Learning Intro

Example: Bayesian Network Graph

- P(G) - Gender
- P(C) - Cholesterol
- P(BP) - BloodPressure
- P(I) - Irritants

| 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      |

Figure 1: image
Bayesian Networks: Parameters

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

\[ P(\theta)(X=x|X_{pa}(X) = y) = \theta^X_{x|y} \]

Today’s Problem

▶ How do we choose the parameter values for a Bayesian network given a dataset?

▶ The maximum likelihood estimate for \( \theta^X_{x|y} \) is just the number of times \( X \) takes value \( x \) when its parents take value \( y \), divided by the number of times its parents take the value \( y \):

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

How can we derive this result?

Example: Smoker and Cancer
Maximum-Likelihood Estimation (MLE)

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

Given data \(x^{(1)}, \ldots, x^{(N)}\), how do we choose \(p_\theta\)? (Notation: \(x^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)})\))

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

For an observed value \(x\), the **log-likelihood** is
\[
L(\theta | x) = \log p_\theta(x)
\]

For a data set \(x^{(1:N)} = (x^{(1)}, \ldots, x^{(N)})\), the log-likelihood is
\[
L(\theta | x^{(1:N)}) = \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x^{(n)})
\]

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

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

What does this likelihood function look like?
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
- 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 \( \theta \) 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

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

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

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

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

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

\[ = \sum_{v=1}^{V-1} \frac{\#(X = v)}{N} \log(\theta_v) + \frac{\#(X = V)}{N} \log \left( 1 - \sum_{v=1}^{V-1} \theta_v \right) \]

Example: Multinomial Parameter Learning

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

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

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

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

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

Example: Bernoulli Parameter Learning

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

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

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

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

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

\[
\begin{array}{ccc}
\text{HD} & \text{G} & \text{BP} & \text{C} & P(\text{HD}|G, BP, C) \\
\hline
\text{No} & M & \text{Low} & \text{Low} & \theta_{\text{HD}|\text{M}, \text{L}, \text{L}} \\
\text{Yes} & M & \text{Low} & \text{Low} & \theta_{\text{HD}|\text{M}, \text{L}, \text{L}} \\
\text{No} & F & \text{Low} & \text{Low} & \theta_{\text{HD}|\text{F}, \text{L}, \text{L}} \\
\text{Yes} & F & \text{Low} & \text{Low} & \theta_{\text{HD}|\text{F}, \text{L}, \text{L}} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

\[
\log P(\text{HD} = h|G = g, BP = b, C = b) = \log \theta_{\text{HD}|g,b,p}
\]

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:

\[
\mathcal{L}(\theta|x^{(1:N)}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} \log \theta_{X_d|X_{\text{pa}(d)}}^{(n)}
\]

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

\[
= \sum_{d=1}^{D} \sum_{x_d} \sum_{x_{\text{pa}(d)}} \frac{\#(X_d = x_d, X_{\text{pa}(d)} = x_{\text{pa}(d)})}{N} \log \theta_{X_d|X_{\text{pa}(d)}}
\]
Example: Heart Disease Joint Distribution

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

Example: Heart Disease Log Likelihood

\[
\mathcal{L}(\theta|x^{1:N}) = \sum_g \frac{\#(G = g)}{N} \log \theta_G^g + \sum_b \frac{\#(BP = b)}{N} \log \theta_{BP}^b + \sum_c \frac{\#(C = c)}{N} \log \theta_C^c \\
+ \sum_{g,b,c} \frac{\#(HD = h,G = g,BP = b,C = c)}{N} \log \theta_{HD}^{h|g,b,c}
\]

Example: Heart Disease Parameter Learning

\[
\max_{\theta \in \Theta} \mathcal{L}(\theta|x^{1:N}) \\
\text{Subject to } \sum_g \theta_G^g = 1
\]

Example: Heart Disease Parameter De-Coupling

\[
\max_{\theta_G} \sum_g \frac{\#(G = g)}{N} \log \theta_G^g \\
\text{Subject to } \sum_g \theta_G^g = 1
\]
Example: Heart Disease Parameter De-Coupling

\[
P(G) \quad P(C) \quad P(BP) \\
\text{Gender} \quad \text{Cholesterol} \quad \text{BloodPressure} \\
P(HD|G,C,BP) \\
\max_{\theta} \sum_{h} \frac{\#(HD = h, G = g, BP = b, C = c)}{N} \cdot \log \theta_{HD}^{h,g,b,c} \\
\text{Subject to } \sum_{h} \theta_{HD}^{h,g,b,c} = 1
\]

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 $Y = y$. 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:

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

Bayesian Network Learning Algorithm

- 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 $p_*$?
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 \( p \) 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

\[ KL(p^*||p_\theta) = \sum_{x \in \text{Val}(X)} p_\star(x) \log \left( \frac{p_\star(x)}{p_\theta(x)} \right) \]

\[ = \sum_{x \in \text{Val}(X)} p_\star(x) (\log p_\star(x) - \log p_\theta(x)) \]

\[ = \sum_{x \in \text{Val}(X)} p_\star(x) \log p_\star(x) - \sum_{x \in \text{Val}(X)} p_\star(x) \log p_\theta(x) \]

\[ = - \sum_{x \in \text{Val}(X)} p_\star(x) \log p_\theta(x) + C \]

Minimizing \( KL(p^*||p_\theta) \) 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

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

\[ \Rightarrow \text{maximum-likelihood estimation minimizes the KL-divergence from the empirical data distribution to } p_\theta. \]

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