \ldots, X_n)$$

is a normalizing constant called the partition function

15

Parameterizations

We say that a distribution P_Φ with $\Phi = \{\phi_1(\boldsymbol{D_1}), \ldots, \phi_K(\boldsymbol{D_K})\}$ factorizes over a Markov network $\mathcal{H}$ if each $\boldsymbol{D_k}$ (k=1, …, K) is a complete subgraph (or clique) of $\mathcal{H}$

A complete subgraph (or clique) is a fully connected subgraph

16

Parameterizations

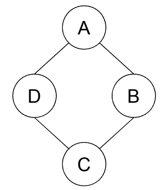
The terms that you multiply together for the joint distribution of a Markov network are often called clique potentials

$$P(X_1, \ldots, X_N) = \frac{1}{Z} \underbrace{\phi_1(\boldsymbol{C_1})}_{\text{Clique Potential}} \times \phi_2(\boldsymbol{C_2}) \times \ldots \times \phi_K(\boldsymbol{C_K})$$

Confusing point: A clique potential can be made up of a product of factors. Suppose clique C_1 has scope A, B and C. The clique potential for C_1 could be $\phi_1(A,B) \times \phi_2(B,C) \times \phi_3(A,C)$.

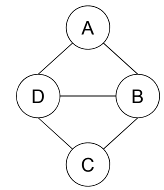
Parameterizations

Examples of Markov networks and their cliques



Cliques:

{A,B}, {B,C}, {C,D}, {A,D}



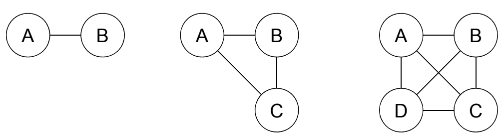
Cliques:

{A,B,D}, {B,C,D},

{A,D},{C,D},{A,B},{B,C}{B,D}

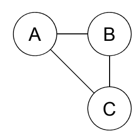
Parameterizations

Note: every complete subgraph is a subset of some (maximal) clique eg.



Because of this, we can reduce the number of factors in our parameterization by allowing factors only for maximal cliques

Parameterizations



The maximal clique for this graph has scope A, B, C.

You can parameterize this in two ways:

1. $P_\Phi(A,B,C) = \phi_1(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})$

or

2. $P_\Phi(A,B,C) = \phi_1(\boldsymbol{A},\boldsymbol{B}) \times \phi_2(\boldsymbol{B},\boldsymbol{C}) \times \phi_3(\boldsymbol{A},\boldsymbol{C})$

Finer-Grained Parameterization

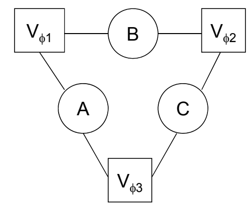
- Markov network structure does not reveal whether the factors in the parameterization involve maximal cliques or subsets of these cliques
- Factor graph makes this explicit in the structure.

Finer-Grained Parameterizations

A factor graph F is an undirected graph containing two types of nodes:

- Variable nodes (denoted as ovals) and
- Factor nodes (denoted as squares).

The graph only contains edges between variable nodes and factor nodes.

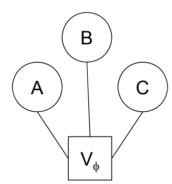


Finer-Grained Parameterizations

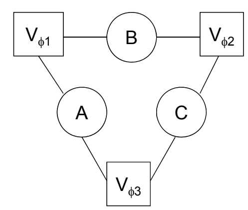
A factor graph F is parameterized by a set of factors, where each factor node V_ϕ is associated with only one factor ϕ, whose scope is the set of variables that are neighbors of V_ϕ in the graph.

A distribution P factorizes over F if it can be represented as a set of factors of this form.

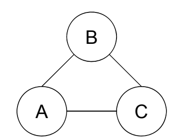
Finer-grained Parameterization



A single factor over all three variables



3 pairwise factors



The induced Markov network

Finer-grained Parameterizations

Rather than encoding factors as complete tables over the scope of the factor, we can use a log-linear model:

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

Where $\varepsilon(\mathbf{D}) = -\ln \phi(\mathbf{D})$ is an energy function (which you want to minimize)

$$P(X_1,\ldots,X_n) \propto \exp\left[-\sum_{i=1}^{m}\varepsilon_i(\mathbf{D}_i)\right]$$

Noote: log representation makes sure the distribution is positive

25

Finer-grained Parameterizations

Let ***D*** be a subset of variables. We define a feature f(***D***) to be a function from ***D*** → *R*.

eg. an indicator feature takes on value 1 for some values $\mathbf{y} \in \mathrm{Val}(\mathbf{D})$ and 0 otherwise

26

Finer-grained Parameterizations

Features provide a compact way to specify certain types of interactions

Example: Suppose A_1 and A_2 can take on l possible values $a^1, \ldots, a^l$. A_1 and A_2 prefer situations when they take on the same value, and have no preference otherwise. The energy function might take the following:

$$\varepsilon(A_1, A_2) = \begin{cases} -10 & A_1 = A_2 \\ 0 & \text{otherwise} \end{cases}$$

27

Finer-grained Parameterizations

(example continued)

Two options for representing the factor:

- As a table, it requires l^2 values
- Log-linear function in terms of a feature $f(A_1,A_2)$ that is an indicator function for the event $A_1=A_2$. The energy function looks like:

$$\varepsilon(A_1, A_2) = 3 * I(A_1 = A_2)$$

We just replaced a table with a function

28

Finer-grained Parameterizations

A distribution P is a log-linear model over a Markov network $\mathcal{H}$ if it is associated with:

- A set of features $F = \{f_1(\mathbf{D}_1), \ldots, f_k(\mathbf{D}_k)\}$, where each $\boldsymbol{D}_i$ is a complete subgraph in $\mathcal{H}$
- A set of weights $w_1, \ldots, w_k$

Such that

$$P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{k} w_i f_i(\boldsymbol{D}_i)\right]$$

29

Finer-grained Parameterizations

3 representations of the parameterization of a Markov network:

1. Markov network: product over potentials on cliques
2. Factor graph: product of factors
3. Set of features: product over feature weights

Finer-grained

Which is most appropriate? Depends on the nature of the problem…

30