Learning Intro

COMPSCI 688: Probabilistic Graphical Models
Lecture 5: Learning in Directed Graphical Models

Dan Sheldon
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Learning Intro

Example: Bayesian Network Graph

Example: Conditional Probability Table
Bayesian Networks: Parameters

The default parameterization in a discrete Bayesian network simply uses a separate parameter for each element of each CPT:

\[
P(\mathbf{X} = \mathbf{x} | \mathbf{X}_{\text{pa}}(\mathbf{X}) = \mathbf{y}) = \theta_{\mathbf{x}\mathbf{y}}
\]

\[
\Theta = (\ldots \ldots )
\]

\[
\Theta_{\mathbf{x}\mathbf{y}}
\]

Today's Problem

- How do we choose the parameter values for a Bayesian network given a data set?
- The maximum likelihood estimate for \( \theta_{\mathbf{x}\mathbf{y}} \) is just the number of times \( \mathbf{X} \) takes value \( \mathbf{x} \) when its parents take value \( \mathbf{y} \), divided by the number of times its parents take the value \( \mathbf{y} \):

\[
P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = \frac{\#(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}{\#(\mathbf{Y} = \mathbf{y})}
\]

How can we derive this result?
For a data set $x^{(1:N)} = (x^{(1)}, \ldots, x^{(N)})$, the log-likelihood is

$$\mathcal{L}(\theta|x^{(1:N)}) = \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x^{(n)})$$

(assumes independence)

Goal: find $\theta$ to maximize $\mathcal{L}(\theta|x^{(1:N)})$.

$$p_\theta(x): \mathcal{X} \to \mathbb{R}^+$$

Example: Bernoulli Model

Suppose $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ are drawn from a Bernoulli distribution:

$$p_\theta(x) = \begin{cases} 1 - \theta, & x = 0, \\ \theta, & x = 1. \end{cases}$$

The log-likelihood is

$$\mathcal{L}(\theta|x^{(1:N)}) = \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x^{(n)}) + \sum_{x^{(n)}} \log(1 - \theta) \cdot \mathbb{I}[x^{(n)} = 0] + \log \theta \cdot \mathbb{I}[x^{(n)} = 1]$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[x^{(n)} = 0] \log(1 - \theta) + \mathbb{I}[x^{(n)} = 1] \log \theta$$

$$= \frac{\#(X = 0)}{N} \log(1 - \theta) + \frac{\#(X = 1)}{N} \log \theta.$$

What does this likelihood function look like?
Learning Intro

Example: Bernoulli Likelihood

Learning as Likelihood Maximization

How can we find the model parameters $\theta$ that maximize the likelihood?
- The derivative of a function is zero at every local maximum

Learning as Likelihood Maximization

How can we find the model parameters $\theta$ that maximize the likelihood?
- The derivative of a function is zero at every local maximum
Learning as Likelihood Maximization

How can we find the model parameters $\theta$ that maximize the likelihood?

- The derivative of a function is zero at every local maximum
- Zero derivative points are not local maxima in general.

Maximum Likelihood and Optimization

How can we find the model parameters $\theta$ that maximize the likelihood?

- The derivative of a function is zero at every local maximum
- Zero derivative points are not local maxima in general.

To be a local maximum, the curvature must be negative

$$ \det \text{Hessian} < 0 $$

(Determine which solutions are local maxima by checking second derivatives)
Example: Bernoulli Likelihood

The maximum likelihood estimates for the simple Bernoulli model are easy to derive:

\[ \mathcal{L}(\theta | x^{1:N}) = \frac{\#(X = 0)}{N} \log(1 - \theta) + \frac{\#(X = 1)}{N} \log \theta \]

\[ \frac{\partial}{\partial \theta} \mathcal{L}(\theta | x^{1:N}) = -\frac{\#(X = 0)}{N} \frac{1}{1-\theta} + \frac{\#(X = 1)}{N} \frac{1}{\theta} = 0 \]

Setting the derivative equation equal to zero and solving yields the maximum likelihood estimate:

\[ \hat{\theta} = \frac{\#(X = 1)}{N} \]

\[ \hat{\theta} - \frac{1}{1-\theta} = 0 \quad \Rightarrow \quad \frac{\hat{\theta}}{1-\theta} = 1 \quad \Rightarrow \quad \theta = \hat{\theta} \]

Example: Multinomial Model

Consider a Multinomial model for a discrete random variable \( X \) that takes \( V \) values \( \{1,...,V\} \).

The likelihood function for the Multinomial model is:

\[ \mathcal{L}(\Theta | x^{1:N}) = \prod_{v=1}^{V} \left( \frac{x_v \Theta_v}{\sum_{v=1}^{V} \Theta_v} \right)^{x_v} \]

For \( \Theta = \{\Theta_1, \ldots, \Theta_V\} \), the log-likelihood function is:

\[ \log \mathcal{L}(\Theta | x^{1:N}) = \sum_{v=1}^{V} \log \left( \frac{x_v \Theta_v}{\sum_{v=1}^{V} \Theta_v} \right) \]

The maximum likelihood estimates are:

\[ \hat{\Theta}_v = \frac{x_v}{\sum_{v=1}^{V} x_v} \]

The log-likelihood function can also be written as:

\[ \log \mathcal{L}(\Theta | x^{1:N}) = \sum_{v=1}^{V} x_v \log \Theta_v - \log \left( \sum_{v=1}^{V} \Theta_v \right) \]

For \( \Theta = \{\Theta_1, \ldots, \Theta_V\} \), the log-likelihood function is:

\[ \log \mathcal{L}(\Theta | x^{1:N}) = \sum_{v=1}^{V} x_v \log \Theta_v - \log \left( \sum_{v=1}^{V} \Theta_v \right) \]

The maximum likelihood estimates are:

\[ \hat{\Theta}_v = \frac{x_v}{\sum_{v=1}^{V} x_v} \]

The log-likelihood function can also be written as:

\[ \log \mathcal{L}(\Theta | x^{1:N}) = \sum_{v=1}^{V} x_v \log \Theta_v - \log \left( \sum_{v=1}^{V} \Theta_v \right) \]

For \( \Theta = \{\Theta_1, \ldots, \Theta_V\} \), the log-likelihood function is:

\[ \log \mathcal{L}(\Theta | x^{1:N}) = \sum_{v=1}^{V} x_v \log \Theta_v - \log \left( \sum_{v=1}^{V} \Theta_v \right) \]

The maximum likelihood estimates are:

\[ \hat{\Theta}_v = \frac{x_v}{\sum_{v=1}^{V} x_v} \]
Example: Multinomial Parameter Learning

- \( \mathcal{L}(\theta|X^{1:N}) = \sum_{v=1}^{V-1} \frac{\#(X = v)}{N} \log(\theta_v) + \frac{\#(X = V)}{N} \log(1 - \sum_{v=1}^{V-1} \theta_v) \)

- Setting the partial derivatives to zero, we require, for each \( i < V \):
  \[
  \frac{\partial}{\partial \theta_i} \mathcal{L}(\theta|X^{1:N}) = \frac{\#(X = i)}{N \theta_i} - \frac{\#(X = V)}{N(1 - \sum_{v=1}^{V-1} \theta_v)} = 0
  \]
- It’s easy to check that this is solved by setting
  \[
  \theta_i = \frac{\#(X = i)}{N}
  \]

Bayesian Network Parameters
In a Bayesian network, each CPT is a collection of multinomial distributions with distinct parameters. There is one multinomial distribution for each joint setting of the parents of each variable.

| HD | G  | BP   | C       | \( P(HD|G, BP, C) \) |
|----|----|------|---------|---------------------|
| No | M  | Low  | Low     | \( \theta_{HD}^{N|M,L,L} \) |
| Yes| M  | Low  | Low     | \( \theta_{HD}^{Y|M,L,L} \) |
| No | F  | Low  | Low     | \( \theta_{HD}^{N|F,L,L} \) |
| Yes| F  | Low  | Low     | \( \theta_{HD}^{Y|F,L,L} \) |
|    |    |      |         |                     |

\[
\log P(HD = h|G = g, BP = b, C = e) = \log \theta_{HD}^{h|g,b,c}
\]
Log Likelihood Decomposition

The log likelihood of a dataset \( x^{(1:N)} \) for a Bayesian network decomposes into a sum of terms that depend only on the parameters for one conditional distribution:

\[
\mathcal{L}(\theta | x^{(1:N)}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} \log \theta_{x_d}(x_d^{(n)}) \sum_{\text{pa}(d)} \mathbb{I}[x_d^{(n)} = x_d, x_{\text{pa}(d)}^{(n)} = x_{\text{pa}(d)}] \log \theta_{x_d}(x_d | x_{\text{pa}(d)}).
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{x_d} \sum_{x_{\text{pa}(d)}} \mathbb{I}[x_d^{(n)} = x_d, x_{\text{pa}(d)}^{(n)} = x_{\text{pa}(d)}] \log \theta_{x_d}(x_d | x_{\text{pa}(d)}).
\]

\[
= \frac{1}{N} \sum_{d=1}^{D} \sum_{x_d} \sum_{x_{\text{pa}(d)}} \#(X_d = x_d, X_{\text{pa}(d)} = x_{\text{pa}(d)}) \log \theta_{x_d}(x_d | x_{\text{pa}(d)}).
\]

Example: Heart Disease Log Likelihood

\[
\mathcal{L}(\theta | x^{(1:N)}) = \sum_{g} \#(G = g) \frac{1}{N} \log \theta_{g}^G + \sum_{B} \#(BP = b) \frac{1}{N} \log \theta_{b}^{BP} + \sum_{C} \#(C = c) \frac{1}{N} \log \theta_{c}^C
\]

\[
+ \sum_{h} \sum_{g,b,c} \#(HD = h, G = g, BP = b, C = c) \frac{1}{N} \log \theta_{h|g,b,c}^{HD}.
\]
Bayesian Network Learning Algorithm

Example: Heart Disease Parameter De-Coupling

\[
\text{Learning Intro}
\]

\[
\text{MLE Examples}
\]

\[
\text{Learning Bayesian Networks}
\]

\[
\text{Estimation Theory}
\]

\[
\text{Learning Intro}
\]

\[
\text{MLE Examples}
\]

\[
\text{Learning Bayesian Networks}
\]

\[
\text{Estimation Theory}
\]

\[
\text{Learning Intro}
\]

\[
\text{MLE Examples}
\]

\[
\text{Learning Bayesian Networks}
\]

\[
\text{Estimation Theory}
\]

\[
\text{Learning Intro}
\]

\[
\text{MLE Examples}
\]

\[
\text{Learning Bayesian Networks}
\]

\[
\text{Estimation Theory}
\]

Example: Heart Disease Parameter De-Coupling

\[
\begin{align*}
\text{P}(G) & \quad \text{Gender} \\
\text{P}(C) & \quad \text{Cholesterol} \\
\text{P}(BP) & \quad \text{BloodPressure} \\
\hline
\text{P}(HD(G,C,BP)) & \quad \text{HeartDisease}
\end{align*}
\]

\[
\begin{align*}
& \max_{g^G} \sum_g \frac{\#(G = g)}{N} \cdot \log \theta_g^G \\
\text{Subject to} \quad & \sum_g \theta_g^G = 1
\end{align*}
\]

Bayesian Network Learning Summary

- The only parameters that must be jointly optimized in a Bayesian network are those in the same sum-to-one constraint with the same setting of the parent variables.
- For any random variable \( X \), consider a specific setting of its parent variables \( Y = y \). We just need to jointly optimize the parameters \( \theta_{x|y}^X \) for each value \( x \in \text{Val}(X) \).
- This is just multinomial parameter estimation applied to each variable \( X \) for each setting \( y \) of its parents:

\[
P(\theta_{x|y}^X | X = x, Y = y) = \frac{\#(X = x, Y = y)}{\#(Y = y)}
\]

For each random variable \( X_d \):
- For each joint configuration \( x_{pa(d)} \in \text{Val}(X_{pa(d)}) \):
  - For each value \( x_d \in \text{Val}(X_d) \). Set

\[
\theta_{x_d|x_{pa(d)}}^X \leftarrow \frac{\#(X_d = x_d, X_{pa(d)} = x_{pa(d)})}{\#(X_{pa(d)} = x_{pa(d)})}
\]
Here is a more general problem: suppose we have an arbitrary target distribution $p_*$ and a parametric model $M = \{p_\theta | \theta \in \Theta\}$.

How can we select $p_\theta^* \in M$ that is as close as possible to $p_*$?
Parameter Selection: Case 2

KL Divergence Minimization

- If \( p_\ast \in M \) then there exists a \( \theta^\ast \) such that \( p_\ast = p_{\theta^\ast} \).

Kullback-Leibler Divergence

One of the most used divergence criteria is the Kullback-Leibler divergence.

\[
KL(p||q) = \sum_{x \in \text{Val}(X)} p(x) \log \left( \frac{p(x)}{q(x)} \right)
\]

The KL divergence is a pre-metric. It satisfies:
- \( KL(p||q) \geq 0 \) for all \( p, q \)
- \( KL(p||q) = 0 \) if and only if \( p = q \)

It does not satisfy:
- \( KL(p||q) = KL(q||p) \) for all \( p, q \)
- \( KL(p||q) \leq KL(p||s) + KL(s||q) \) for all \( p, q, s \)
KL Divergence Minimization

- If $p_\star \in M$ then there exists a $\theta^\star$ such that $p_\star = p_{\theta^\star}$.
- If $p_\star$ is not in $M$ then we select the $\theta^\star$ that minimizes $KL(p_\star||p_{\theta^\star})$ over the parameter space $\Theta$.

KL Divergence Minimization Simplification

Minimizing $KL(p_\star||p_{\theta^\star})$ is the same as maximizing $\mathcal{L}(\theta|p_\star) = \sum_{x \in \text{Val}(X)} p_\star(x) \log p_{\theta}(x)$.

Maximum Likelihood = KL Minimization

Suppose $p_\star$ is the empirical distribution of a data set $x^{(1)}, \ldots, x^{(N)}$, meaning it places $\frac{1}{N}$ probability on each data point. Then

$$\mathcal{L}(\theta|p_\star) = \sum_{x \in \text{Val}(X)} p_\star(x) \log p_{\theta}(x) = \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(x^{(n)}) = \mathcal{L}(\theta|x^{(1:N)})$$

$\implies$ maximum-likelihood estimation minimizes the KL-divergence from the empirical data distribution to $p_{\theta}$.

This is a reasonable behavior even when the data comes from a distribution that does not belong to the parametric model.