COMPSCI 688: Probabilistic Graphical Models

Lecture 5: Learning in Directed Graphical Models

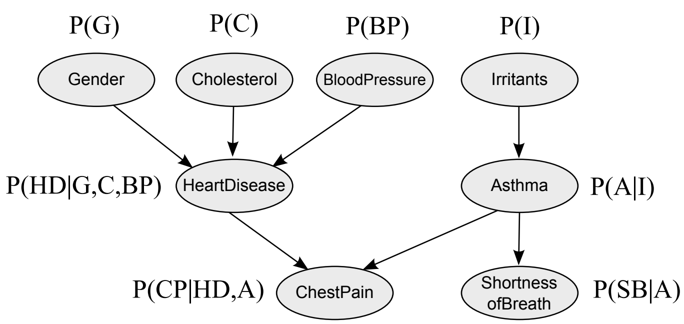
Dan Sheldon

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

HD	G	BP	C	$P(HD\|G,BP,C)$
No	M	Low	Low	0.95
Yes	M	Low	Low	0.05
No	F	Low	Low	0.99
Yes	F	Low	Low	0.01
⋮	⋮	⋮	⋮	⋮

Bayesian Networks: Parameters

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

$$P_\theta(X = x | \mathbf{X}_{\mathsf{pa}(X)} = \mathbf{y}) = \theta^X_{x|\mathbf{y}}$$

Bayesian Networks: Parameters

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	$\theta^{HD}_{Y\|F,L,L}$
⋮	⋮	⋮	⋮	⋮

Today's Problem

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

$$P_\theta(X = x | \mathbf{Y} = \mathbf{y}) = \theta^X_{x|\mathbf{y}} = \frac{\#(X = x, \mathbf{Y} = \mathbf{y})}{\#(\mathbf{Y} = \mathbf{y})}$$

How can we derive this result?

Example: Smoker and Cancer

Estimation

Maximum-Likelihood Estimation (MLE)

A parametric model $\{p_\theta | \theta \in \Theta\}$ is a family of probability distributions indexed by parameters θ

Given data $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$, how do we choose p_θ? (Notation: $\mathbf{x}^{(n)} = (x_1^{(n)}, \dots, x_d^{(n)})$)

Principle of maximum likelihood: choose the distribution that assigns the highest probability to the data

For an observed value $\mathbf{x}$, the **log-likelihood** is

$$\mathcal{L}(\theta|\mathbf{x}) = \log p_\theta(\mathbf{x})$$

For a data set $\mathbf{x}^{(1:N)} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$, the log-likelihood is

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

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

Example: Bernoulli Model

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

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

The log-likelihood is

$$\begin{aligned} \mathcal{L}(\theta|x^{(1:N)}) &= \frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x^{(n)}) \\ &= \frac{1}{N}\sum_{n=1}^{N} \left(\mathbb{I}[x^{(n)} = 0]\log(1-\theta) + \mathbb{I}[x^{(n)} = 1]\log\theta\right) \\ &= \frac{\#(X=0)}{N}\log(1-\theta) + \frac{\#(X=1)}{N}\log\theta. \end{aligned}$$

What does this likelihood function look like?

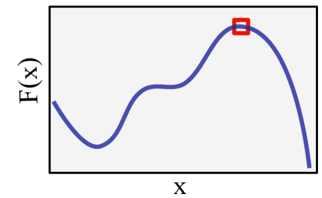
Example: Bernoulli Likelihood

Demo:
Likelihood Function

Learning as Likelihood Maximization

How can we find the model parameters θ 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



Maximum Likelihood and Optimization

How can we find the model parameters θ that maximize the likelihood?

- Compute the (partial) derivatives of the log likelihood
- Set them equal to zero
- Solve derivative equations for the parameters
- (Determine which solutions are local maxima by checking second derivatives)

MLE Examples

Example: Bernoulli Likelihood

Demo:
Likelihood Function

Example: Bernoulli Parameter Learning

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

- $\mathcal{L}(\theta|x^{(1:N)}) = \frac{\#(X=0)}{N}\log(1-\theta) + \frac{\#(X=1)}{N}\log\theta$
- $\frac{\partial}{\partial\theta}\mathcal{L}(\theta|x^{(1:N)}) = \frac{\#(X=1)}{N\theta} - \frac{\#(X=0)}{N(1-\theta)}$
- Setting the derivative equation equal to zero and solving yields the maximum likelihood estimate:
$$\theta = \frac{\#(X=1)}{N}$$

Example: Multinomial Model

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

$$p_\theta(x) = \begin{cases} \theta_1 & x = 1 \\ \vdots & \\ \theta_{V-1} & x = V-1 \\ 1 - \sum_{v=1}^{V-1}\theta_v & x = V \end{cases}$$

Then

$$\mathcal{L}(\theta|x^{(1:N)}) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{v=1}^{V-1}\mathbb{I}[x^{(n)}=v]\log(\theta_v) + \mathbb{I}[x^{(n)}=V]\log\left(1-\sum_{v=1}^{V-1}\theta_v\right)\right)$$
$$= \sum_{v=1}^{V-1}\frac{\#(X=v)}{N}\log(\theta_v) + \frac{\#(X=V)}{N}\log\left(1-\sum_{v=1}^{V-1}\theta_v\right)$$

Example: Multinomial Parameter Learning

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

$$\theta_i = \frac{\#(X=i)}{N}$$

Learning Bayesian Networks

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	$\theta^{HD}_{Y\|F,L,L}$
⋮	⋮	⋮	⋮	⋮

$$\log P(HD = h|G = g, BP = b, C = c) = \log \theta^{HD}_{h|g,b,c}$$

Joint Probability in Terms of Parameters

The joint probability in a Bayesian network is a product of conditional multinomial distribution for each node:

$$p_\theta(\mathbf{x}) = \prod_{d=1}^{D} p_\theta(x_d|\mathbf{x}_{\mathsf{pa}(d)}) = \prod_{d=1}^{D} \theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}}$$

$\Longrightarrow$ log-likelihood is a sum of terms:

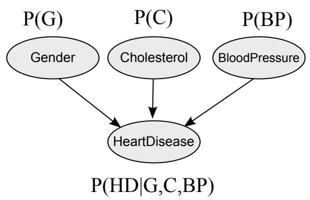
$$\log p_\theta(\mathbf{x}) = \sum_{d=1}^{D} \log \theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}}$$

Log Likelihood Decomposition

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

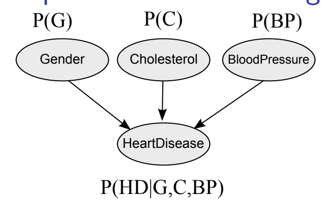
$$\begin{aligned}\mathcal{L}(\theta|\mathbf{x}^{(1:N)}) &= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\log \theta^{X_d}_{x_d^{(n)}|\mathbf{x}^{(n)}_{\mathsf{pa}(d)}} \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{x_d}\sum_{\mathbf{x}_{\mathsf{pa}(d)}} \mathbb{I}[x_d^{(n)} = x_d, \mathbf{x}^{(n)}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)}] \log \theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}} \\ &= \sum_{d=1}^{D}\sum_{x_d}\sum_{x_{\mathsf{pa}(d)}} \frac{\#(X_d = x_d, \mathbf{X}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)})}{N} \log \theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}}\end{aligned}$$

Example: Heart Disease Joint Distribution



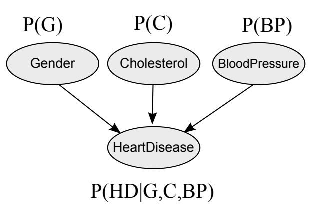
$$p_\theta(g, c, b, h) = p_\theta(g)p_\theta(b)p_\theta(c)p_\theta(h|g, b, c)$$

Example: Heart Disease Log Likelihood



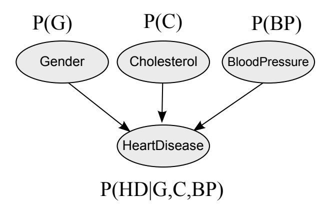
$$\mathcal{L}(\theta|\mathbf{x}^{(1:N)}) = \sum_g \frac{\#(G=g)}{N} \log \theta_g^G + \sum_b \frac{\#(BP=b)}{N} \log \theta_b^{BP} + \sum_c \frac{\#(C=c)}{N} \log \theta_c^C$$
$$+ \sum_{g,b,c} \sum_h \frac{\#(HD=h, G=g, BP=b, C=c)}{N} \log \theta_{h|g,b,c}^{HD}$$

Example: Heart Disease Parameter Learning



$$\max_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{x}^{(1:N)})$$

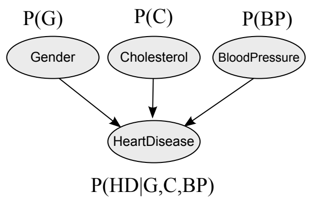
Example: Heart Disease Parameter De-Coupling



$$\max_{\theta^G} \sum_g \frac{\#(G=g)}{N} \cdot \log \theta_g^G$$

Subject to $\sum_g \theta_g^G = 1$

Example: Heart Disease Parameter De-Coupling



$$\max_{\theta^{HD}_{\cdot|g,b,c}} \sum_h \frac{\#(HD=h, G=g, BP=b, C=c)}{N} \cdot \log \theta^{HD}_{h|g,b,c}$$

Subject to $\sum_h \theta^{HD}_{h|g,b,c} = 1$

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 $\mathbf{Y} = \mathbf{y}$. We just need to jointly optimize the parameters $\theta^X_{x|\mathbf{y}}$ for each value $x \in \text{Val}(X)$.
- This is just multinomial parameter estimation applied to each variable X for each setting $\mathbf{y}$ of it's parents:

$$P_\theta(X = x|\mathbf{Y} = \mathbf{y}) = \theta^X_{x|\mathbf{y}} = \frac{\#(X = x, \mathbf{Y} = \mathbf{y})}{\#(\mathbf{Y} = \mathbf{y})}$$

Bayesian Network Learning Algorithm

- For each random variable X_d:
  - For each joint configuration $\mathbf{x}_{\mathsf{pa}(d)} \in \text{Val}(\mathbf{X}_{\mathsf{pa}(d)})$:
    - For each value $x_d \in \text{Val}(X_d)$. Set

$$\theta^X_{x_d|\mathbf{x}_{\mathsf{pa}(d)}} \leftarrow \frac{\#(X_d = x_d, \mathbf{X}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)})}{\#(\mathbf{X}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)})}$$

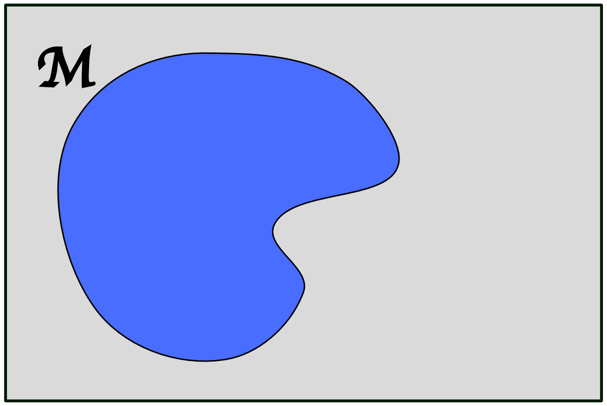
Estimation Theory

Estimation Theory

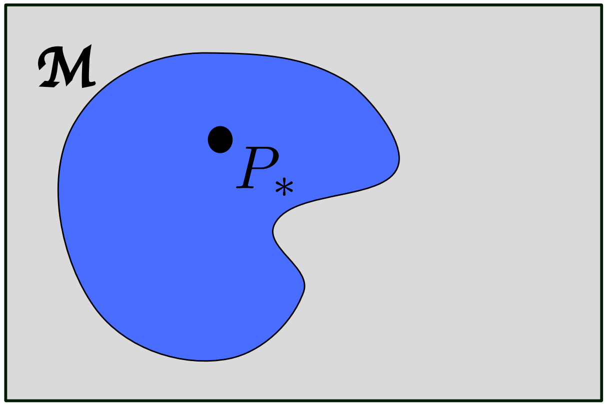
Here is a more general problem: suppose we have an arbitrary target distribution p_* and a parametric model $M = \{p_\theta | \theta \in \Theta\}$.

How can we select $p_{\theta^*} \in M$ that is as close as possible to p_*?

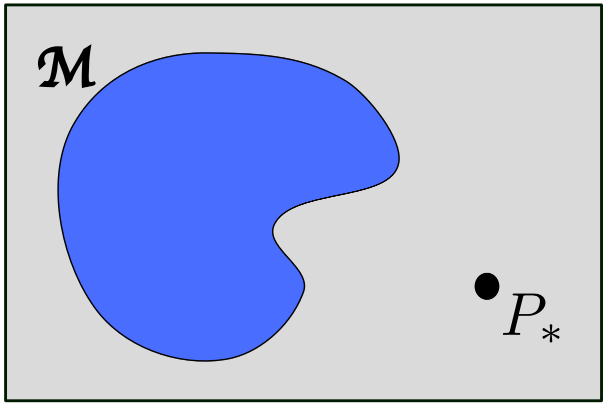
Parametric Probability Model



Parameter Selection: Case 1



Parameter Selection: Case 2



Kullback-Leibler Divergence

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

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

The KL divergence is a pre-metric. It satisfies:

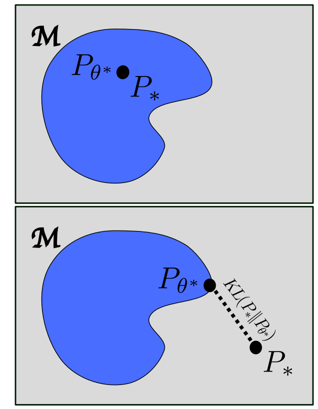
- $KL(p||q) \geq 0$ for all p and q
- $KL(p||q) = 0$ if and only if $p = q$

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



KL Divergence Minimization Simplification

$$\begin{aligned} KL(p_*||p_\theta) &= \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \log \left(\frac{p_*(\mathbf{x})}{p_\theta(\mathbf{x})} \right) \\ &= \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \left(\log p_*(\mathbf{x}) - \log p_\theta(\mathbf{x}) \right) \\ &= \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \log p_*(\mathbf{x}) - \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \log p_\theta(\mathbf{x}) \\ &= - \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \log p_\theta(\mathbf{x}) + C \end{aligned}$$

Minimizing $KL(p^*\|p_\theta)$ is the same as maximizing

$$\mathcal{L}(\theta|p_*) = \sum_{\mathbf{x} \in \text{Val}(\mathbf{X})} p_*(\mathbf{x}) \log p_\theta(\mathbf{x})$$

Maximum Likelihood = KL Minimization

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

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

$\implies$ maximum-likelihood estimation minimizes the KL-divergence from the empirical data distribution to p_θ.

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