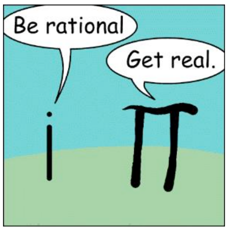
BTRY 6790, Probabilistic Graphical Models



Sep. 5, 2013

Plan for Today

- Finish statistics (quick!)
- Directed graphical models
- Factorization of joint distributions
- Conditional independence
- Terminology and notation

MLE Example #3

- Suppose we have a DNA sequence of length *n*

$$\mathbf{x} = \text{CGATCTAG...} = (x_1, x_2, \ldots, x_n)$$

- Assume bases are iid from a multinomial distribution

$$f(x_i) = \begin{cases} \pi_A & x_i = A \\ \pi_C & x_i = C \\ \pi_G & x_i = G \\ \pi_T & x_i = T \end{cases}$$

- We wish to estimate the parameters of this distribution by maximum likelihood

The Likelihood

$$\begin{aligned} L(\boldsymbol{\pi}|\mathbf{x}) &= \prod_{i=1}^{n} \pi_A^{I(x_i=A)} \pi_C^{I(x_i=C)} \pi_G^{I(x_i=G)} \pi_T^{I(x_i=T)} \\ &= \pi_A^{\sum_{i=1}^n I(x_i=A)} \pi_C^{\sum_{i=1}^n I(x_i=C)} \pi_G^{\sum_{i=1}^n I(x_i=G)} \pi_T^{\sum_{i=1}^n I(x_i=T)} \\ &= \pi_A^{n_A} \pi_C^{n_C} \pi_G^{n_G} \pi_T^{n_T} \end{aligned}$$

$$\begin{aligned} \ln L(\boldsymbol{\pi}|\mathbf{x}) &= n_A \ln \pi_A + n_C \ln \pi_C + n_G \ln \pi_G + n_T \ln \pi_T \\ &= \sum_{b \in \mathcal{A}} n_b \ln \pi_b \qquad \text{where } \mathcal{A} = \{A, C, G, T\} \end{aligned}$$

Solving for the MLEs

- Define *Lagrangian*

$$\ln L(\boldsymbol{\pi}|\mathbf{x}) = \sum_{b \in \mathcal{A}} n_b \ln \pi_b$$

$$\tilde{l}(\boldsymbol{\pi}|\mathbf{x}) = \sum_{b \in \mathcal{A}} n_b \ln \pi_b + \lambda \left(1 - \sum_{b \in \mathcal{A}} \pi_b\right)$$

$$\frac{\partial}{\partial \pi_b} \tilde{l}(\boldsymbol{\pi}|\mathbf{x}) = \frac{n_b}{\pi_b} - \lambda = 0$$

- Solve for "dummy" variable

$$n_b = \lambda \pi_b$$

$$\sum_{b \in \mathcal{A}} n_b = \sum_{b \in \mathcal{A}} \lambda \pi_b$$

$$n = \lambda$$

- The MLEs are the relative frequencies

$$\Longrightarrow \quad \pi_A = \frac{n_A}{n}, \quad \pi_C = \frac{n_C}{n}, \quad \pi_G = \frac{n_G}{n}, \quad \pi_T = \frac{n_T}{n}$$

ML Estimation for Complex Models

- Theta may have very high dimension (tens, hundreds, even thousands of parameters)
- Even if the (negative) likelihood function is *convex,* it may not be possible to solve for the MLE analytically
- Often multiple local maxima
- Numerical optimization methods are used: gradient descent, Newton's method, quasi-Newton methods, conjugate gradients
- Stochastic methods can also be used

Bayesian Inference

- Bayes' formula:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

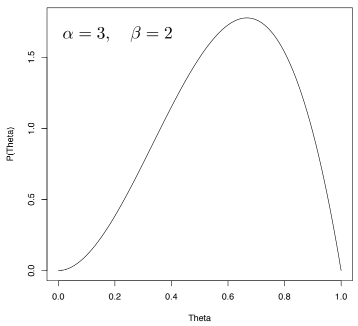
- Combination of likelihood and *prior*
- Parameters are treated like random variables
- Idea is to infer *posterior* distributions for parameters, given the data

Bayesian Coin Flipping

- Suppose coin with weight θ. Huckster at fair is taking bets on outcomes. What is θ?
- You have a weak prior belief that the coin is not fair ($\theta > 0.5$)
- Prior distribution: Beta(α=3, β=2). Reason: mathematical convenience

$$p(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad 1 \leq \theta \leq 0$$

Beta Prior



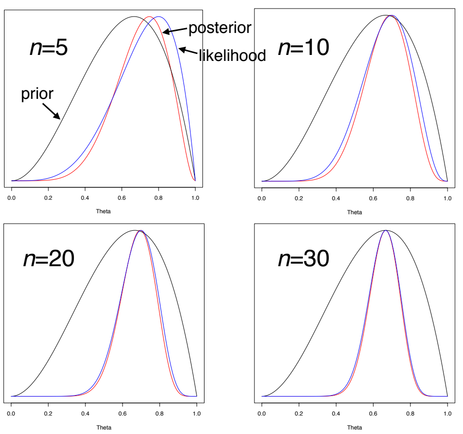
Solving for the Posterior

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta} \propto p(\mathbf{x}|\theta)p(\theta)$$

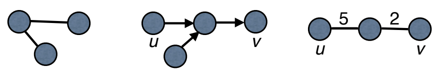
$$\propto \theta^s(1-\theta)^{n-s}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}$$

$$= \mathrm{Beta}(s+\alpha, n-s+\beta)$$



First: Graphs

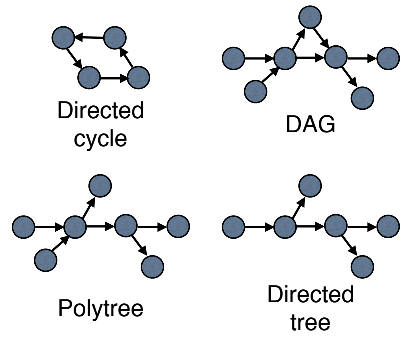


- A graph consists of ***nodes*** and ***edges***. The edges may be ***directed*** or ***undirected***, and may be ***weighted*** or ***unweighted***.
- A ***path*** from node *u* to node *v* is a sequence of connected edges leading from *u* to *v*
- The ***length*** of a path is its total number of edges. The ***weight*** of a path is the sum of the weights of all edges.
- A ***cycle*** is a path (of nonzero length) from a node to itself. An undirected graph without cycles is called a ***tree***.

Directed Acyclic Graphs (DAGs)

- A ***DAG*** is a directed graph that does not contain (directed) cycles
- A ***directed tree*** is a DAG in which every node has at most one parent
- A ***polytree*** is a DAG whose underlying undirected graph is a tree

Examples



Directed Graphical Models

(Bayesian Networks)

- Let $X = \{X_1, ..., X_n\}$ be a set of (discrete) ***random variables*** of interest.
- Let $G = (V, E)$ be a directed acyclic graph. Nodes in G correspond one-to-one with variables in X.
- Let X_v be the variable associated with $v \in V$, let X_U be associated with $U \subseteq V$
- The graph defines the ***joint distribution***, $p(X_1, ..., X_n)$. From this we can obtain various ***marginal*** or ***conditional*** distributions of interest

Marginals and Conditionals

- By the ***law of total probability*** a marginal probability $p(x_U) = p(X_U = x_U)$ is given by,

$$p(x_U) = \sum_{x_T : T = V - U} p(x_U, x_T)$$

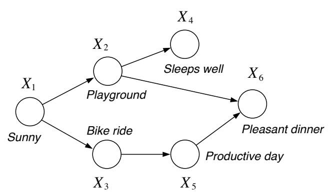
- By the definition of conditional probability,

$$p(x_U | x_W) = \frac{p(x_{U \cup W})}{p(x_W)}$$

where: $$p(x_{U \cup W}) = \sum_{x_S : S = V - (U \cup W)} p(x_{U \cup W}, x_S)$$

$$p(x_W) = \sum_{x_{S'} : S' = V - W} p(x_W, x_{S'})$$

Example

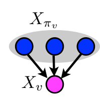


$$p(x_1, x_2 | x_3, x_4) = \frac{\sum_{x_5, x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)}{\sum_{x_1, x_2, x_5, x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)}$$

May be expensive!

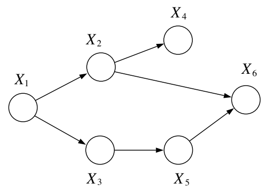
Local Conditionals

- Let π_v be the set of *parents* of v. The corresponding set of variables is $X_{\pi v}$.
- Let $p(x_v \mid x_{\pi v})$ be the *local conditional* distribution of v given π_v
- The local conditional distributions together define a joint distribution:



$$p(x_1, \ldots, x_n) = \prod_v p(x_v | x_{\pi_v})$$

Factorization Example



$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

Theorem

- Suppose associated with every node v and its parents π_v is an arbitrary function, $f_v(x_v, x_{\pi v})$, such that:

$$f_v(x_v, x_{\pi_v}) \geq 0 \quad \forall x_v, \qquad \sum_{x_v} f_v(x_v, x_{\pi_v}) = 1$$

- Let:

$$f(x_1, \ldots, x_n) = \prod_v f_v(x_v, x_{\pi_v})$$

- Then it must be true that:

$$f(x_1, \ldots, x_n) \geq 0 \quad \forall x_1, \ldots, x_n$$
$$\sum_{x_1, \ldots, x_n} f(x_1, \ldots, x_n) = 1$$

Theorem, cont.

- Furthermore, the joint distribution

$$p(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)$$

has marginals:

$$p(x_v|x_{\pi_v}) = f_v(x_v, x_{\pi_v})$$

Sketch of Proof

- Nonnegativity follows from nonnegativity of the f_vs
- The sum of one can be seen by listing the variables in topological order, sliding summations to the right, and replacing sums with 1s from right to left, e.g.,

$$\sum_{x_1, \ldots, x_n} f(x_1, \ldots, x_n) = 1$$
$$\sum_{x_1} \cdots \sum_{x_n} f_1(x_1, x_{\pi_1}) \cdots f_n(x_n, x_{\pi_n}) = 1$$
$$\sum_{x_1} f_1(x_1, x_{\pi_1}) \cdots \sum_{x_n} f_n(x_n, x_{\pi_n}) = 1$$
$$1 = 1$$

Sketch of Proof, cont.

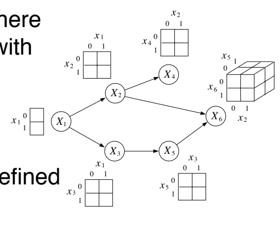
- To show that the marginals have to be the f_vs, start with the root nodes, e.g.,

$$p(x_1|.) = \sum_{x_2, \ldots, x_n} f(x_1, \ldots, x_n)$$
$$= f_1(x_1, .) \sum_{x_2} f_2(x_2, x_{\pi_2}) \cdots \sum_{x_n} f_n(x_n, x_{\pi_n})$$
$$= f_1(x_1, .) \cdot 1 \cdots 1$$
$$= f_1(x_1, .)$$

- The proofs for the downstream nodes proceed in a similar way, by induction.

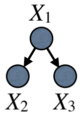
Tables

- The graph defines a *family* of joint distributions, all of which factor in the same way
- Each member has an economic representation in terms of its local conditional distributions
- If discrete and finite, there is a *table* associated with each edge of G
- Now exponential in $|\pi_v|$ rather than in $|V|$
- Degree of reduction defined by factorization



Conditional Independence

- The graph *G* also represent a set of *conditional independence* statements
- We say X_2 and X_3 are conditionally independent given X_1 if



$$p(x_2, x_3|x_1) = p(x_2|x_1)p(x_3|x_1)$$

or

$$p(x_2|x_1, x_3) = p(x_2|x_1)$$

for all x_1, x_2, and x_3 such that $p(x_1) > 0$

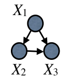
- Thus, by assuming: $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$

instead of: $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$

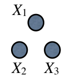
we assume CI of x_2 and x_3 given x_1

Examples

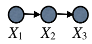
- No conditional independence assertions = fully connected graph



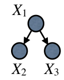
- Complete independence = fully unconnected graph



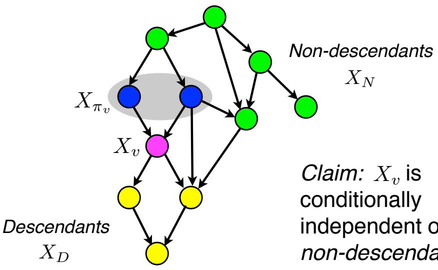
- First-order Markov dependencies = linear chain



- Branching Markov dependencies = directed tree

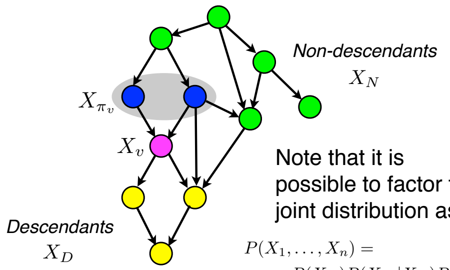


Graph Separation & CI



Claim: X_v is conditionally independent of its *non-descendants* X_N given its *parents* X_{π_v}

Factorization



Note that it is possible to factor the joint distribution as:

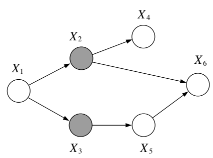
$$P(X_1, \ldots, X_n) = P(X_N)P(X_{\pi_v}|X_N)P(X_v|X_{\pi_v}) \times P(X_D|X_v, X_{\pi_v}, X_N)$$

Theorem

- Claim: $P(X_v|X_{\pi_v}, X_N) = P(X_v|X_{\pi_v})$
- Proof:

$$P(X_v|X_{\pi_v}, X_N) = \frac{\sum_{X_D} P(X_N)P(X_{\pi_v}|X_N)P(X_v|X_{\pi_v})P(X_D|X_v, X_{\pi_v}, X_N)}{\sum_{X_D}\sum_{X_v} P(X_N)P(X_{\pi_v}|X_N)P(X_v|X_{\pi_v})P(X_D|X_v, X_{\pi_v}, X_N)}$$

Blocking of Dependency



$$\begin{aligned}
p(x_6|x_1, x_2, x_3) &= \frac{\sum_{x_4, x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)}{\sum_{x_4, x_5, x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)\sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)\sum_{x_4} p(x_4|x_2)}{p(x_1)p(x_2|x_1)p(x_3|x_1)\sum_{x_4} p(x_4|x_2)\sum_{x_5} p(x_5|x_3)\sum_{x_6} p(x_6|x_2, x_5)} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)\sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)}{p(x_1)p(x_2|x_1)p(x_3|x_1)} \\
&= \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5) \\
&= p(x_6|x_2, x_3) \qquad \Longrightarrow \; X_1 \perp\!\!\!\perp X_6 \,|\, X_2, X_3
\end{aligned}$$

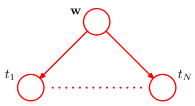
Next Time

- More general blocking of dependency (Bayes ball algorithm and D-separation)
- Relationship between a particular factorization and a particular set of conditional independence assumptions

Continuous vs. Discrete Models

- So far, emphasis on discrete random variables, but most points hold with continuous variables
- In particular, factorization, conditional independence, and blocking are unchanged
- Proofs remain the same but with summations replaced by integrals
- Algorithms for inference do change

A Word About Notation



Series Notation

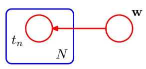
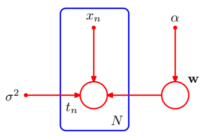
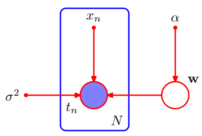


Plate Notation



Parameters



Observed Variables

That's All

- The class is now full
- Everyone should be signed up for Piazza: https://piazza.com/cornell/fall2013/btry6790cs6782/home
- The time for the discussion section is set at Wed 3:30-4:30, but the room will change
- Keep up with readings!
  - Bishop chapter 8 (8.0–8.3), Jordan chapter 2
  - Jordan & Weiss, Kevin Murphy reviews
- First assignment posted tomorrow or Sat