CSC412 – Probabilistic Learning & Reasoning — Sam Roweis

Lecture 5:

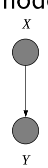
Parameter Estimation & Learning

January 23, 2006

Basic Statistical Problems 1

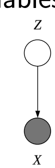
- Let's remind ourselves of the basic problems we discussed on the first day: *density estimation*, *clustering classification* and *regression*.
- Can always do joint density estimation and then condition:
  Regression: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y},\mathbf{x})/p(\mathbf{x}) = p(\mathbf{y},\mathbf{x})/\int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

Fundamental Operations 2

- 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.* (today)
  Set the parameters of the local functions given some (partially) observed data to maximize the probability of seeing that data.

Learning Graphical Models from Data 3

- In AI the bottleneck is often knowledge acquisition.
- Human experts are rare, expensive, unreliable, slow.
  But we have lots of machine readable data.
- Want to build systems automatically based on data and a small amount of prior information (e.g. from experts).

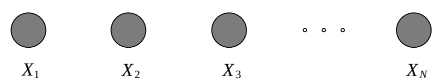

⇒ Sam Roweis


⇒ Geoff Hinton

- In this course, our "systems" will be probabilistic graphical models.
- Assume the prior information we have specifies type & structure of the GM, as well as the mathematical form of the parent-conditional distributions or clique potentials.
- In this case learning $\equiv$ setting parameters.
  ("Structure learning" is also possible but we won't consider it now.)

Multiple Observations, Complete IID Data

- A single observation of the data $\mathbf{X}$ is rarely useful on its own.
- Generally we have data including many observations, which creates a *set of random variables*: $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M\}$
- We will assume two things:
  1. Observations are independently and identically distributed according to joint distribution of graphical model: IID samples.
  2. We observe all random variables in the domain on each observation: complete data.
- We shade the nodes in a graphical model to indicate they are observed. (Later you will see unshaded nodes corresponding to missing data or latent variables.)



Likelihood Function

- So far we have focused on the (log) probability function $p(\mathbf{x}|\theta)$ which assigns a probability (density) to any joint configuration of variables $\mathbf{x}$ given fixed parameters θ.
- But in learning we turn this on its head: we have some fixed data and we want to find parameters.
- Think of $p(\mathbf{x}|\theta)$ as a function of θ for fixed $\mathbf{x}$:

$$L(\theta; \mathbf{x}) = p(\mathbf{x}|\theta)$$
$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta)$$

  This function is called the (log) "likelihood".
- Chose θ to maximize some cost function $c(\theta)$ which includes $\ell(\theta)$:

$$c(\theta) = \ell(\theta; \mathcal{D}) \qquad \text{maximum likelihood (ML)}$$
$$c(\theta) = \ell(\theta; \mathcal{D}) + r(\theta) \qquad \text{maximum a posteriori (MAP)/penalizedML}$$

  (also cross-validation, Bayesian estimators, BIC, AIC, ...)

Maximum Likelihood

- For IID data, the log likelihood is a sum of identical functions:

$$p(\mathcal{D}|\theta) = \prod_m p(\mathbf{x}^m|\theta)$$
$$\ell(\theta; \mathcal{D}) = \sum_m \log p(\mathbf{x}^m|\theta)$$

- Idea of maximum likelihod estimation (MLE): pick the setting of parameters most likely to have generated the data we saw:

$$\theta^*_{\mathrm{ML}} = \mathrm{argmax}_\theta \ \ell(\theta; \mathcal{D})$$

- Very commonly used in statistics.
  Often leads to "intuitive", "appealing", or "natural" estimators.
- For a start, the IID assumption makes the log likelihood into a sum, so its derivative can be easily taken term by term.

Sufficient Statistics

- A statistic is a (possibly vector valued) function of a (set of) random variable(s).
- $T(\mathbf{X})$ is a "sufficient statistic" for $\mathbf{X}$ if

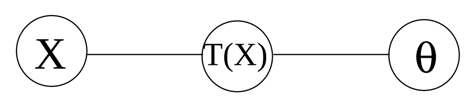
$$T(\mathbf{x}^1) = T(\mathbf{x}^2) \quad \Rightarrow \quad L(\theta; \mathbf{x}^1) = L(\theta; \mathbf{x}^2) \quad \forall \theta$$

- Equivalently (by the Neyman factorization theorem) we can write:

$$p(\mathbf{x}|\theta) = h\left(\mathbf{x}, T(\mathbf{x})\right) g\left(T(\mathbf{x}), \theta\right)$$

- Example: exponential family models:

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



Example: Bernoulli Trials 8

- We observe M iid coin flips: $\mathcal{D}$=H,H,T,H,...
- Model: $p(H) = \theta \quad p(T) = (1-\theta)$
- Likelihood:

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= \log \prod_m \theta^{\mathbf{x}^m} (1-\theta)^{1-\mathbf{x}^m} \\ &= \log\theta \sum_m \mathbf{x}^m + \log(1-\theta) \sum_m (1-\mathbf{x}^m) \\ &= \log\theta N_{\mathrm{H}} + \log(1-\theta) N_{\mathrm{T}}\end{aligned}$$

- Take derivatives and set to zero:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= \frac{N_{\mathrm{H}}}{\theta} - \frac{N_{\mathrm{T}}}{1-\theta} \\ \Rightarrow \theta^*_{\mathrm{ML}} &= \frac{N_{\mathrm{H}}}{N_{\mathrm{H}} + N_{\mathrm{T}}}\end{aligned}$$

Example: Multinomial 9

- We observe M iid die rolls (K-sided): $\mathcal{D}$=3,1,K,2,...
- Model: $p(k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood (for binary indicators $[\mathbf{x}^m = k]$):

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= \log \prod_m \theta_{\mathbf{x}^m} = \log \prod_m \theta_1^{[\mathbf{x}^m=1]} \ldots \theta_k^{[\mathbf{x}^m=k]} \\ &= \sum_k \log\theta_k \sum_m [\mathbf{x}^m = k] = \sum_k N_k \log\theta_k\end{aligned}$$

- Take derivatives and set to zero (enforcing $\sum_k \theta_k = 1$):

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_k} &= \frac{N_k}{\theta_k} - M \\ \Rightarrow \theta^*_k &= \frac{N_k}{M}\end{aligned}$$

Example: Univariate Normal 10

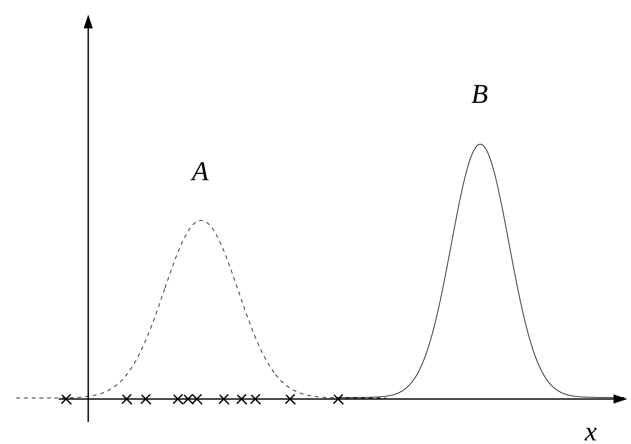
- We observe M iid real samples: $\mathcal{D}$=1.18,-.25,.78,...
- Model: $p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/2\sigma^2\}$
- Likelihood (using probability density):

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= -\frac{M}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_m \frac{(x^m-\mu)^2}{\sigma^2}\end{aligned}$$

- Take derivatives and set to zero:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= (1/\sigma^2) \textstyle\sum_m (x_m - \mu) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4} \textstyle\sum_m (x_m-\mu)^2 \\ \Rightarrow \mu_{\mathrm{ML}} &= (1/M) \textstyle\sum_m x_m \\ \sigma^2_{\mathrm{ML}} &= (1/M) \textstyle\sum_m x_m^2 - \mu^2_{\mathrm{ML}}\end{aligned}$$

Example: Univariate Normal 11



Example: Linear Regression

- At a linear regression node, some parents (covariates/inputs) and all children (responses/outputs) are continuous valued variables.
- For each child and setting of discrete parents we use the model:

$$p(y|\mathbf{x}, \theta) = \mathrm{gauss}(y|\theta^\top \mathbf{x}, \sigma^2)$$

- The likelihood is the familiar "squared error" cost:

$$\ell(\theta; \mathcal{D}) = -\frac{1}{2\sigma^2} \sum_m (y^m - \theta^\top \mathbf{x}^m)^2$$

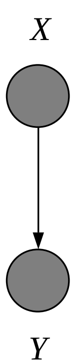
- The ML parameters can be solved for using linear least-squares:

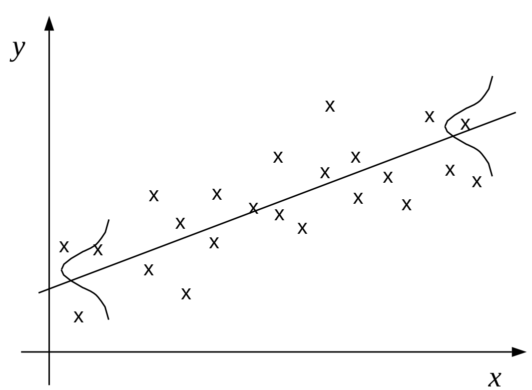
$$\frac{\partial \ell}{\partial \theta} = -\sum_m (y^m - \theta^\top \mathbf{x}^m)\mathbf{x}^m$$

$$\Rightarrow \theta^*_{\mathrm{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$$

- Sufficient statistics are input correlation matrix and input-output cross-correlation vector.

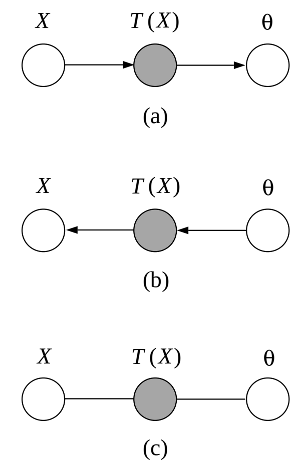
Example: Linear Regression





Sufficient Statistics are Sums

- In the examples above, the sufficient statistics were merely sums (counts) of the data:
  Bernoulli: # of heads, tails
  Multinomial: # of each type
  Gaussian: mean, mean-square
  Regression: correlations
- As we will see, this is true for all exponential family models: sufficient statistics are the *average natural parameters*.
- Only* exponential family models have simple sufficient statistics.

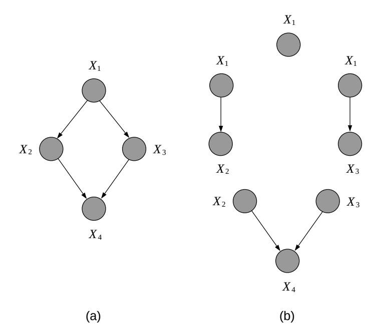


MLE for Directed GMs

- For a directed GM, the likelihood function has a nice form:

$$\log p(\mathcal{D}|\theta) = \log \prod_m \prod_i p(\mathbf{x}_i^m|\mathbf{x}_{\pi_i}, \theta_i) = \sum_m \sum_i \log p(\mathbf{x}_i^m|\mathbf{x}_{\pi_i}, \theta_i)$$

- The parameters *decouple*; so we can maximize likelihood independently for each node's function by setting θ_i.
- Only need the values of $\mathbf{x}_i$ and its parents in order to estimate θ_i.
- Furthermore, if $\mathbf{x}_i, \mathbf{x}_{\pi_i}$ have sufficient statistics only need those.
- In general, for fully observed data if we know how to estimate params at a single node we can do it for the whole network.

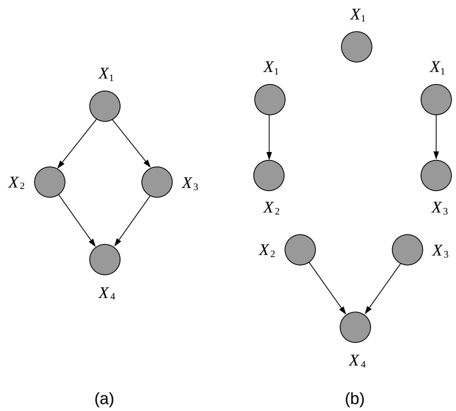


Example: A Directed Model 16

- Consider the distribution defined by the DAGM:

$$p(\mathbf{x}|\theta) = p(\mathbf{x}_1|\theta_1)p(\mathbf{x}_2|\mathbf{x}_1, \theta_2)p(\mathbf{x}_3|\mathbf{x}_1, \theta_3)p(\mathbf{x}_4|\mathbf{x}_2, \mathbf{x}_3, \theta_4)$$

- This is exactly like learning four separate small DAGMs, each of which consists of a node and its parents (not its Markov blanket).



MLE for Multinomial Networks 17

- Assume our DAGM contains only discrete nodes, and we use the (general) multinomial form for the conditional probabilities.
- Sufficient statistics involve counts of joint settings of $\mathbf{x}_i, \mathbf{x}_{\pi_i}$ summing over all other variables in the table.
- Likelihood for these special "fully observed multinomial networks":

$$\begin{aligned}
\ell(\theta; \mathcal{D}) &= \log \prod_{m,i} p(\mathbf{x}_i^m | \mathbf{x}_{\pi_i}{}^m, \theta_i) \\
&= \log \prod_{i,\mathbf{x}_i,\mathbf{x}_{\pi_i}} p(\mathbf{x}_i|\mathbf{x}_{\pi_i}, \theta_i)^{N(\mathbf{x}_i,\mathbf{x}_{\pi_i})} = \log \prod_{i,\mathbf{x}_i,\mathbf{x}_{\pi_i}} \theta_{\mathbf{x}_i|\mathbf{x}_{\pi_i}}^{N(\mathbf{x}_i,\mathbf{x}_{\pi_i})} \\
&= \sum_i \sum_{\mathbf{x}_i,\mathbf{x}_{\pi_i}} N(\mathbf{x}_i, \mathbf{x}_{\pi_i}) \log \theta_{\mathbf{x}_i|\mathbf{x}_{\pi_i}}
\end{aligned}$$

$$\Rightarrow \theta^*_{\mathbf{x}_i|\mathbf{x}_{\pi_i}} = \frac{N(\mathbf{x}_i, \mathbf{x}_{\pi_i})}{N(\mathbf{x}_{\pi_i})}$$

MLE for General Exponential Family Models 18

- Recall the probability function for models in the exponential family:

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

- For iid data, the sufficient statistic vector is $\sum_m T(\mathbf{x}^m)$:

$$\ell(\eta; \mathcal{D}) = \log p(\mathcal{D}|\eta) = \left(\sum_m \log h(\mathbf{x}^m)\right) - MA(\eta) + \left(\eta^\top \sum_m T(\mathbf{x}^m)\right)$$

- Take derivatives and set to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \eta} &= \textstyle\sum_m T(\mathbf{x}^m) - M\frac{\partial A(\eta)}{\partial \eta} \\
\Rightarrow \frac{\partial A(\eta)}{\partial \eta} &= \frac{1}{M}\textstyle\sum_m T(\mathbf{x}^m) \\
\eta_{\mathrm{ML}} &= \frac{1}{M}\textstyle\sum_m T(\mathbf{x}^m)
\end{aligned}$$

recalling that the natural moments of an exponential distribution are the derivatives of the log normalizer.