Lecture 4:

Probability Models

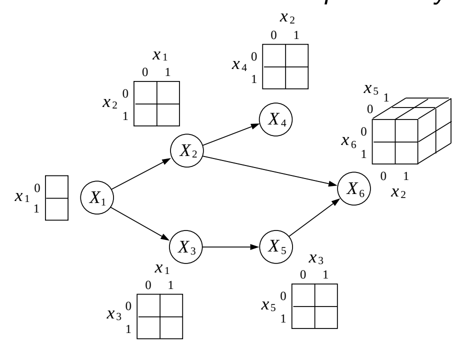
January 18, 2006

What's Inside the Nodes/Cliques?

- We've focused a lot on the structure of the graphs in directed and undirected models. Today we'll look at specific functions that can live inside the nodes (directed) or on the cliques (undirected).
- For directed models we need prior functions $p(\mathbf{x}_i)$ for root nodes and parent-conditionals $p(\mathbf{x}_i|\mathbf{x}_{\pi_i})$ for interior nodes.
- For undirected models we need clique potentials $\psi_C(\mathbf{x}_C)$ on the maximal cliques (or log potentials/energies $H_C(\mathbf{x}_C)$).
- We'll consider various types of nodes: binary/discrete (categorical), continuous, interval, and integer counts.
- We'll see some basic *probability models* (parametrized families of distributions); these models live inside nodes of directed models.
- We'll also see a variety of potential/energy functions which take multiple node values as arguments and return a scalar compatibility; these live on the cliques of undirected models.

Probability Tables & CPTs

- For discrete (categorical) variables, the most basic parametrization is the probability table which lists $p(x = k^{th} \text{ value})$.
- Since PTs must be nonnegative and sum to 1, for k-ary nodes there are $k-1$ free parameters.
- If a discrete node has discrete parent(s) we make one table for each setting of the parents: this is a *conditional probability table* or CPT.



Exponential Family

- For a numeric random variable $\mathbf{x}$

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\}$$
$$= \frac{1}{Z(\eta)} h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x})\}$$

is an exponential family distribution with *natural parameter* η.
- Function $T(\mathbf{x})$ is a *sufficient statistic*.
- Function $A(\eta) = \log Z(\eta)$ is the log normalizer.
- Key idea: all you need to know about the data in order to estimate parameters is captured in the summarizing function $T(\mathbf{x})$.
- Examples: Bernoulli, binomial/geometric/negative-binomial, Poisson, gamma, multinomial, Gaussian, ...

Bernoulli Distribution

- For a binary random variable $x = \{0, 1\}$ with $p(x = 1) = \pi$:

$$p(x|\pi) = \pi^x (1-\pi)^{1-x}$$
$$= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}$$

- Exponential family with:

$$\eta = \log \frac{\pi}{1-\pi}$$
$$T(x) = x$$
$$A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$$
$$h(x) = 1$$

- The *logistic* function links natural parameter and chance of heads

$$\pi = \frac{1}{1+e^{-\eta}} = \text{logistic}(\eta)$$

Poisson

- For an integer count variable with *rate* λ:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$
$$= \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$

- Exponential family with:

$$\eta = \log \lambda$$
$$T(x) = x$$
$$A(\eta) = \lambda = e^\eta$$
$$h(x) = \frac{1}{x!}$$

- e.g. number of photons $\mathbf{x}$ that arrive at a pixel during a fixed interval given mean intensity λ
- Other count densities: (neg)binomial, geometric.

Multinomial

- For a categorical (discrete), random variable taking on K possible values, let π_k be the probability of the k^{th} value. We can use a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_k, \ldots, x_K)$ in which $x_k = 1$ if and only if the variable takes on its k^{th} value. Now we can write,

$$p(\mathbf{x}|\pi) = \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_K^{x_K} = \exp\left\{ \sum_i x_i \log \pi_i \right\}$$

Exactly like a probability table, but written using binary vectors.

- If we observe this variable several times $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, the (iid) probability depends on the *total observed counts* of each value:

$$p(\mathbf{X}|\pi) = \prod_n p(\mathbf{x}^n|\pi) = \exp\left\{ \sum_i \left(\sum_n x_i^n \right) \log \pi_i \right\} = \exp\left\{ \sum_i c_i \log \pi_i \right\}$$

Multinomial as Exponential Family

- The multinomial parameters are constrained: $\sum_i \pi_i = 1$.
Define (the last) one in terms of the rest: $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$

$$p(\mathbf{x}|\pi) = \exp\left\{ \sum_{i=1}^{K-1} \log\left(\frac{\pi_i}{\pi_K}\right) x_i + k \log \pi_K \right\}$$

- Exponential family with:

$$\eta_i = \log \pi_i - \log \pi_K$$
$$T(x_i) = x_i$$
$$A(\eta) = -k \log \pi_K = k \log \sum_i e^{\eta_i}$$
$$h(\mathbf{x}) = 1$$

- The *softmax* function relates direct and natural parameters:

$$\pi_i = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}$$

Gaussian (normal) 8

- For a continuous univariate random variable:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
$$= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}$$

- Exponential family with:

$$\eta = [\mu/\sigma^2\ ;\ -1/2\sigma^2]$$
$$T(x) = [x\ ;\ x^2]$$
$$A(\eta) = \log\sigma + \mu^2/2\sigma^2$$
$$h(x) = 1/\sqrt{2\pi}$$

- Note: a univariate Gaussian is a two-parameter distribution with a two-component vector of sufficient statistics. (also maxent)

Multivariate Gaussian Distribution 9

- For a continuous vector random variable:

$$p(x|\mu,\Sigma) = |2\pi\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

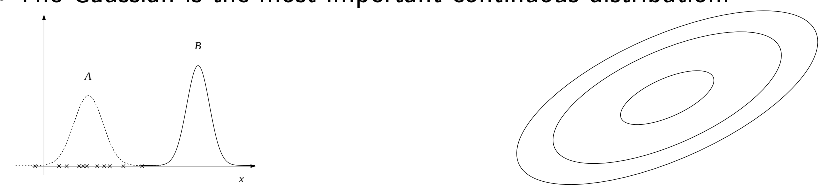
- Exponential family with:

$$\eta = [\Sigma^{-1}\mu\ ;\ -1/2\Sigma^{-1}]$$
$$T(x) = [\mathbf{x}\ ;\ \mathbf{x}\mathbf{x}^\top]$$
$$A(\eta) = \log|\Sigma|/2 + \mu^\top\Sigma^{-1}\mu/2$$
$$h(x) = (2\pi)^{-n/2}$$

- Note: a d-dimensional Gaussian is a d+d^2-parameter distribution with a d+d^2-component vector of sufficient statistics (but because of symmetry and positivity, parameters are constrained)

Gaussians 10

- The Gaussian is the most important continuous distribution.



- You should know how to manipulate these, and condition on subsets of variables given others. Mostly linear algebra.
- Other continuous densities: Student-t, Laplacian.
- Nonnegative densities: exponential, Gamma, log-normal.

Moments 11

- For numeric nodes, moment calculations are important.
- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer $A(\eta)$.
- The q^{th} derivative gives the q^{th} centred moment.

$$\frac{dA(\eta)}{d\eta} = \text{mean}$$
$$\frac{d^2A(\eta)}{d\eta^2} = \text{variance}$$
$$\cdots$$

- When the sufficient statistic is a vector, partial derivatives need to be considered.

Nodes with Parents — 12

- When the parent is discrete, we just have one probability model for each setting of the parent. Examples:
  – table of natural parameters (exponential model for cts. child)
  – table of tables (CPT model for discrete child)
- When the parent is numeric, some or all of the parameters for the child node become *functions* of the parent's value.
- A very common instance of this for regression is the "linear-Gaussian": $p(\mathbf{y}|\mathbf{x}) = \mathrm{gauss}(\theta^\top \mathbf{x}; \Sigma)$.
- For classification, often use Bernoulli/Multinomial densities whose parameters π are some function of the parent: $\pi_j = f_j(\mathbf{x})$.

GLMs and Canonical Links — 13

- Generalized Linear Models: $p(\mathbf{y}|\mathbf{x})$ is exponential family with conditional mean $\mu_i = f_i(\theta^\top \mathbf{x})$.
- The function f is called the *response function*.
- If we chose f to be the inverse of the mapping b/w conditional mean and natural parameters then it is called the *canonical response function* or *canonical link*:

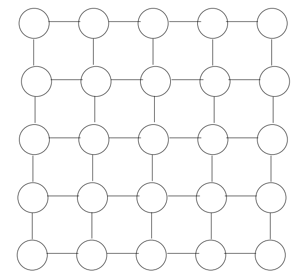
$$\eta = \psi(\mu)$$
$$f(\cdot) = \psi^{-1}(\cdot)$$

- Example: logistic function is canonical link for Bernoulli variables; softmax function is canonical link for multinomials

Potential Functions — 14

- We are much less constrained with potential functions, since they can be any positive function of the values of the clique nodes.
- Recall $\psi_C(\mathbf{x}_C) = \exp\{-H_C(\mathbf{x}_C)\}$
- A common (redundant) choice for cliques which are pairs is:

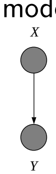
$$H(\mathbf{x}) = \sum_i a_i \mathbf{x}_i + \sum_{\text{pairs } ij} w_{ij} \mathbf{x}_i \mathbf{x}_j$$



Basic Statistical Problems — 15

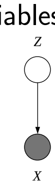
- Let's remind ourselves of the basic problems we discussed on the first day: *density estimation*, *clustering classification* and *regression*.
- Can always do joint density estimation and then condition:
  Regression: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y},\mathbf{x})/p(\mathbf{x}) = p(\mathbf{y},\mathbf{x})/\int p(\mathbf{y},\mathbf{x})d\mathbf{y}$
  Classification: $p(c|\mathbf{x}) = p(c,\mathbf{x})/p(\mathbf{x}) = p(c,\mathbf{x})/\sum_c p(c,\mathbf{x})$
  Clustering: $p(c|\mathbf{x}) = p(c,\mathbf{x})/p(\mathbf{x})$ c unobserved
  Density Estimation: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y},\mathbf{x})/p(\mathbf{x})$ $\mathbf{x}$ unobserved

In general, if certain nodes are *always* observed we may not want to model their density:



Regression/Classification

If certain nodes are *always* unobserved they are called *hidden* or *latent* variables (more later):



Clustering/Density Est.

- What can we do with a probabilistic graphical model?
- *Generate data.*
  For this you need to know how to sample from local models (directed) or how to do Gibbs or other sampling (undirected).
- *Compute log probabilities.*
  When all nodes are either observed or marginalized the result is a single number which is the log prob of the configuration.
- *Inference.*
  Compute expectations of some nodes given others which are observed or marginalized.
- *Learning.*
  Set the parameters of the local functions given some (partially) observed data to maximize the probability of seeing that data.