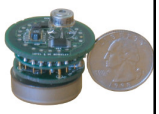
Readings:
K&F: 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7

Markov networks, Factor graphs, and an unified view

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

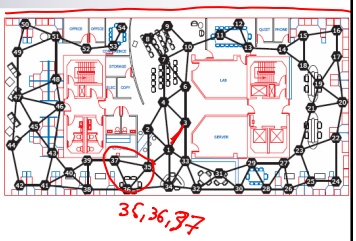
October 27th, 2006

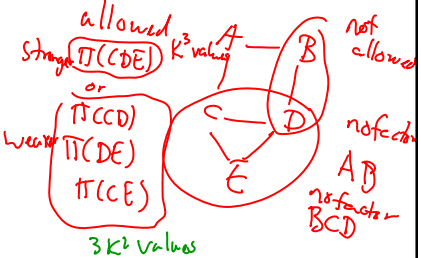
Factorization in Markov networks



- Given an undirected graph *H* over variables $\mathbf{X}=\{X_1,\ldots,X_n\}$
- A distribution *P* **factorizes** over *H* if ∃
  - subsets of variables $\mathbf{D}_1 \subseteq \mathbf{X},\ldots, \mathbf{D}_m \subseteq \mathbf{X}$, such that the $\mathbf{D}_i$ are *fully connected* in *H*
  - *non-negative potentials* (or factors) $\pi_1(\mathbf{D}_1),\ldots, \pi_m(\mathbf{D}_m)$
    - also known as clique potentials
  - such that

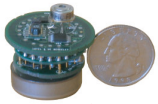
$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i)$$



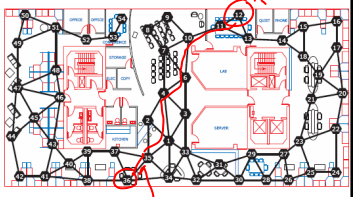


- Also called Markov random field *H*, or Gibbs distribution over *H*

Global Markov assumption in Markov networks



- A path $X_1 - \ldots - X_k$ is **active** when set of variables **Z** are observed if none of $X_i \in \{X_1,\ldots,X_k\}$ are observed (are part of **Z**)



- Variables **X** are **separated** from **Y** given **Z** in graph *H*, $sep_H(\mathbf{X};\mathbf{Y}|\mathbf{Z})$, if there is no active path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given **Z**
- The **global Markov assumption** for a Markov network *H* is

Representation Theorem for Markov Networks

If joint probability distribution *P*:

$$P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m}\pi_i(\mathbf{D}_i)$$

Then

H* is an I-map for *P

If *H* is an I-map for *P* and *P* is a positive distribution

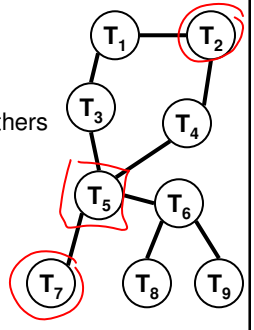
Then

joint probability distribution *P*:

$$P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_{i=1}^{m}\pi_i(\mathbf{D}_i)$$

Local independence assumptions for a Markov network

- **Separation** defines global independencies
- **Pairwise Markov Independence**:
  - Pairs of non-adjacent variables are independent given all others
- **Markov Blanket**:
  - Variable independent of rest given its neighbors



Equivalence of independencies in Markov networks

- **Soundness Theorem**: For all positive distributions *P*, the following three statements are equivalent:
  - *P* entails the global Markov assumptions
  - *P* entails the pairwise Markov assumptions
  - *P* entails the local Markov assumptions (Markov blanket)

Minimal I-maps and Markov Networks

- A fully connected graph is an I-map
- Remember minimal I-maps?
  - A "simplest" I-map → Deleting an edge makes it no longer an I-map
- In a BN, there is no unique minimal I-map
- Theorem: **In a Markov network, minimal I-map is unique!!**
- Many ways to find minimal I-map, e.g.,
  - Take pairwise Markov assumption:
  - If *P* doesn't entail it, add edge:

How about a perfect map?

- Remember perfect maps?
  - independencies in the graph are exactly the same as those in *P*
- For BNs, doesn't always exist
  - counter example: Swinging Couples
- How about for Markov networks?

Unifying properties of BNs and MNs

- BNs:
  - give you: V-structures, CPTs are conditional probabilities, can directly compute probability of full instantiation
  - but: require acyclicity, and thus no perfect map for swinging couples
- MNs:
  - give you: cycles, and perfect maps for swinging couples
  - but: don't have V-structures, cannot interpret potentials as probabilities, requires partition function
- Remember PDAGS???
  - skeleton + immoralities
  - provides a (somewhat) unified representation
  - see book for details

What you need to know so far about Markov networks

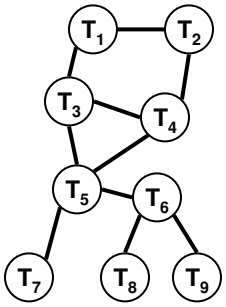
- Markov network representation:
  - undirected graph
  - potentials over cliques (or sub-cliques)
  - normalize to obtain probabilities
  - need partition function
- Representation Theorem for Markov networks
  - if P factorizes, then it's an I-map
  - if P is an I-map, only factorizes for positive distributions
- Independence in Markov nets:
  - active paths and separation
  - pairwise Markov and Markov blanket assumptions
  - equivalence for positive distributions
- Minimal I-maps in MNs are unique
- Perfect maps don't always exist

Some common Markov networks and generalizations

- Pairwise Markov networks
- A very simple application in computer vision
- Logarithmic representation
- Log-linear models
- Factor graphs

Pairwise Markov Networks

- All factors are over single variables or pairs of variables:
  - Node potentials
  - Edge potentials
- Factorization:



- Note that there may be bigger cliques in the graph, but only consider pairwise potentials

A very simple vision application

- Image segmentation: separate foreground from background
- Graph structure:
  - pairwise Markov net
  - grid with one node per pixel
- Node potential:
  - "background color" v. "foreground color"
- Edge potential:
  - neighbors like to be of the same class

Logarithmic representation

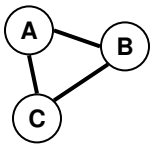
- Standard model: $P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$
- Log representation of potential (assuming positive potential):
  - also called the energy function
- Log representation of Markov net:

Log-linear Markov network (most common representation)

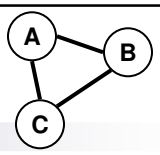
- **Feature** is some function $\phi[\mathbf{D}]$ for some subset of variables **D**
  - e.g., indicator function
- **Log-linear model** over a Markov network *H*:
  - a set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two ϕ's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[\sum_{i=1}^{k} w_i\phi_i(\mathbf{D}_i)\right]$

Structure in cliques

- Possible potentials for this graph:



Factor graphs



- Very useful for approximate inference
  - Make factor dependency explicit
- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i - \phi_j$ if $X_i \in$ Scope$[\phi_j]$
- More explicit representation, but exactly equivalent

Exact inference in MNs and Factor Graphs

- Variable elimination algorithm presented in terms of factors $\rightarrow$ exactly the same VE algorithm can be applied to MNs & Factor Graphs
- Junction tree algorithms also applied directly here:
  - triangulate MN graph as we did with moralized graph
  - each factor belongs to a clique
  - same message passing algorithms

Summary of types of Markov nets

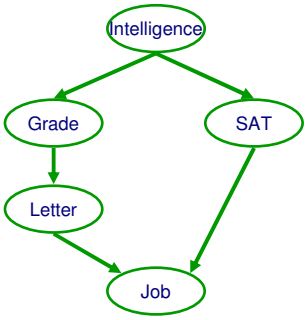
- Pairwise Markov networks
  - very common
  - potentials over nodes and edges
- Log-linear models
  - log representation of potentials
  - linear coefficients learned from data
  - most common for learning MNs
- Factor graphs
  - explicit representation of factors
    - you know exactly what factors you have
  - very useful for approximate inference

What you learned about so far

- Bayes nets
- Junction trees
- (General) Markov networks
- Pairwise Markov networks
- Factor graphs

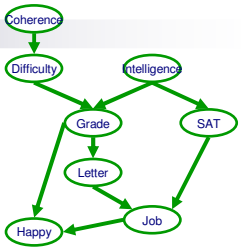
- How do we transform between them?
- More formally:
  - I give you an graph in one representation, find an **I-map** in the other

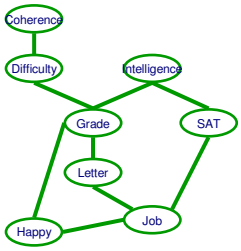
From Bayes nets to Markov nets



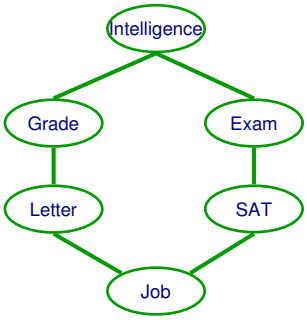
BNs → MNs: Moralization

- **Theorem**: Given a BN *G* the Markov net *H* formed by moralizing *G* is the *minimal I-map* for I(*G*)
- **Intuition**:
  - in a Markov net, each factor must correspond to a subset of a clique
  - the factors in BNs are the CPTs
  - CPTs are factors over a node and its parents
  - thus node and its parents must form a clique
- **Effect**:
  - **some** independencies that could be read from the BN graph become hidden



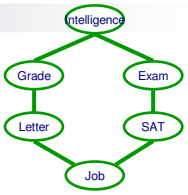


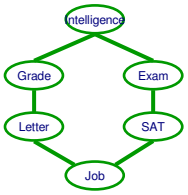
From Markov nets to Bayes nets



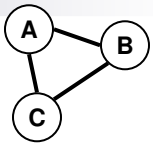
MNs → BNs: Triangulation

- **Theorem**: Given a MN *H,* let *G* be the Bayes net that is a *minimal I-map* for I(*H)* then *G* must be **chordal**
- **Intuition**:
  - v-structures in BN introduce immoralities
  - these immoralities were not present in a Markov net
  - the triangulation eliminates immoralities
- **Effect**:
  - **many** independencies that could be read from the MN graph become hidden





Markov nets v. Pairwise MNs



- Every Markov network can be transformed into a Pairwise Markov net
  - introduce extra "variable" for each factor over three or more variables
  - domain size of extra variable is exponential in number of vars in factor
- **Effect**:
  - any local structure in factor is lost
  - a chordal MN doesn't look chordal anymore

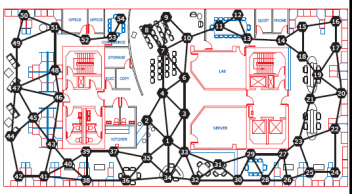
Overview of types of graphical models and transformations between them

Approximate inference overview

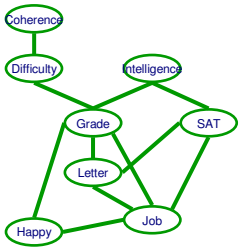
- So far: VE & junction trees
  - exact inference
  - exponential in tree-width
- There are many many many many approximate inference algorithms for PGMs
- We will focus on three representative ones:
  - sampling
  - variational inference
  - loopy belief propagation and generalized belief propagation
- There will be a special recitation by Pradeep Ravikumar on more advanced methods

Approximating the posterior v. approximating the prior

- Prior model represents entire world
  - world is complicated
  - thus prior model can be very complicated



- Posterior: after making observations
  - sometimes can become much more sure about the way things are
  - sometimes can be approximated by a simple model



- First approach to approximate inference: **find *simple* model that is "close" to posterior**
- Fundamental problems:
  - **what is close?**
  - **posterior is intractable result of inference, how can we approximate what we don't have?**

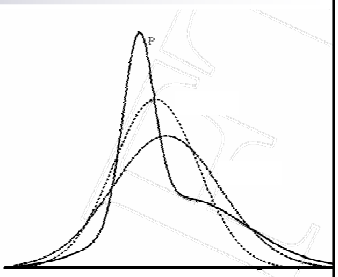
KL divergence: Distance between distributions

- Given two distributions *p* and *q* KL divergence:

- D(p||q) = 0 iff p=q
- Not symmetric – p determines where difference is important
  - p(x)=0 and q(x)≠0
  - p(x)≠0 and q(x)=0

Find simple approximate distribution

- Suppose *p* is intractable posterior
- Want to find simple *q* that approximates *p*
- KL divergence not symmetric
- D(p||q)
  - true distribution p defines support of diff.
  - the "correct" direction
  - will be intractable to compute
- D(q||p)
  - approximate distribution defines support
  - tends to give overconfident results
  - will be tractable



Back to graphical models

- Inference in a graphical model:
  - P(**x**) =
  - want to compute P(X_i|**e**)
  - our *p*:
- What is the simplest *q*?
  - every variable is independent:
  - mean-fields approximation
  - can compute any prob. very efficiently

D(p||q) for mean-fields – KL the right way

- p:
- q:
- D(p||q)=

D(q||p) for mean-fields – KL the reverse direction

- p:
- q:
- D(p||q)=

What you need to know so far

- Goal:
  - Find an efficient distribution that is close to posterior
- Distance:
  - measure distance in terms of KL divergence
- Asymmetry of KL:
  - $D(p||q) \neq D(q||p)$
- The right KL is intractable, so we use the reverse KL