Lecture 3:

Undirected Graphical Models

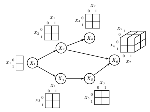
Sam Roweis

January 13, 2003

Review: Directed Models (Bayes Nets)

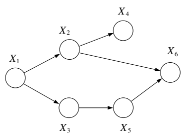
- Semantics: $\mathbf{x} \perp \mathbf{y} | \mathbf{z}$ if $\mathbf{z}$ d-separates $\mathbf{x}$ and $\mathbf{y}$
- d-separation: $\mathbf{z}$ d-separates $\mathbf{x}$ from $\mathbf{y}$ if along every undirected path between $\mathbf{x}$ and $\mathbf{y}$ there is a node $\mathbf{w}$ such that either:
  1. $\mathbf{w}$ has converging arrows along the path ($\rightarrow \mathbf{w} \leftarrow$) and neither $\mathbf{w}$ nor its descendents 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.
- It is always possible to find a distibution consistent with the graph. Most general such distribution is a product of parent-conditionals:

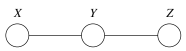
$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$



Review: Goal of Graphical Models

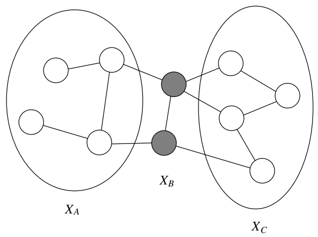
- Graphical models aim to provide *compact factorizations* of large joint probability distributions.
- These factorizations are achieved using *local functions* which exploit *conditional independencies* in the models.
- The graph tells us a basic set of *conditional independencies* that must be true. From these we can derive more that also must be true. These independencies are crucial to developing efficient algorithms valid for *all* numerical settings of the local functions.
- Local functions tell us the quantitative details of the distribution.
- Certain numerical settings of the distribution may have more independencies present, but these do not come from the *graph*.





Undirected Models

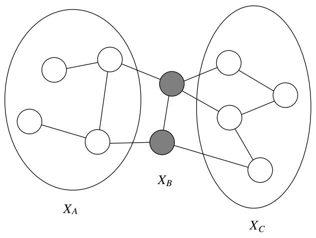
- Also graphs with one node per random variable and edges that connect pairs of nodes, but now the edges are undirected.
- Semantics: every node is conditionally independent from its non-neighbours given its neighbours, i.e.
  $\mathbf{x}_A \perp \mathbf{x}_C \,|\, \mathbf{x}_B$ if every path b/w $\mathbf{x}_A$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



- Can model symmetric interactions that directed models cannot.
- aka Markov Random Fields, Markov Networks, Boltzmann Machines, Spin Glasses, Ising Models

Simple Graph Separation

- In undirected models, simple graph separation (as opposed to d-separation) tells us about conditional independencies.
- $\mathbf{x}_A \perp \mathbf{x}_C | \mathbf{x}_B$ if every path between $\mathbf{x}_A$ and $\mathbf{x}_C$ is blocked by some node in $\mathbf{x}_B$.



- "Markov Ball" algorithm:
  remove $\mathbf{x}_B$ and see if there is any path from $\mathbf{x}_A$ to $\mathbf{x}_C$.

Conditional Parameterization?

- 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.
- Undirected models: have the semantics, need parametrization.
- What about this "conditional parameterization"?

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

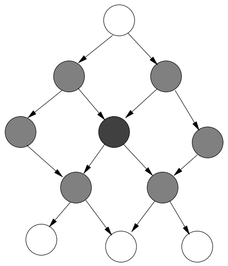
- Good: product of local functions.
  Good: each one has a simple conditional interpretation.
  Bad: local functions cannot be arbitrary, but must agree properly in order to define a valid distribution.

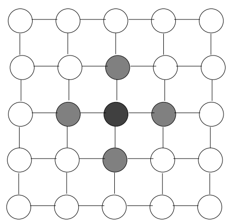
Markov Blanket

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

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

- "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}





Marginal Parameterization?

- OK, what about this "marginal parameterization"?

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

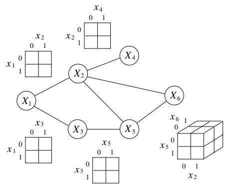
- Good: product of local functions.
  Good: each one has a simple marginal interpretation.
  Bad: only very few pathalogical marginals on overalpping nodes can be multiplied to give a valid joint.

Clique Potentials

- Whatever factorization we pick, we know that only connected nodes can be arguments of a single local function.
- A *clique* is a fully connected subset of nodes.
- Thus, consider using a *product of positive clique potentials*:

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

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



Partition Function

- Normalizer $Z(\mathbf{X})$ above 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.

Boltzmann Distributions

- We often represent the clique potentials using their logs:

$$\psi_C(\mathbf{x}_C) = \exp\{-H_C(\mathbf{x}_C)\}$$

  for arbitrary real valued "energy" functions $H_C(\mathbf{x}_C)$.
  The negative sign is a standard convention.
- This gives the joint a nice additive structure:

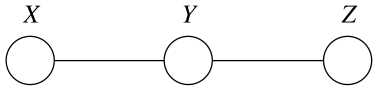
$$\mathsf{P}(\mathbf{X}) = \frac{1}{Z} \exp\{-\sum_{\text{cliques } C} H_C(\mathbf{x}_c)\} = \frac{1}{Z} \exp\{-H(\mathbf{X})\}$$

  where the sum in the exponent is called the "free energy":

$$H(\mathbf{X}) = \sum_C H_C(\mathbf{x}_c)$$

- This way of defining a probability distribution based on energies is the "Boltzmann distribution"from statistical physics.

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:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}, \mathbf{y})p(\mathbf{z}|\mathbf{y}) = \psi_{\mathbf{xy}}(\mathbf{x}, \mathbf{y})\psi_{\mathbf{yz}}(\mathbf{y}, \mathbf{z})$$
$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{z}, \mathbf{y}) = \psi_{\mathbf{xy}}(\mathbf{x}, \mathbf{y})\psi_{\mathbf{yz}}(\mathbf{y}, \mathbf{z})$$

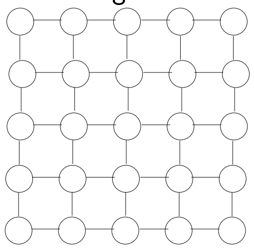
  cannot have all potentials be marginals
  cannot have all potentials be conditionals
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

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.
- For directed models, there is a version of this theorem which tells us that the family of distributions defined by the conditional independences semantics of the directed graph and the family defined by products of parent-conditionals are the same.
- Notice the crucial difference between *graphs*, which tells us independencies that are true *no matter what* local functions we choose, and *numerical functions* which could introduce some extra independencies, once we know them.

Example: Ising Models

- Common model for binary nodes: spin-glass/ Ising lattice.
- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbours.

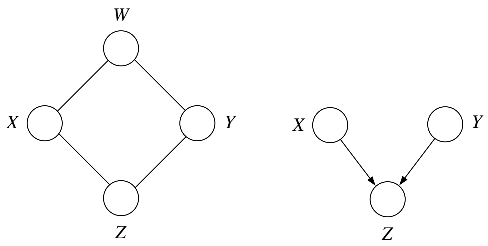


- For example, if we think of each node as a pixel, we might want to encourage nearby pixels to have similar intensities.
- Energy is of the form:

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

Expressive Power

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



(a) (b)

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

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

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.