Learning Graphical Models from Data

- 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).
- 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 ≡ setting parameters. (“Structure learning” is also possible but we won’t consider it now.)

Likelihood Function

- So far we have focused on the (log) probability function \( p(x|\theta) \) which assigns a probability (density) to any joint configuration of variables \( x \) given fixed parameters \( \theta \).
- But in learning we turn this on its head: we have some fixed data and we want to find parameters.
- Think of \( p(x|\theta) \) as a function of \( \theta \) for fixed \( x \):
  \[
  L(\theta; x) = p(x|\theta) \\
  \ell(\theta; x) = \log p(x|\theta)
  \]
  This function is called the (log) “likelihood”.
- Chose \( \theta \) to maximize some cost function \( c(\theta) \) which includes \( \ell(\theta) \):
  \[
  c(\theta) = \ell(\theta; D) \\
  c(\theta) = \ell(\theta; D) + r(\theta)
  \]
  maximum likelihood (ML)  \\
  maximum a posteriori (MAP)/penalizedML 
  (also cross-validation, Bayesian estimators, BIC, AIC, ...)

Multiple Observations, Complete Data, IID Sampling

- A single observation of the data \( X \) is rarely useful on its own.
- Generally we have data including many observations, which creates a set of random variables: \( D = \{x^1, x^2, \ldots, x^M\} \)
- 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.)
**Maximum Likelihood**

- For IID data:
  \[ p(D|\theta) = \prod_{m} p(x^m|\theta) \]
  \[ \ell(\theta; D) = \sum_{m} \log p(x^m|\theta) \]

- Idea of maximum likelihood estimation (MLE): pick the setting of parameters most likely to have generated the data we saw:
  \[ \theta^*_{ML} = \arg\max_{\theta} \ell(\theta; D) \]

- 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).
- \( T(X) \) is a “sufficient statistic” for \( X \) if
  \[ T(x^1) = T(x^2) \Rightarrow T(x) \in \mathcal{L}(\theta; x^1) = L(\theta; x^2) \quad \forall \theta \]
- Equivalently (by the factorization theorem) we can write:
  \[ p(x|\theta) = h(x) T(x) g(T(x), \theta) \]

- Example: exponential family models:
  \[ p(x|\theta) = h(x) \exp \left\{ \sum_{i}^n \theta_i \tau_i(x) - A(\theta) \right\} \]

- Example: Bernoulli Trials
  - We observe \( M \) iid coin flips: \( D=H,H,T,H,\ldots \)
  - Model: \( p(H) = \theta \quad p(T) = (1-\theta) \)
  - Likelihood:
    \[ \ell(\theta; D) = \log p(D|\theta) \]
    \[ = \log \prod_{m} \theta^{x^m} (1-\theta)^{1-x^m} \]
    \[ = \log \theta \sum_{m} x^m + \log(1-\theta) \sum_{m} (1-x^m) \]
    \[ = \log \theta N_H + \log(1-\theta) N_T \]
  - Take derivatives and set to zero:
    \[ \frac{\partial \ell}{\partial \theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} \]
    \[ \Rightarrow \theta^*_{ML} = \frac{N_H}{N_H + N_T} \]

- Example: Multinomial
  - We observe \( M \) iid die rolls (K-sided): \( D=3,1,K,2,\ldots \)
  - Model: \( p(k) = \theta_k \quad \sum_k \theta_k = 1 \)
  - Likelihood (for binary indicators \([x^m=k] \)):
    \[ \ell(\theta; D) = \log p(D|\theta) \]
    \[ = \log \prod_{m} \theta_{x^m} \]
    \[ = \sum_{k} \log \theta_k \sum_{m} [x^m = k] \]
    \[ = \sum_{k} N_k \log \theta_k \]
  - Take derivatives and set to zero (enforcing \( \sum_k \theta_k = 1 \)):
    \[ \frac{\partial \ell}{\partial \theta_k} = \frac{N_k}{\theta_k} - M \]
    \[ \Rightarrow \theta^*_{ML} = \frac{N_k}{M} \]
Example: Univariate Normal

- We observe $M$ iid real samples: $D = 1.18, -0.25, 0.78, \ldots$
- Model: $p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x - \mu)^2/2\sigma^2\}$
- Likelihood (using probability density):
  $$\ell(\theta; D) = \log p(D|\theta)$$
  $$= -\frac{M}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_m (x^m - \mu)^2$$
- Take derivatives and set to zero:
  $$\frac{\partial \ell}{\partial \mu} = 1/\sigma^2 \sum_m (x^m - \mu)$$
  $$\frac{\partial \ell}{\partial \sigma^2} = -M/2\sigma^2 + \frac{1}{2\sigma^4} \sum_m (x^m - \mu)^2$$
  $$\Rightarrow \mu_{ML} = (1/M) \sum_m x^m$$
  $$\sigma^2_{ML} = (1/M) \sum_m x^2_m - \mu^2_{ML}$$

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:
  $$p(y|x, \theta) = \text{gauss}(y|\theta^T x, \sigma^2)$$
- The likelihood is the familiar “squared error” cost:
  $$\ell(\theta; D) = -\frac{1}{2\sigma^2} \sum_m (y^m - \theta^T x^m)^2$$
- The ML parameters can be solved for using linear least-squares:
  $$\frac{\partial \ell}{\partial \theta} = -\sum_m (y^m - \theta^T x^m)x^m$$
  $$\Rightarrow \theta_{ML}^* = (X^T X)^{-1} X^T Y$$
- Sufficient statistics are input correlation matrix and input-output cross-correlation vector.
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.

*The parameters decouple; so we can maximize likelihood independently for each node’s function by setting $\theta_i$.

MLE for Directed GMs

For a directed GM, the likelihood function has a nice form:
$$
\log p(D|\theta) = \log \prod_{m} \prod_{i} p(x_i^m|\pi_{x_i}, \theta_i) = \sum_{m} \sum_{i} \log p(x_i^m|\pi_{x_i}, \theta_i)
$$

The parameters decouple; so we can maximize likelihood independently for each node’s function by setting $\theta_i$.

MLE for Multinomial Networks

Consider the distribution defined by the DAGM:
$$p(x|\theta) = p(x_1|\theta_1)p(x_2|x_1, \theta_2)p(x_3|x_1, \theta_3)p(x_4|x_2, x_3, \theta_4)$$

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

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 $x_i, x_{\pi_i}$ summing over all other variables in the table.

Likelihood for these special "fully observed multinomial networks":
$$
\ell(\theta; D) = \log \prod_{m, i} p(x_i^m|\pi_{x_i}^m, \theta_i)
$$

$$
= \log \prod_{i, x_i, \pi_{x_i}} p(x_i|x_{\pi_i}, \theta_i)^{N(x_i, x_{\pi_i})} = \log \prod_{i, x_i, \pi_{x_i}} \theta_{x_i|x_{\pi_i}}^N(x_i, x_{\pi_i})
$$

$$
= \sum_{i} N(x_i, x_{\pi_i}) \log \theta_{x_i|x_{\pi_i}}
$$

$$
\Rightarrow \theta_{x_i|x_{\pi_i}} = \frac{N(x_i, x_{\pi_i})}{N(x_{\pi_i})}$$
MLE for General Exponential Family Models

- Recall the probability function for models in the exponential family:
  \[ p(x|\theta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\} \]

- For iid data, the sufficient statistic vector is \( \sum_m T(x^m) \):
  \[ \ell(\eta; D) = \log p(D|\eta) = \left( \sum_m \log h(x^m) \right) - MA(\eta) + \left( \eta^\top \sum_m T(x^m) \right) \]

- Take derivatives and set to zero:
  \[ \frac{\partial \ell}{\partial \eta} = \sum_m T(x^m) - MA(\eta) \]
  \[ \Rightarrow \frac{\partial A(\eta)}{\partial \eta} = \frac{1}{M} \sum_m T(x^m) \]
  \[ \eta_{ML} = \frac{1}{M} \sum_m T(x^m) \]
  recalling that the natural moments of an exponential distribution are the derivatives of the log normalizer.