Lecture 6: Examples of Bayesian Networks and Markov Networks

Zhenke Wu
Department of Biostatistics, University of Michigan

September 22, 2016

Lecture 5 Main Points Once Again

- Bayesian network $(\mathcal{G}, P)$
  - Directed acyclic graph (DAG): $\mathcal{G}$, comprised of nodes V and edges E
  - Joint distribution P over $|V|$ random variables
  - P is Markov to $\mathcal{G}$ **if** variables in P satisfy $X_A \perp X_B \mid X_C$ whenever C d-separates A and B as read off from $\mathcal{G}$
- Markov network $(\mathcal{H}, P)$
  - Undirected graph (UG): $\mathcal{H}$, comprised of nodes V and edges E
  - Joint distribution P over $|V|$ random variables
  - P is Global Markov to $\mathcal{H}$ **if** variables in P satisfy $X_A \perp X_B \mid X_C$ whenever C separates A and B as read off from the graph
- Roughly, given Markov properties, graph $\mathcal{G}$, or $\mathcal{H}$ is a valid guide to understand the variable relationships in distribution P

Lecture 5 Main Points Once Again (continued)

- **Question:** Given a distribution P that is Markov to a DAG $\mathcal{G}$, can we find an UG $\mathcal{H}$ with the same set of nodes so that P is also Markov to it? (Yes, by **moralization**—"marrying the parents". But UG could lose some d-separations, e.g., v-structure; won't lose any if $\mathcal{G}$ is already moralized.)
- (Question above, but with DAG and UG reversed) (Yes, by constructing directed edges following certain node ordering. But DAG could lose some separations, e.g., four-node loop)
- Are there distributions representable by both DAG and UG, but without loss of (d-)separations? (Yes.) If so, under what conditions? (Those distributions either are Markov to a **chordal Markov network**, or to a DAG without immoralities.)
- **Definition** (chordal Markov network): every one of its loops of length ≥ 4 possesses a chord, where a chord in the loop is an edge (from the original graph) connecting X_i and X_j for two nonconsecutive nodes (with respect to the loop).

Markov Network Example: Ising Model

- A mathematical model of ferromagnetism in statistical mechanics; Named after physicist Ernst Ising;
- The model consists of discrete variables that represent magnetic dipole moments of atomic spins that can be in one of two states (+1 or −1).
- The spins are arranged in a graph, usually a lattice, allowing each spin to interact with its neighbors.

Markov Network Example: Ising Model

- **Formulation**: Let $\mathcal{H} = (V, E)$ be an undirected graph, e.g., (lattice or non-lattice). Let the binary random variables $X_i \in \{-1, +1\}$. The Ising model takes the form

$$P(\mathbf{x}; \theta) \propto \exp\left(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right)$$

- From the model form, Ising model is positive and Markov to $\mathcal{H}$. Using the local Markov property, and code the -1 into 0, the conditional distribution for a node X_i given all its neighbors is given by a logisitic regression:

$$Pr(X_i = 1 \mid X_j, j \neq i; \theta) = Pr(X_i = 1 \mid X_j, (i,j) \in E; \theta)$$

$$= sigmoid(\theta_i + \sum_{j:(i,j) \in E} \theta_{ij} x_j)$$

Markov Network Example: Special case of Ising Model

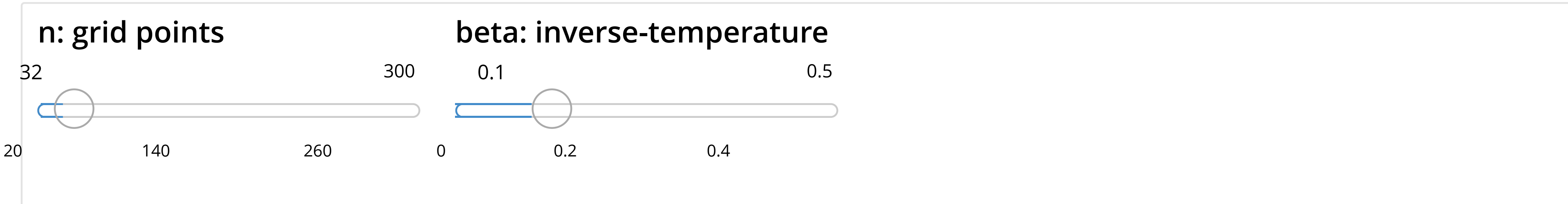
- No external field: $\theta_i = 0, X_i \in V$
- $\theta_{ij} = \beta J$, $\forall i, j$.
- We have

$$P(\mathbf{x}; \theta) \propto \exp\left(\beta \cdot J \cdot \sum_{(i,j)\in E} x_i x_j\right)$$

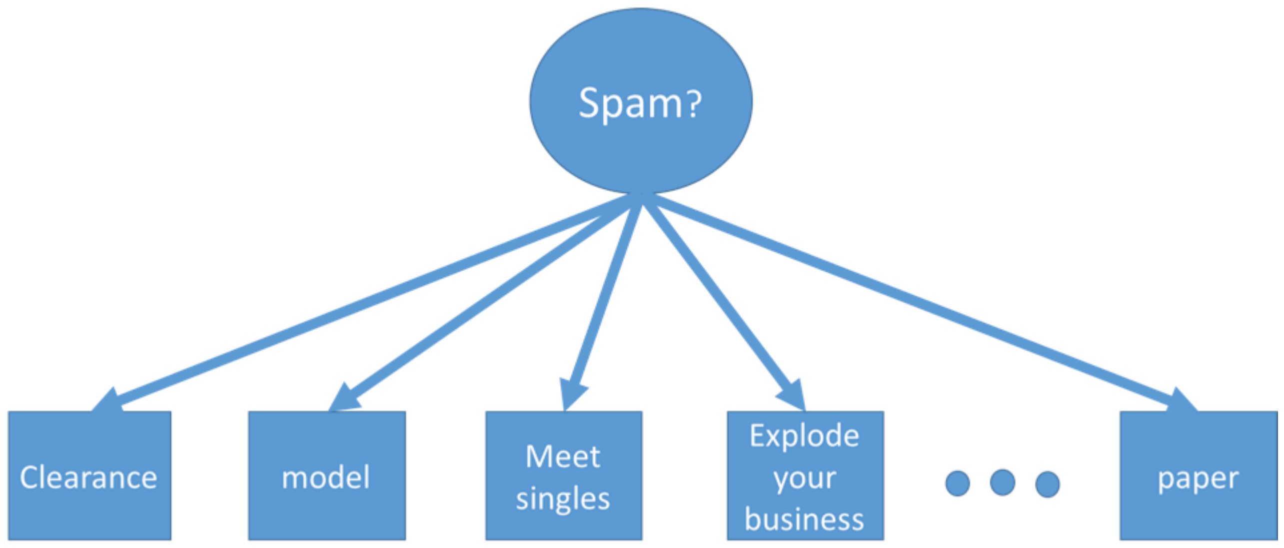
- β: inverse temperature; large β, lower temperature (colder)
- $J > 0$: neighboring nodes tend to align, so-called ferromagnetic model; $J < 0$: anti-ferromagnetic.

Square-Lattice Ising Model under Different Temperatures

- $P(\mathbf{x};\theta) \propto \exp\left(\beta \cdot J \cdot \sum_{(i,j)\in E} x_i x_j\right)$
  - Set $J = 2$, ferromagnetic
  - (Run `Lecture6.Rmd` in RStudio)
    - Vary inverse temperature: β
    - Try different graph size: n^2

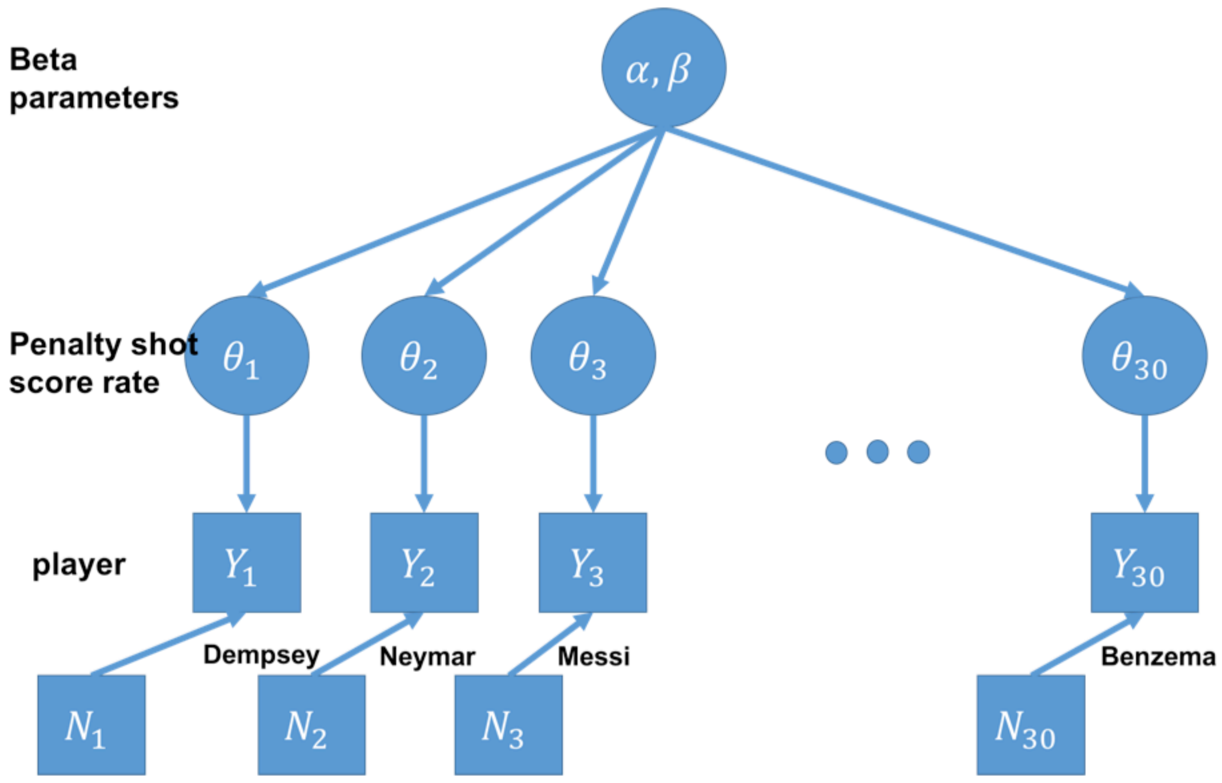


Bayesian Network Example: Naive Bayes for SPAM classification



- Features (words) assumed **independent** given SPAM or HAM status, hence "naive"
- Infer the SPAM status given observed evidence from the email
- Very fast, low storage requirements, robust to irrelevant features, good for benchmarking

Bayesian Network Example: Beta-Binomial Model



- 30 soccer players' penalty shot score rates and the actual number of shots
- What's the best estimate of a player's scoring rate? (empirical Bayes estimate)
- Information from other players could contribute to a given player's score rate estimate. Use moralized graph to explain.

Inference for Bayesian Network: Moralization

- **Question:** given observed evidence, what's the updated probability distribution for those unobserved variables? Or more specifically, which conditional independencies still hold, which don't?
- **Proposition 4.7** Let $\mathcal{G}$ be a Bayesian Network over $\mathbf{V}$ and $\mathbf{Z} = \mathbf{z}$ an observation. Let $\mathbf{W} = \mathbf{V} - \mathbf{Z}$. Then $P_{\mathcal{G}}(\mathbf{W} \mid \mathbf{Z} = \mathbf{z})$ is a Gibbs distribution defined by factors $\Phi = \{\phi_{X_i}\}_{X_i \in \mathbf{V}}$, where $\phi_{X_i} = P_{\mathcal{G}}(X_i \mid Pa_{X_i})[\mathbf{Z} = \mathbf{z}]$. The partition function for this Gibbs distribution is $P_{\mathcal{G}}(\mathbf{Z} = \mathbf{z})$, the marginal probability.
- Use the moralized graph to identify conditional independencies given observed data.
- Because the Gibbs distribution above factorizes according to a moralized graph $M(\mathcal{G})$ which creates cliques for a family (parents and a child).
- And P factorizing with respect to $M(\mathcal{G})$ amounts to P satisfying the Markov property. This means you can use the moralized graph as a "map", albeit it could miss some original conditional independence information.

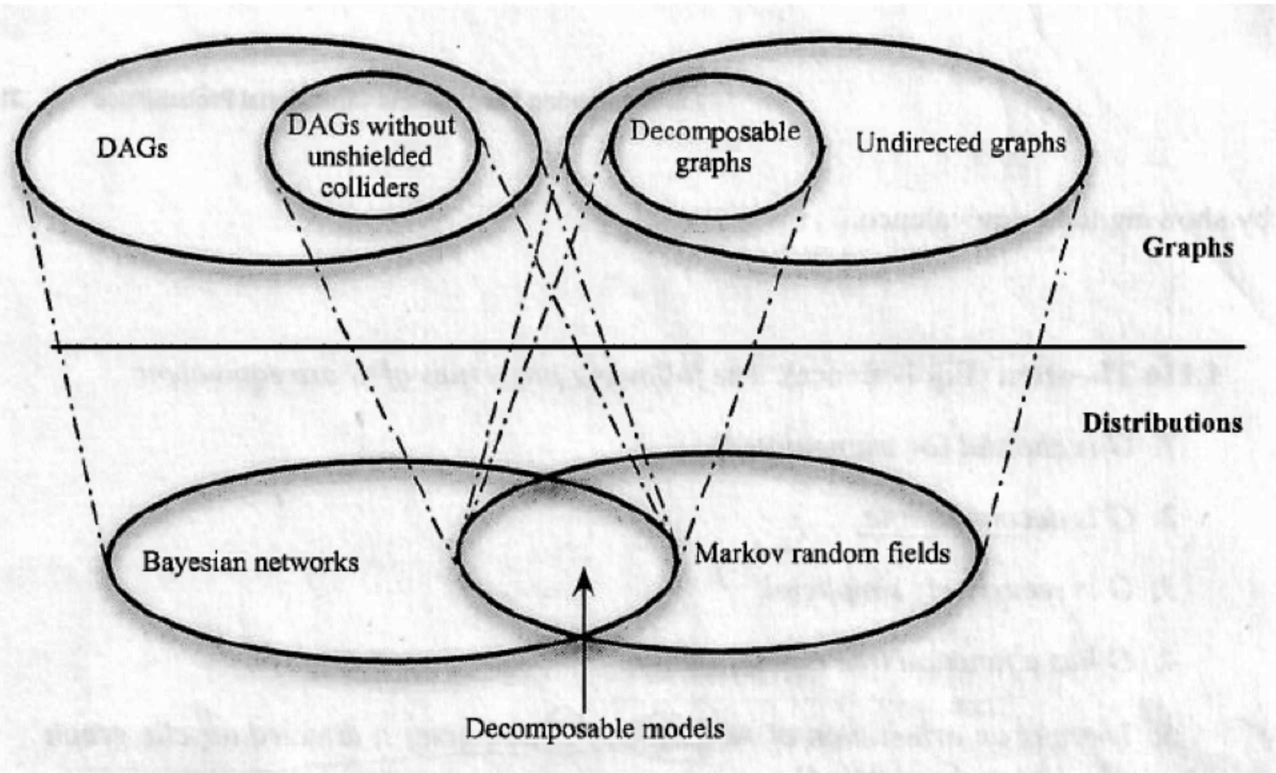
Moralized Graph

- Naturally, if a Bayesian network is already moral (parents are connected by directed edges), then moralization will not add extra edges and conditional independencies will not be lost.
- So in this case separations in UG $M(\mathcal{G})$ correspond one-to-one for d-separations in the original DAG $\mathcal{G}$.

Chordal Graph

- If $\mathcal{H}$ is an UG, and let $\mathcal{G}$ be any DAG that is minimal I-map for $\mathcal{H}$, then $\mathcal{G}$ must have no immoralities. [Proof]
- Nonchordal DAGs must have immoralities
- $\mathcal{G}$ then must be chordal
- The conditional independencies encoded by an undirected chordal graph can be perfectly encoded by a directed graph. (Use clique tree proof)
- If $\mathcal{H}$ is nonchordal, no DAG can encode **perfectly** the same set of conditional independencies as in $\mathcal{H}$. (Use the third bullet point.)

The connections among graphs and distributions (note from Lafferty, Liu and Wasserman)



- The intersection of Bayesian networks and Markov networks (or random fields) are those distributions Markov to a chordal Markov network or to a DAG without immoralities.
- Chordal graph ⇔ decomposable graph

Comment

- **Next Lecture**: Overview of Module 2 that discusses inference: more algorithmic-flavored and exciting ideas. Begin exact inference.
- **No required reading**.
- **Homework 1** due 11:59PM, October 3rd, 2016 to Instructor's email.