Lecture 3:

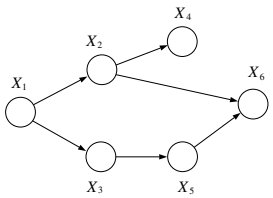
Undirected Graphical Models

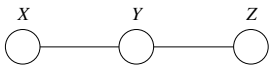
Sam Roweis

January 14, 2002

Review: Goal of Graphical Models

- Graphical models aim to provide a *compact factorization* of large joint probability distributions.
- The factorization is achieved using *local functions* which exploit *conditional independencies* in the models.
- The graph tells us a basic set of *conditional independencies* that must be true. (Although more may be true). From these we can derive more: these are crucial to developing efficient algorithms valid for *all* numerical settings of the local functions.
- The local functions tell us the quantitative details of the distribution.





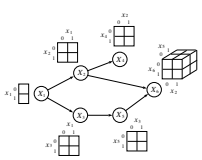
Outline

- Review
- Undirected Model Semantics
- Clique Potentials
- Boltzmann Distribution and Partition Function
- Examples: Ising and Gaussian

Review: Directed Models (Bayes Nets)

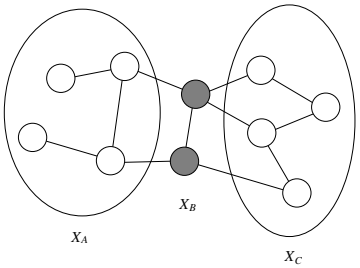
- Semantics: $\mathbf{x} \perp \mathbf{y} | \mathbf{z}$ if $\mathbf{z}$ d-separates $\mathbf{x}$ and $\mathbf{y}$
- d-separation: $\mathbf{z}$ d-separates $\mathbf{x}$ from $\mathbf{y}$ if along every undirected path between $\mathbf{x}$ and $\mathbf{y}$ there is a node $\mathbf{w}$ such that either:
  1. $\mathbf{w}$ has converging arrows along the path $(\rightarrow \mathbf{w} \leftarrow)$ and neither $\mathbf{w}$ nor its descendatns are in $\mathbf{z}$ or
  2. $\mathbf{w}$ does not have converging arrows along the path and $\mathbf{w} \in \mathbf{z}$.
- The "Bayes-Ball" algorithm can be used to check d-separation.
- Always possible to find a distibution consistent with the graph. The most general such distribution is a product of parental conditionals:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$



Undirected Models

- Once again we have a graph with one node per random variable and edges that connect pairs of nodes, but now the edges are undirected.
- Semantics: every node is conditionally independent from its non-neighbours given its neighbours, i.e. $\mathbf{x}_A \perp \mathbf{x}_C \mid \mathbf{x}_B$ if every path b/w $\mathbf{x}_A$ and $\mathbf{x}_C$ goes through $\mathbf{x}_B$



- "Markov Ball" algorithm: remove $\mathbf{x}_B$ and see if there is any path from $\mathbf{x}_A$ to $\mathbf{x}_C$.
- Can model symmetric interactions in a way that directed models cannot.
- aka Markov Random Fields, Markov Networks, Boltzmann Machines, Spin Glasses, Ising Models

Parametrization

- In directed models, we started with $p(\mathbf{X}) = \prod_i p(\mathbf{x}_i|\mathbf{x}_{\pi_i})$ and we derived the d-separation semantics from that.
- Now: have semantics, need parametrization.
- What about this:

$$p(\mathbf{X}) = \prod_i p(\mathbf{x}_i|\mathbf{x}_{\text{neighbours}(i)})$$

- Good, because it is the product of local functions each with a simple conditional interpretation. Tricky, because now the local functions cannot be arbitrary, but must agree properly in order to define a valid distribution.
- OK, what about:

$$p(\mathbf{X}) = \prod_i p(\mathbf{x}_i, \mathbf{x}_{\text{neighbours}(i)})$$

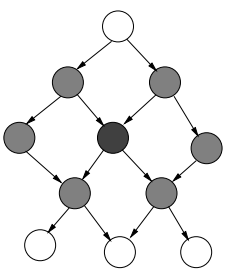
- Good: product of local functions with simple marginal interpretations.
  Bad, because only very few pathalogical marginals on overalpping nodes can be multiplied to give a valid joint.

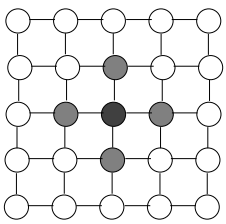
Markov Blanket

- $\mathbf{b}$ is a "markov blanket" for $\mathbf{x}$ iff

$$\mathbf{x} \perp \mathbf{y} \mid \mathbf{b} \quad \forall\, \mathbf{y} \notin \mathbf{b}$$

- "Markov boundary": minimal markov blanket
- For undirected models, this is set of neighbours.
- Q: What is markov blanket in directed model?
  A: {parents+children+parents-of-children}





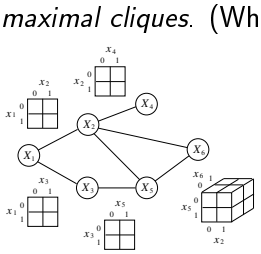
Clique Potentials

- Whatever factorization we pick, only connected nodes can be arguments of a single local fn.
- A *clique* is a fully connected subset of nodes.
- Thus, consider the form

$$\mathrm{P}(\mathbf{X}) = \frac{1}{Z} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{\mathbf{X}} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c)$$

- This parametrization is the product of functions that don't need to agree with each other.
- This parametrization still factors in the way that the graph semantics demand.
- Without loss of generality we can restrict ourselves to *maximal cliques*. (Why?)



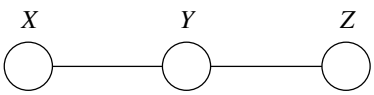
Boltzmann Distrib. & Partition Fn.

- We often represent the clique potentials as
$$\psi_C(\mathbf{x}_C) = \exp\{-H_C(\mathbf{x}_C)\}$$
for arbitrary real valued functions $H_C(\mathbf{x}_C)$. The negative sign is a standard convention.
- This gives the joint a nice additive structure:
$$\mathsf{P}(\mathbf{X}) = \frac{1}{Z}\exp\{-\sum_{\text{cliques } C} H_C(\mathbf{x}_c)\}$$
where the sum in the exponent is called the "energy" $H(\mathbf{x}) = \sum_C H_C(\mathbf{x}_c)$; this is the "Boltzmann distribution" from statistical physics.
- Normalizer Z is called the "partition function".
- Computing the normalizer and its derivatives can often be the hardest part of inferene and learning in undirected models.
- Often the factored structure of the distribution makes it possible to efficiently do the sums/integrals required to compute Z.
- Don't *always* have to compute Z, e.g. for conditional probabilities.

Hammersley-Clifford Theorem (1971)

- H-C theorem tells us that the family of distributions defined by the conditional independence semantics on the graph and the family defined by products of potential functions on maximal cliques are the same.
- Tricky point: the potential functions are arbitrary real valued, but strictly positive.

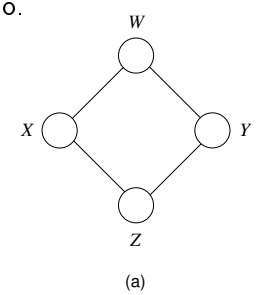
Interpretation of Clique Potentials



- The model implies $\mathbf{x} \perp \mathbf{z} \mid \mathbf{y}$
$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})p(\mathbf{z}|\mathbf{y})$$
- We can write this as:
$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x},\mathbf{y})p(\mathbf{z}|\mathbf{y}) = \psi_{\mathbf{xy}}(\mathbf{x},\mathbf{y})\psi_{\mathbf{yz}}(\mathbf{y},\mathbf{z})$$
$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{z},\mathbf{y}) = \psi_{\mathbf{xy}}(\mathbf{x},\mathbf{y})\psi_{\mathbf{yz}}(\mathbf{y},\mathbf{z})$$
cannot have all potentials be marginals
cannot have all potentials be conditionals
- The clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables.

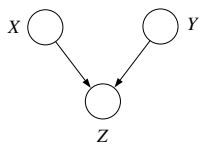
Expressive Power

- Can we always convert directed $\leftrightarrow$ undirected?
- No.



(a)

No directed model can represent these and only these independencies.
$\mathbf{x} \perp \mathbf{y} \mid \{\mathbf{w},\mathbf{z}\}$
$\mathbf{w} \perp \mathbf{z} \mid \{\mathbf{x},\mathbf{y}\}$

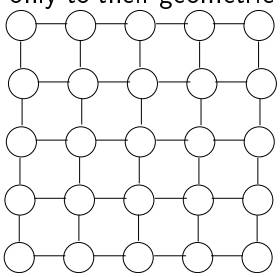


(b)

No undirected model can represent these and only these independencies.
$\mathbf{x} \perp \mathbf{y}$

Example: Ising Models

- For binary valued nodes, a very common model is a spin-glass or Ising lattice.
- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbours.

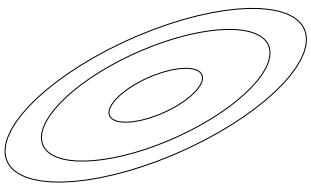


- For example, if we think of each node as a pixel, we might want to encourage nearby pixels to have similar intensities.
- Energy is of the form:

$$H(\mathbf{x}) = \sum_{ij} \beta_{ij} \mathbf{x}_i \mathbf{x}_j + \sum_i \alpha_i \mathbf{x}_i$$

Example: Gaussian Distribution

- The most common and important undirected graphical model on a set of continuous valued nodes is the Gaussian (normal).
- It uses pairwise potentials between every pair of nodes to define an energy identical to the Ising model, but for continuous values:

$$H(\mathbf{x}) = \sum_{ij} (\mathbf{x}_i - \mu_i) V_{ij} (\mathbf{x}_j - \mu_j)$$

where μ is the mean and V is the inverse covariace matrix.