Lecture 5

Learning with Graphs

- We look at the basic theory behind learning graphical models.
- We wont look at the causal side.

Overview

- **The Search Space**
- Independence
- Evidence
- The Exponential Family
- Putting it All Together

The Parametric Model Space

- For K binary variables:
  - 2^K data vectors
  - 2^{2^K} sets of data vectors
- The *saturated* model (graph with all $K(K-1)/2$ arcs present) has 2^K parameters.
- Worst case analysis shows $O(2^K)$ examples needed to learn DAGs, and the factor a low order polynomial in error and confidence (see Hoffgen '93; Friedman and Yakhini '96).
- If we restrict graphs to no more than P parents, this becomes worst case analysis gives $O(2^P)$ (see Hoffgen '93).

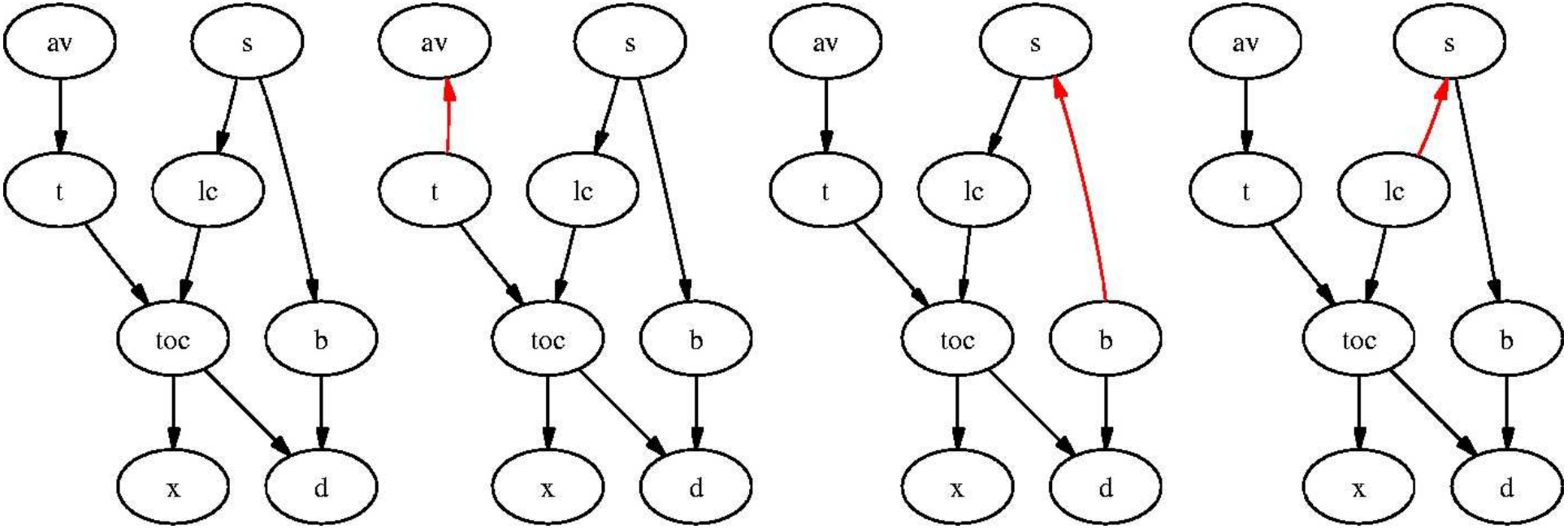
The Space of Graphs

- There are $C_2^K = K(K-1)/2$ arcs on a graph with K variables, thus there are $2^{K(K-1)/2}$ possible undirected graphs.
- For DAGs it is more complex than $3^{K(K-1)/2}$ (i.e., each potential arc has 3 possibilities, left,right or absent) as covered by Volf and Studeny 1999:
  - Different DAGs can represent equivalent functional forms.
  - DAGs cannot have cycles.

Number of vertices	2	3	4	5	K
Number of UGs	2	8	64	1024	$2^{K(K-1)/2}$
Number of equivalent DAGs	2	11	185	8728	$< 3^{K(K-1)/2}$

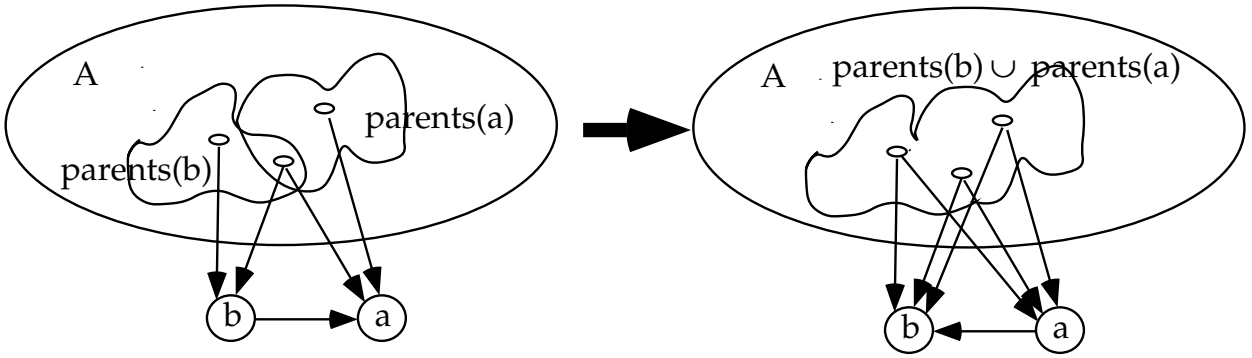
Equivalent DAGS, example

Here are some of the equivalent graphs for our old favorite.



Equivalent DAGS

Theorem (quiet easy, various authors): Two DAGs have the same independence properties iff they have the same shape (ignoring directions), and moralize to the same undirected graph.



Note: consider two nodes a and b adjacent in the partial order (so $b \in \mathrm{par}(a)$) represented by the DAG. Lets switch their order. The new conditional distribution for a, now before b, is

$$\sum_b p(a|\mathrm{par}(a))p(b|\mathrm{par}(b)) = \; p(a|\mathrm{par}(b) \cup \mathrm{par}(a) - \{b\})$$

(1) switch their order iff they share the same common parents.

(2) can do initial separate switches independently

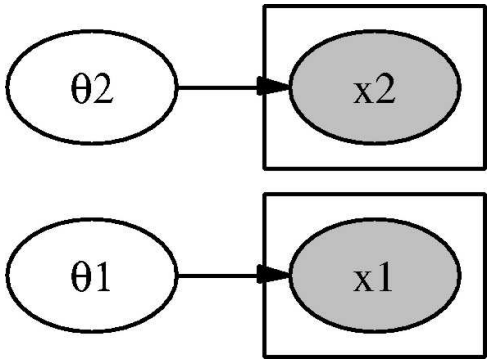
The Search Space

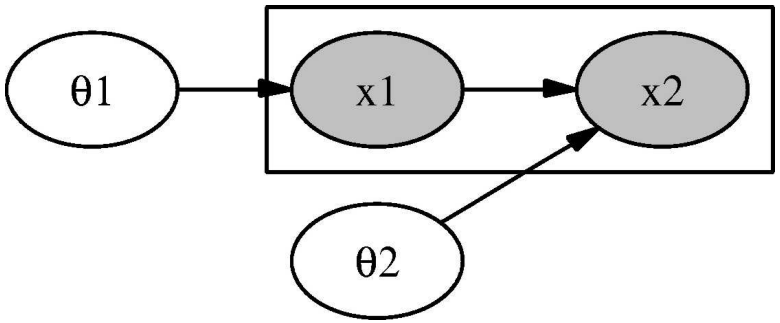
The parametric space for the saturated model has less dimension than the number of graphs!

- Different graphs may share stand-alone sub-structures and parametric models.
- Only bother with learning graphs if you really expect an answer much simpler than the saturated graph.
- Many ways to restrict this:
  - Penalize more complex graphs.
  - Allow no more than P parents.
  - Fix an ordering of variables to restrict parents.

Simple Example

Lets look at N samples of boolean data according to one of two models. The data is totalled into the table below as "sufficient statistics".

$\mathscr{M}_1 =$ 

$\mathscr{M}_2 =$ 

	$x_1 = 0$	$x_1 = 1$	marginal
$x_2 = 0$	$n_{0,0}$	$n_{1,0}$	$n_{\cdot,0}$
$x_2 = 1$	$n_{0,1}$	$n_{1,1}$	$n_{\cdot,1}$
marginal	$n_{0,\cdot}$	$n_{1,\cdot}$	N

Simple Example, Model $\mathscr{M}_1$

Model $\mathscr{M}_1$ is as follows:

$$x_1 \sim \text{Boolean}(\theta_0),\ x_2 \sim \text{Boolean}(\theta_2)$$

and θ_1, θ_2 aprior independent. **Likelihood** is

$$p(\vec{x}_1, \vec{x}_2 \,|\, \theta_1, \theta_2, \mathscr{M}_1) \ = \ \theta_1^{n_{0,\cdot}} (1-\theta_1)^{n_{1,\cdot}} \quad \theta_2^{n_{\cdot,0}} (1-\theta_2)^{n_{\cdot,1}}$$

Using Beta priors on θ_1, θ_2 with parameters α_1, α_2, we get **model likelihood** of

$$p(\vec{x}_1, \vec{x}_2 \,|\, \mathscr{M}_1) \ = \ \frac{\Gamma(\alpha_1 + n_{0,\cdot})\Gamma(\alpha_2 + n_{1,\cdot})\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1 + \alpha_2 + n_{\cdot,\cdot})} \\ \frac{\Gamma(\alpha_1 + n_{\cdot,0})\Gamma(\alpha_2 + n_{\cdot,1})\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1 + \alpha_2 + n_{\cdot,\cdot})}$$

Simple Example, Model $\mathscr{M}_2$

Model $\mathscr{M}_2$ is as follows: $x_1 \sim \text{Boolean}(\theta_0)$, $x_2|x_1=0 \sim \text{Boolean}(\theta_{2,0})$, $x_2|x_1=1 \sim \text{Boolean}(\theta_{2,1})$, where $\theta_2 = (\theta_{2,0}, \theta_{2,1})$, and θ_1, θ_2 aprior independent. **Likelihood** $p(\vec{x}_1, \vec{x}_2 \,|\, \theta_1, \theta_2, \mathscr{M}_2)$ is

$$\theta_1^{n_{0,\cdot}}(1-\theta_1)^{n_{1,\cdot}} \quad \theta_{2,0}^{n_{0,0}}(1-\theta_{2,0})^{n_{0,1}} \; \theta_{2,1}^{n_{1,0}}(1-\theta_{2,1})^{n_{1,1}}$$

Using Beta priors on $\theta_1, \theta_{2,0}, \theta_{2,1}$ with parameters α_1, α_2, we get **model likelihood** of

$$\begin{aligned} p(\vec{x}_1, \vec{x}_2 \,|\, \mathscr{M}_2) \;=\; & \frac{\Gamma(\alpha_1+n_{0,\cdot})\Gamma(\alpha_2+n_{1,\cdot})\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1+\alpha_2+n_{\cdot,\cdot})} \\ & \frac{\Gamma(\alpha_1+n_{0,0})\Gamma(\alpha_2+n_{0,1})\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1+\alpha_2+n_{0,\cdot})} \\ & \frac{\Gamma(\alpha_1+n_{1,0})\Gamma(\alpha_2+n_{1,1})\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1+\alpha_2+n_{1,\cdot})} \end{aligned}$$

Simple Example, cont.

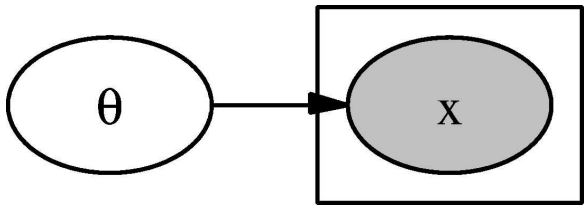
- $\mathscr{M}_1$ has full independence and the likeliihoods only use marginal statistics.
- $\mathscr{M}_2$ is the same for x_1 but the x_2 component now splits into two parts, one for each partition $x_1 = 0$ and $x_1 = 1$.
- Both likelihoods $p(\vec{x}_1, \vec{x}_2 \,|\, \mathscr{M}_1)$ and $p(\vec{x}_1, \vec{x}_2 \,|\, \mathscr{M}_2)$ are made up of terms on Gamma functions which form the normalizing constant for the Beta distribution.
- Their ratio $\mathscr{M}_1$ on $\mathscr{M}_2$ is a good independence test ($>>1$ supports independence):

$$\frac{\Gamma(\alpha_1 + n_{.,0})\Gamma(\alpha_2 + n_{.,1})}{\Gamma(\alpha_1 + \alpha_2 + n_{.,.})} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2 + n_{0,.})}{\Gamma(\alpha_1 + n_{0,0})\Gamma(\alpha_2 + n_{0,1})} \frac{\Gamma(\alpha_1 + \alpha_2 + n_{1,.})}{\Gamma(\alpha_1 + n_{1,0})\Gamma(\alpha_2 + n_{1,1})}$$

Overview

- The Search Space
- **Independence**
- Evidence
- The Exponential Family
- Putting it All Together

Independence in Learning: Simple



Likelihood of sample of size N as a vector of data $\vec{x}$, given model $\mathscr{M}$ and parameters θ (possibly a vector), assuming *i.i.d.* data:

$$p(\vec{x}\,|\,\theta,\mathscr{M}) \;=\; \prod_{i=1,\ldots,N} p(x_i\,|\,\theta,\mathscr{M})$$

The model-level likelihood $p(\vec{x}\,|\,\mathscr{M})$ plays a special role in Bayesian work, and is called the **Evidence**:

$$p(\vec{x}\,|\,\mathscr{M}) \;=\; \int_\theta p(\theta\,|\,\mathscr{M})\,p(\vec{x}\,|\,\theta,\mathscr{M})\,\mathrm{d}\,\theta$$

Note it requires a parameter prior.

Independence: Advanced

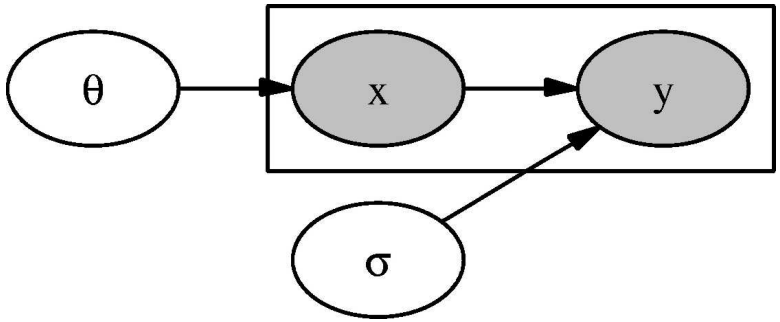
Likelihood of sample of size N as **(1)** a vector of data $\vec{x}$, given model $\mathscr{M}_x$ and parameters θ (possibly a vector), and **(2)** a vector of data $\vec{y}|\vec{x}$, given model $\mathscr{M}_y$ and parameters ϕ and **(3)** assuming *i.i.d.* data:

$$p(\vec{x},\vec{y}|\,\theta,\phi,\mathscr{M}_x,\mathscr{M}_y) = \prod_{i=1,\ldots,N} p(x_i|\,\theta,\mathscr{M}_x) \cdot \prod_{i=1,\ldots,N} p(y_i|\,x_i,\phi,\mathscr{M}_y)$$

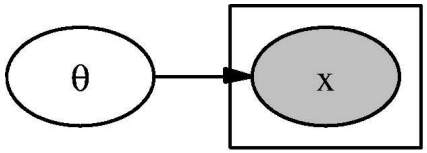
If we further assume parameters for $\mathscr{M}_x$ and $\mathscr{M}_y$ are *aprior independent* then

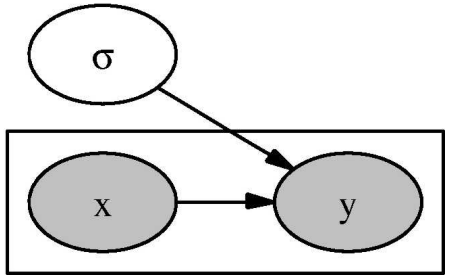
$$\begin{aligned} p(\theta,\phi\,|\vec{x},\vec{y},\mathscr{M}_x,\mathscr{M}_y) &= p(\theta\,|\vec{x},\mathscr{M}_x)p(\phi\,|\vec{x},\vec{y},\mathscr{M}_y) \\ p(\vec{x},\vec{y}|\,\mathscr{M}_x,\mathscr{M}_y) &= p(\vec{x}|\,\mathscr{M}_x)p(\vec{y}|\,\vec{x},\mathscr{M}_y) \end{aligned}$$

Independence: Advanced, cont.



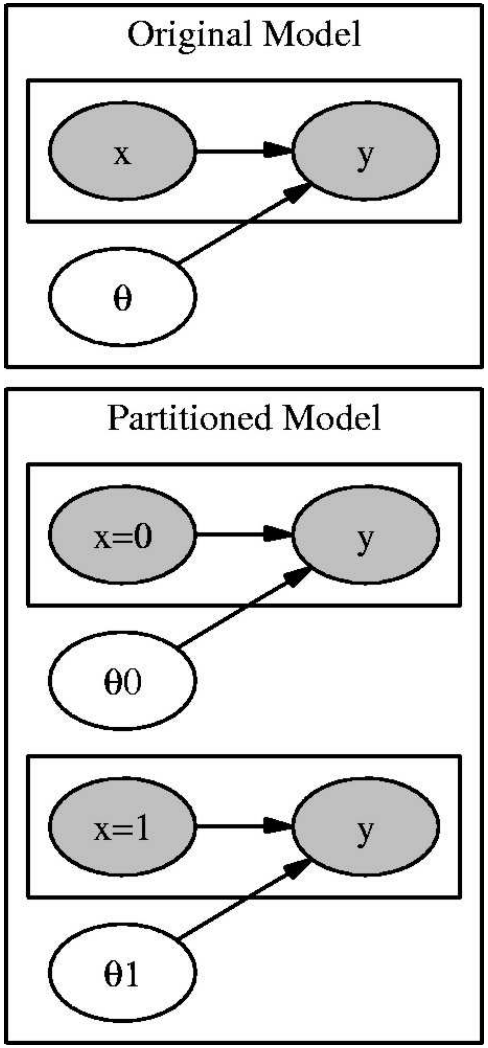
Thus we can treat the independent model case at two separate models.





Aposterior independence of parameters comes from likelihood independence and aprior independence.

Independence: Advanced, cont.



Parameter independence also occurs within a single conditional distribution. Here we have $p(y|x,\theta)$. We have broken up the input space (the domain of x) into two parts $x=0$ and $x=1$ and are building two separate models for each part.

$$p(\theta|\vec{x},\vec{y},\mathscr{M}) \;=\; p(\theta_0|\vec{x}=\vec{0},\vec{y},\mathscr{M})p(\theta_1|\vec{x}=\vec{1},\vec{y},\mathscr{M})$$

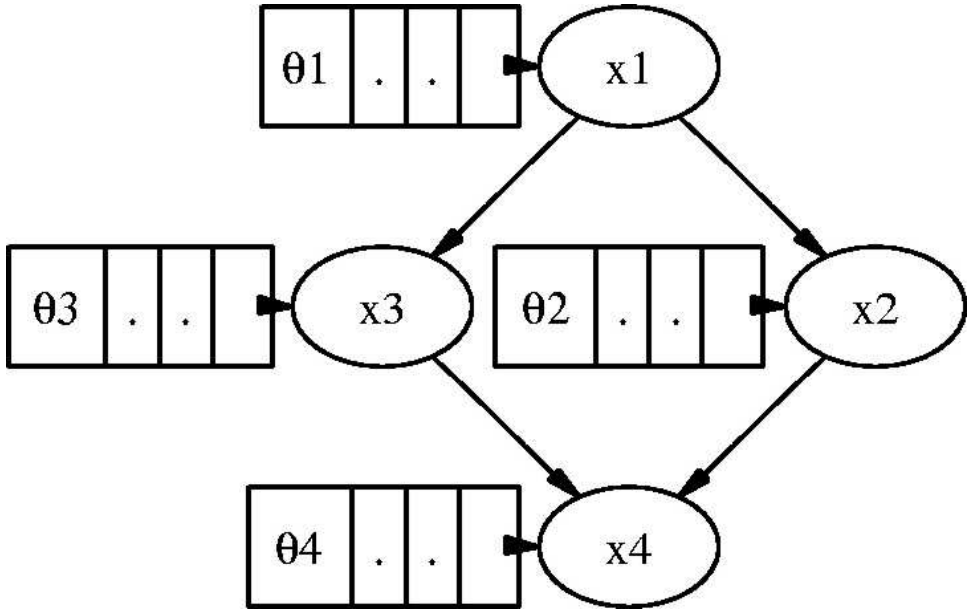
This is known as a **partitioned model**. A *decision tree* (e.g., using Quinlans C4.5) build's such a space. Conditional probability tables in a Bayesian network also form such a space: they are just a set of probability vectors for the partition of the space. Likewise *n-grams* used as language models in speech recognition form a partitioned conditional space.

Independence: Example

Given data $\vec{x}$ sampled from domain X, suppose we have selected a particular Bayesian network as our best model, $\mathscr{M}$. and its various parameters θ_y, one set for the table at each node $y \in X$, and we have aposterior independence for these.

How do we do inference on new variables without assigning to unknown parameters θ_y? *e.g.* whats

$$\begin{aligned} & p(x_1, x_4 \,|\, \vec{x}, \mathscr{M}) \\ = \; & \text{Expec}_{p(\theta \,|\, \vec{x}, \mathscr{M})} \left(p(x_1, x_4 \,|\, \theta) \right) \end{aligned}$$



Independence: Example

This $p(x_1, x_4 | \vec{x}, \mathscr{M})$ can be found as

$$
\begin{aligned}
&= \mathsf{Expec}_{p(\theta|\vec{x},\mathscr{M})}\left(p(x_1, x_4|\theta)\right) \\
&= \int_\theta \left(\sum_{x_2,x_3} p(x_1|\theta_1) p(x_2|x_1,\theta_2) p(x_3|x_1,\theta_3) p(x_4|x_2,x_3,\theta_4) \right) p(\theta_1,\theta_2,\theta_3,\theta_4|\vec{x},\mathscr{M})\, \mathrm{d}\theta \\
&= \sum_{x_2,x_3} \mathsf{Expec}_{p(\theta_1|\vec{x},\mathscr{M})} p(x_1|\theta_1) \mathsf{Expec}_{p(\theta_2|\vec{x},\mathscr{M})} p(x_2|x_1,\theta_2) \\
&\qquad \mathsf{Expec}_{p(\theta_3|\vec{x},\mathscr{M})} p(x_3|x_1,\theta_3) \mathsf{Expec}_{p(\theta_4|\vec{x},\mathscr{M})} p(x_4|x_2,x_3,\theta_4)
\end{aligned}
$$

Note probabilities such as $p(x_4|x_2,x_3,\theta_4)$ just pull a value from the table of entries for θ_4. Thus if $\overline{\theta}_k = \mathsf{Expec}_{p(\theta_k|\vec{x},\mathscr{M})}(\theta_k)$, then

$$
= \sum_{x_2,x_3} p(x_1|\overline{\theta}_1) p(x_2|x_1,\overline{\theta}_2) p(x_3|x_1,\overline{\theta}_3) p(x_4|x_2,x_3,\overline{\theta}_4)
$$

By parameter independence, the expected value of the inference is the same as the inference on the table of expected values.

Independence: Inference Theory

Suppose you want to evaluate $p(U \,|\, \vec{x}, \mathscr{M})$, for $U \subseteq X$. Now, X is a discrete domain, and the parameters θ_y represent the table of values at each node $y \in X$.

$$p(U \,|\, \vec{x}, \mathscr{M}) \;=\; \text{Expec}_{p(\theta \,|\, \vec{x}, \mathscr{M})} \left(\sum_{X/U} \prod_{y \in X} p(y \,|\, \text{parents}(y), \theta_y) \right)$$

Since the entry in the product is just an element from the table θ_y for $y \in X$, and we have aposterior independence of parameters, it follows that

$$p(U \,|\, \vec{x}, \mathscr{M}) \;=\; p(U \,|\, \overline{\theta})$$

where $\overline{\theta}_y = \text{Expec}_{p(\theta \,|\, \vec{x}, \mathscr{M})}(\theta_y)$.

Overview

- The Search Space
- Independence
- **Evidence**
- The Exponential Family
- Putting it All Together

Terminology

For the simple case, a sample of size N as a vector of data $\vec{x}$, given model $\mathscr{M}$ and parameters θ (possibly a vector), assuming *i.i.d.* data:

Likelihood: $p(\vec{x}\,|\,\theta,\mathscr{M})$.
Prior: $p(\theta\,|\,\mathscr{M})$.
Posterior: $p(\theta\,|\,\vec{x},\mathscr{M})$.
Posterior Utility: $\text{Expec}_{p(\theta\,|\,\vec{x},\mathscr{M})}(\text{utility}(\theta))$.
Evidence: $p(\vec{x}\,|\,\mathscr{M})$.

Terminology, cont.

Likelihood: critical assumption, sometimes chosen for computational convenience!

Prior: an unavoidable part of the model and should encode as much information as possible.

Posterior: the logical consequence of the previous two, often computationally unwieldy.

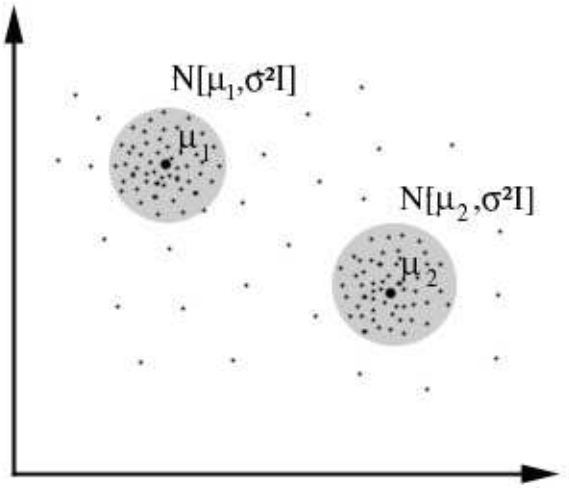
Posterior Utility: the optimization objective; its inclusion can make the posterior manageable.

Evidence: useful tool for evaluating the model and comparing it with others.

- its use sometimes ignores model priors, i.e., $p(\mathcal{M})$;
- is used for comparing, evaluating and combining multiple models;
- is defined via integration and is often only approximable.

Evidence: Example

Our data $\vec{x}$ is a mixture of K Gaussians, for some unknown K. Here's the $K = 2$ case.



Let $\mathscr{M}_K$ be a model mixing exactly K Gaussians. Then:

- Which K do we use? Choosing an $\mathscr{M}_K$ based on data is called *model selection*.
- Why pick just one, how do we use several, $\mathscr{M}_2, \mathscr{M}_3$ and $\mathscr{M}_4$?
- Isn't $\mathscr{M}_K \subset \mathscr{M}_{K+1}$, so why use these, why not stick with $\mathscr{M}_\infty$?

Evidence. Example, cont.

We can play with the $p(\vec{x}|\mathscr{M}_K)$ as follows:

- They are comparable to a posterior:

$$p(\mathscr{M}_K|\vec{x}) \;\propto\; p(\vec{x}|\mathscr{M}_K)p(\mathscr{M}_K)\ .$$

  If priors are nearly constant, we can ignore them.
- They standardize by introducing a prior, and normalizing:

$$p(\mathscr{M}_K|\vec{x}) \;=\; \frac{p(\vec{x}|\mathscr{M}_K)p(\mathscr{M}_K)}{\sum_K p(\vec{x}|\mathscr{M}_K)p(\mathscr{M}_K)}$$

- But their computation can be very hard:

$$p(\vec{x}|\mathscr{M}_K) \;=\; \int_\theta p(\theta\,|\mathscr{M}_K)\,p(\vec{x}|\,\theta,\mathscr{M}_K)\mathrm{d}\,\theta$$

Evidence. Example, cont.

- The quantities $p(\vec{x}|\mathscr{M}_K)$ for $K = 1, \ldots,$ let us compare the worth of different K (assuming we can computer or approximate them).
- If one K has dominant $p(\vec{x}|\mathscr{M}_K)$, and things seem to vanish for larger values, then select that K.
- Strictly speaking, $\mathscr{M}_K$ is a set of measure zero in $\mathscr{M}_{K+1}$, and for most computationally tractible priors, such sets don't register. So we treat separately.
- Use $\mathscr{M}_\infty$ only if you can sort out the priors, math. and algorithms.

Evidence. Example, cont.

If $p(\vec{x}|\mathscr{M}_K), p(\vec{x}|\mathscr{M}_{K+1}), p(\vec{x}|\mathscr{M}_{K+2})$ are all quite big, then use all three and weight predictions with

$$\frac{p(\vec{x}|\mathscr{M}_k)}{\sum_{k=K,K+1,K+2} p(\vec{x}|\mathscr{M}_k)} \quad \text{for } k = K, K+1, K+2$$

This is called *model averaging*. It represents the realistic scenario that you are unsure which model is best, so hedge your bets and pool their responses.

Model Averaging

- Averaging over different θ the posterior parameter values, is expectation or finding the posterior mean. Done with integration, Monte Carlo Markov Chain (MCMC), etc.
- But when we average over different models $\mathscr{M}$ it is called model averaging. Done similarly.
- Early examples: for Decision trees, Buntine 1991, for Bayesian networks, York and Madigan, 1991.
- Can be done with exponentially many models in some cases for decision trees, n-grams for language modelling, Bayesian networks: e.g., Periera and Singer 1997. Produces near state of the art in each case.

Overview

- The Search Space
- Independence
- Evidence
- **The Exponential Family**
- Putting it All Together

Multinomial

$j \sim C$-dim multinomial$(\theta_1, \ldots, \theta_C)$

Functional form	θ_j for $j \in [1, \ldots, C]$ for $\sum_{i=1,\ldots,C} \theta_i = 1$
Conjugate prior	$\theta \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_C)$
Conjugate	$\theta \sim$ Dirichlet$(n_1 + \alpha_1, \ldots, n_C + \alpha_C)$
....posterior	for $n_c = \sum_{i=1}^{N} 1_{j_i=c} = \#< j's = c >$
Evidence	Beta$(n_1 + \alpha_1, \ldots, n_C + \alpha_C)$ / Beta$(\alpha_1, \ldots, \alpha_C)$

$\theta \sim C$-dim Dirichlet$(\alpha_1, \ldots, \alpha_C)$ means

$$p(\theta \,|\, \alpha) = \frac{1}{\text{Beta}(\alpha_1, \ldots, \alpha_C)} \prod_{i=1}^{C} \theta_i^{\alpha_i - 1} \quad \text{for} \quad \alpha_i > 0$$

$$\text{Beta}(\alpha_1, \ldots, \alpha_C) = \prod_{i=1}^{C} \Gamma(\alpha_i) \Big/ \Gamma\left(\sum_{i=1}^{C} \alpha_i\right)$$

Gamma

$x \sim \text{Gamma}(\alpha > 0, \beta > 0)$ means for $x \in \Re^+$,

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Functional form	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x \in \Re^+$
Conjugate prior	$\beta\|\alpha \sim \text{Gamma}(\alpha_0, \beta_0)$
Conjugate ....posterior	$\beta\|\alpha \sim \text{Gamma}(N\alpha + \alpha_0, \sum_{i=1}^N x_i + \beta_0)$
Evidence	$\frac{\beta_0^{\alpha_0}\Gamma(N\alpha+\alpha_0)}{\Gamma(\alpha_0)\left(\sum_{i=1}^N x_i+\beta_0\right)^{N\alpha+\alpha_0}}$ for α fixed

Gaussian

$y|x \sim \text{Gaussian}(x^\dagger\theta, \sigma^2)$

Functional form	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma}(y-x^\dagger\theta)^2\right)$ for $y\in\Re, x\in\Re^d$
Conjugate prior	$\theta\|\sigma \sim d\text{-dim Gaussian}(\theta_0, \frac{1}{\sigma^2}\Sigma_0)$, $\sigma^{-2} \sim \text{Gamma}(\alpha_0/2, \beta_0/2)$
Conjugate ....posterior	$\theta\|\sigma \sim d\text{-dim Gaussian}(\overline{\theta}, \frac{1}{\sigma^2}\Sigma)$, $\sigma^{-2} \sim \text{Gamma}((\alpha_0+N)2, 2/\beta)$, for $\Sigma = \Sigma_0 + \sum_{i=1}^N x_i x_i^\dagger$, $\overline{\theta} = \Sigma^{-1}\left(\Sigma_0\theta_0 + \sum_{i=1}^N y_i x_i\right)$, for $\beta = \sum_{i=1}^N (y_i - \overline{\theta}^\dagger x_i)^2 + (\overline{\theta}-\theta_0)^\dagger\Sigma_0(\overline{\theta}-\theta_0) + \beta_0$
Evidence	$\frac{\det^{1/2}\Sigma_0}{\pi^{N/2}\det^{1/2}\Sigma}\frac{\Gamma((\alpha_0+N)/2)\beta^{(\alpha_0+N)/2}}{\Gamma(\alpha_0/2)\beta_0^{\alpha_0/2}}$

Multi-dimensional Gaussian

$x \sim d$-dim Gaussian(μ, Σ) gives probability

$$\frac{\det^{1/2}\Sigma}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x-\mu)^{\dagger}\Sigma(x-\mu)\right) \text{ for } x \in \Re^d$$

$S \sim d$-dim Wishart(α, Σ) gives probability

$$\frac{\det^{-\alpha/2}\Sigma \det^{\alpha-d-1/2} S}{2^{d\alpha/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((\alpha+1-i)/2)} \exp\left(-\frac{1}{2}\text{trace}\Sigma^{-1}S\right)$$

for S, Σ symmetric positive definite matrices in $\Re$ of dimension d, μ a vector in $\Re$ of dimension d, $\alpha \in \Re$ such that $\alpha \geq d$.

Multi-dimensional Gaussian, cont.

Conjugate prior	$\mu\|\Sigma \sim \text{Gaussian}(\mu_0, N_0\Sigma)$, $\Sigma \sim \text{Wishart}(\delta_0, S_0)$
Conjugate ....posterior	$\mu\|\Sigma \sim \text{Gaussian}(\overline{\mu}, (N+N_0)\Sigma)$, $\Sigma \sim \text{Wishart}(N+\delta_0, S+S_0)$ for $\overline{\mu} = \overline{x} + \frac{N_0}{N+N_0}(\mu_0 - \overline{x})$, for $S = \sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^\dagger$
Evidence	$\frac{\det^{\delta_0/2} S_0}{(\pi)^{dN/2}\det^{(\delta_0+N)/2}(S+S_0)} \frac{N_0^d}{(N+N_0)^d} \prod_{i=1}^{d} \frac{\Gamma((\delta_0+N-1-i)/2)}{\Gamma((\delta_0-1-i)/2)}$

The Exponential Family

A vector of measurements $\vec{x}$, a vector of T functions $\vec{t}(\vec{x})$ and some parameters $\vec{\theta}$ also of dimension T:

$$q(\vec{x}|\,\vec{\theta}) \;=\; \frac{1}{Y_t(\vec{x})Z_t(\vec{\theta})}\exp\left(\vec{t}(\vec{x})^\dagger\vec{\theta}\right) \;\;.$$

where $\vec{\mu}_t \;\equiv\; \text{Expec}_{q(\vec{x}|\,\vec{\theta})}(\vec{t}(\vec{x}))$

Z_t	$t_k(\vec{x})$	θ_k	$\mu_{t,k}$
$\sqrt{2\pi}\sigma\exp(\mu^2/2\sigma^2)$	$x,\ x^2$	$\frac{\mu}{\sigma^2},\ -\frac{1}{2\sigma^2},$	$\mu,\ \sigma^2+\mu^2$
1	x_k	$\log\alpha_k$	$N\alpha_k$
$\frac{\prod_k\Gamma(\alpha_k)}{\Gamma(\sum_k\alpha_k)}$	$\log x_k$	α_k-1	$\Psi_0(\alpha_k)-\Psi_0\left(\sum_k\alpha_k\right)$

The Exponential Family, cont.

Two key definitions:

$$\vec{\mu}_t \equiv E_{q(\vec{x}|\vec{\theta})}\{\vec{t}(\vec{x})\} = \frac{\partial \log Z_t}{\partial \vec{\theta}} ,$$

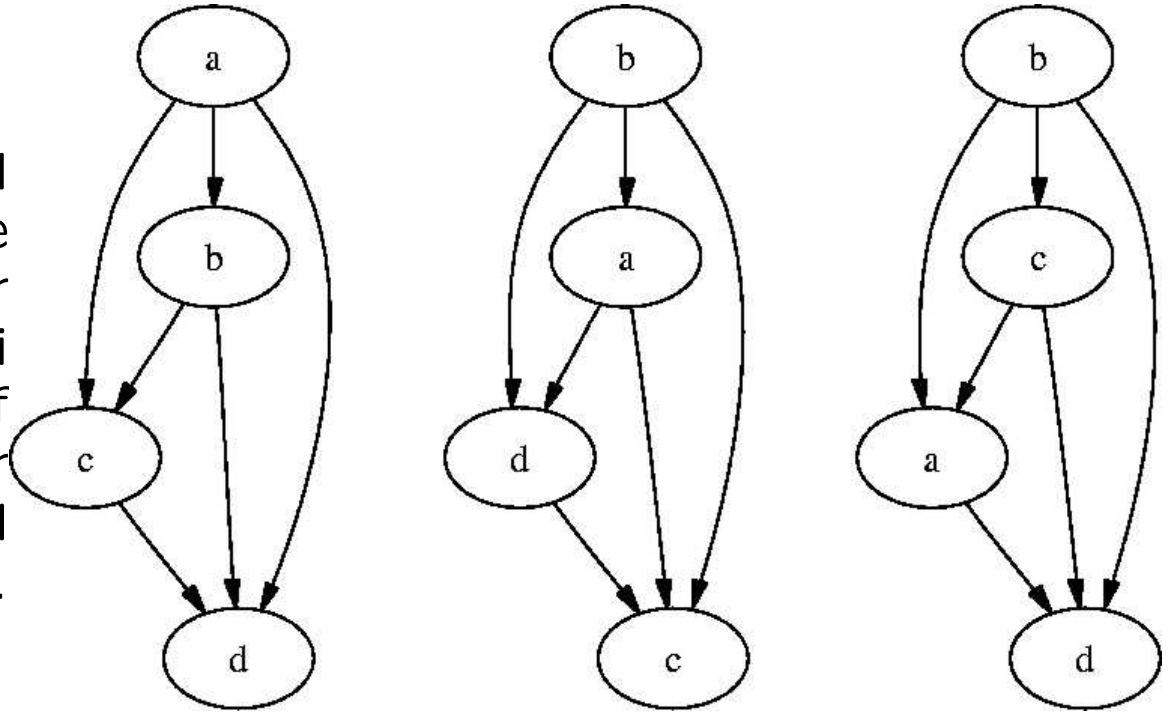
$$\vec{\Sigma}_t \equiv E_{q(\vec{x}|\vec{\theta})}\{(\vec{t}(\vec{x})-\vec{\mu}_t)(\vec{t}(\vec{x})-\vec{\mu}_t)^\dagger\} = \frac{\partial^2 \log Z_t}{\partial \vec{\theta}\partial \vec{\theta}} = \frac{\partial \vec{\mu}_t}{\partial \vec{\theta}} .$$

Exponential family are always unimodel, and moments of $\vec{t}(\vec{x})$ and $\exp(\vec{t}(\vec{x}))$ found by manipulating Z_t. Also,

$$I(q(\vec{x}|\vec{\theta})) = E_{q(\vec{x}|\vec{\theta})}\{\log Y_t(\vec{x})\} + \log Z_t - \vec{\mu}_t^\dagger \vec{\theta}$$

Evidence and Multivariate Models

These graphs all represent the same assumptions. For clean aposteriori independence of the parameters for $\theta_a, \theta_b, \theta_c, \theta_d$, we need apriori independence.



Suppose we have finite discrete variables a, b, c, d. If we want aposteriori independence to happen for *any order* of the graph, then the priors must be Dirichlet. If the variables a, b, c, d are Gaussian, then the priors must be Gaussian-Wishart. (Heckerman and Geiger 1995-1997).

Overview

- The Search Space
- Independence
- Evidence
- The Exponential Family
- **Putting it All Together**

One Algorithm

One algorithm for learning Bayesian networks. Works with all Discrete or all Gaussian variables.

1. Pick and ordering of the variables X.
2. For each variable $x \in X$, search the space of conditional models for it using any parent set from those before it in the order.
3. Evaluate each such conditional model using evidence. Save the best each variable $x \in X$ and record their evidence and values.
4. Pool the individual conditional model to form the full joint model.

One Algorithm, cont

The evidence for the discrete case is given as follows:

Suppose we have the conditional model of x conditioned on parents(x). Let x be a C-valued discrete. If this conditional model is a fully partitioned model with a separate probability vector for each of the values $y \in$ Domain(parents(x)) represented as θ_y. The prior on θ_y is a C-dimensional Dirichlet with C-dimensional parameter vector $\vec{\alpha}_x$ independent of y.

Let the count of data with $x = c$ when parents$(x) = y$ be $n_{c,y}$. The C-dimensional vector of these for inputs y is denoted $\vec{n}_{\cdot,y}$.

One Algorithm, cont

Then the evidence for this conditional model for x is:

$$\prod_{y \in \text{Domain}(\text{parents}(x))} \frac{\text{Beta}(\vec{\alpha}_x + \vec{n_{\cdot,y}})}{\text{Beta}(\vec{\alpha}_x)}$$

You might initialize $\vec{\alpha}_x$ as follows. Let C-dimensional vector $\vec{m}_x$ be the observed probabilities of the C-valued variable x. Let $\vec{\alpha}_x = 2 * \vec{m}_x$. This is called an empirical prior.

We are working on a perfectly partitioned space. We have a separate probability vector for each variable x, and for each assignment to its parents$(x) = y$. Each such vector is evaluated on the merits of the sample it sees according to the standard evidence formula for the multinomial distribution.

Next Week

- Do your project and term paper!