COMPSCI 688: Probabilistic Graphical Models

Lecture 5: Learning in Directed Graphical Models

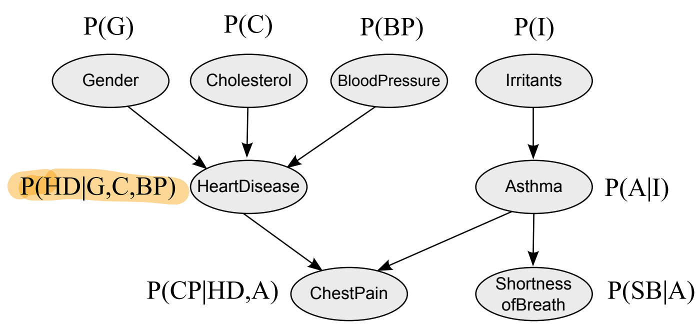
Dan Sheldon

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
⋮	⋮	⋮	⋮	⋮

— sum to one

Bayesian Networks: Parameters

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

target
name of RV

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

values of parents

value of target

$\theta = (\cdot \cdot \cdot \cdot \cdot \cdot \cdot)$

$\theta^x_{1|0,0}$

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	$\theta^{HD}_{Y\|F,L,L}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

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

error order $\frac{1}{\sqrt{n}}$

smoker cancer

$S \longrightarrow C$

smoker	cancer
1	0
0	1 *
1	1
0	0
0	0
0	0
1	0
1	1

$P(S=1) = \theta^S_1 = \frac{1}{2}$

$\theta^S_0 = \frac{1}{2}$

$P(C|S=0) \Rightarrow \theta^C_{0|0} = \frac{3}{4}$

$\theta^C_{1|0} = \frac{1}{4}$

$P(C|S=1) \Rightarrow \theta^C_{0|1} = \frac{2}{4}$

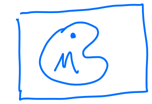
$\theta^C_{1|1} = \frac{2}{4}$

Estimation

$A \to B \to C$

Maximum-Likelihood Estimation (MLE)

$S \to C$



$\mathcal{M} =$

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

data index

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

fixed

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

function of θ

$$\frac{1}{N} \log p_\theta(x^{(1)}, \dots, x^{(N)}) = \frac{1}{N} \log \prod_n p_\theta(x^{(n)}) = \frac{1}{N} \sum_n \log p_\theta(x^{(n)})$$

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

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

(assumes independence)

parameters   data set

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

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

Example: Bernoulli Model

X

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

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

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

The log-likelihood is

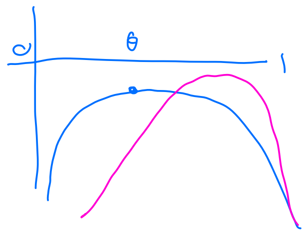
$$\begin{aligned} \mathcal{L}(\theta|x^{(1:N)}) &= \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x^{(n)}) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{I}[x^{(n)} = 0] \log(1-\theta) + \mathbb{I}[x^{(n)} = 1] \log\theta \right) \\ &= \frac{\overset{66}{\#(X=0)}}{N} \log(1-\theta) + \frac{\overset{34}{\#(X=1)}}{N} \log\theta. \end{aligned}$$

indicator $= \mathbb{I}[x=0] \cdot \log(1-\theta) + \mathbb{I}[x=1] \cdot \log\theta$

What does this likelihood function look like?

Example: Bernoulli Likelihood

Demo:
Likelihood Function



N=100
#(X=0)=66
#(X=1)=34

10% 0
90% 1
N=10000

Learning as Likelihood Maximization

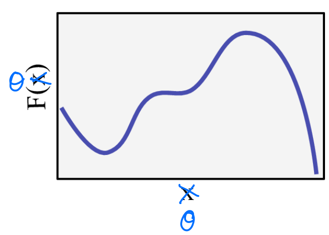
How can we find the model parameters θ that maximize the likelihood?

- The derivative of a function is zero at every local maximum

Learning as Likelihood Maximization

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

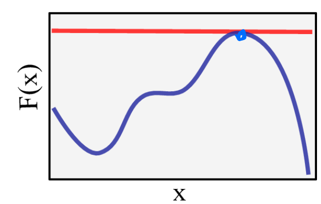
- The derivative of a function is zero at every local maximum



Learning as Likelihood Maximization

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

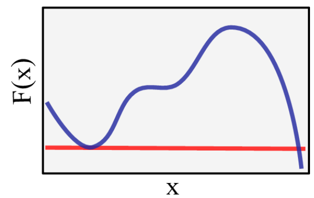
- The derivative of a function is zero at every local maximum



Learning as Likelihood Maximization

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

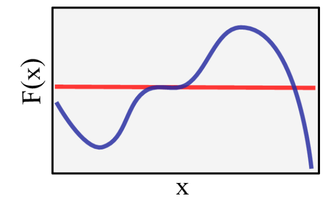
- The derivative of a function is zero at every local maximum
- Zero derivative points are not local maxima in general.



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.



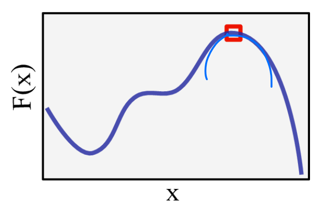
Learning as Likelihood Maximization

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

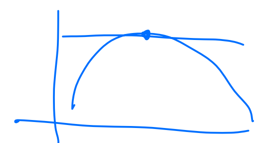
- The derivative of a function is zero at every local maximum
- Zero derivative points are not local maxima in general.
- To be a local maximum, the curvature must be negative

neg 2nd derivative

negative definite Hessian matrix



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)

concavity $\Rightarrow$ (zero gradient $\Rightarrow$ global maximum)

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:

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

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

$$\frac{\hat{p}}{\theta} - \frac{1-\hat{p}}{1-\theta} = 0 \Leftrightarrow \frac{\hat{p}}{\theta} = \frac{1-\hat{p}}{1-\theta} \Leftrightarrow \theta = \hat{p}$$

Example: Multinomial Model

$\theta_1, \dots, \theta_{V-1}$

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

$$p_\theta(x) = \begin{cases} \theta_1 & x=1 \\ \theta_2 & x=2 \\ \vdots & \\ \theta_{V-1} & x=V-1 \\ 1-\sum_{v=1}^{V-1}\theta_v & x=V \end{cases}$$

$$\mathcal{L}(\theta|x) = \log p_\theta(x) = \sum_{v=1}^{V-1}\mathbb{I}[x=v]\cdot\log\theta_v + \mathbb{I}[x=V]\cdot\log\left(1-\sum_{v=1}^{V-1}\theta_v\right)$$

$$\mathcal{L}(\theta|x^{(1:N)}) = \frac{1}{N}\sum_{n=1}^{N}\left(\quad \right)$$

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

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

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

– Quiz 2 → Fri
– HW1 → next Wed
– no class Thu

Learning Bayesian Networks

Bayesian Network Parameters

$\theta = (\, . \, . \, . \, . \, . \,)$

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.

θ^{HD} • $1, \ldots$

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	$\theta^{HD}_{Y\|F,L,L}$
⋮	⋮	⋮	⋮	⋮

sum-to-one ≈ multinomial

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

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:

$(\log p(a) + \log p(b|a))$

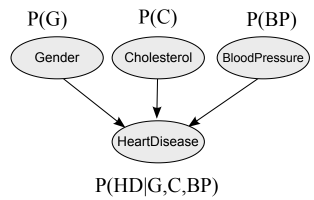
$$\mathcal{L}(\theta|\mathbf{x}^{(1:N)}) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{d=1}^{D}\log\theta^{X_d}_{x_d^{(n)}|\mathbf{x}^{(n)}_{\mathsf{pa}(d)}}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{x_d}\sum_{\mathbf{x}_{\mathsf{pa}(d)}}\mathbb{I}[x_d^{(n)} = x_d, \mathbf{x}^{(n)}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)}]\log\theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}}$$

$$= \sum_{d=1}^{D}\sum_{x_d}\sum_{x_{\mathsf{pa}(d)}}\frac{\#(X_d = x_d, \mathbf{X}_{\mathsf{pa}(d)} = \mathbf{x}_{\mathsf{pa}(d)})}{N}\log\theta^{X_d}_{x_d|\mathbf{x}_{\mathsf{pa}(d)}}$$

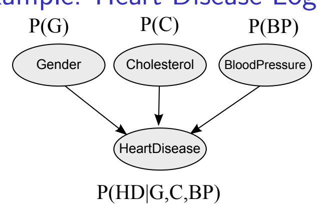
$\max_{\theta} \mathcal{L}(\theta|x^{(1:N)})$

Example: Heart Disease Joint Distribution



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

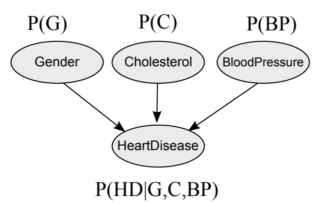
Example: Heart Disease Log Likelihood



$$\mathcal{L}(\theta|\mathbf{x}^{(1:N)}) = \sum_g \frac{\#(G=g)}{N}\log\theta^G_g + \sum_b \frac{\#(BP=b)}{N}\log\theta^{BP}_b + \sum_c \frac{\#(C=c)}{N}\log\theta^C_c$$

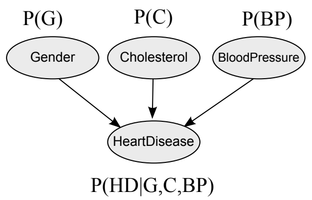
$$+ \sum_{g,b,c}\sum_h \frac{\#(HD=h, G=g, BP=b, C=c)}{N}\log\theta^{HD}_{h|g,b,c}$$

Example: Heart Disease Parameter Learning



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

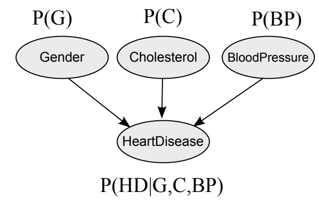
Example: Heart Disease Parameter De-Coupling



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

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 $x \in \mathrm{Val}(X)$.
- 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 $\mathbf{x}_{\mathsf{pa}(d)} \in \mathrm{Val}(\mathbf{X}_{\mathsf{pa}(d)})$:
    - For each value $x_d \in \mathrm{Val}(X_d)$. 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

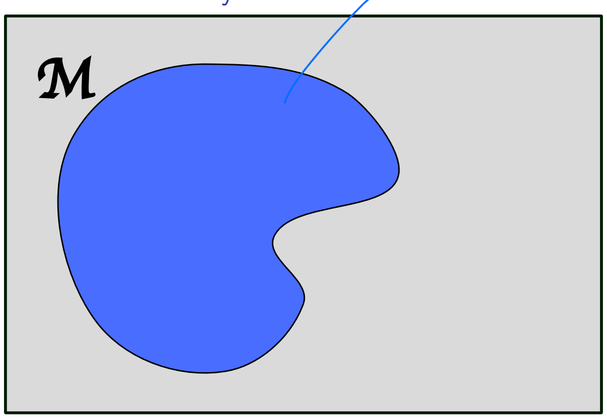
Estimation Theory

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

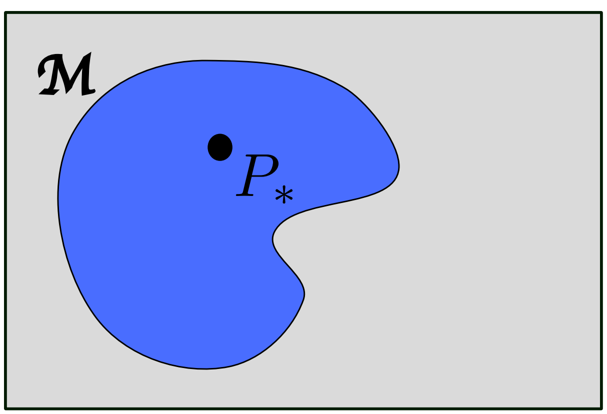
How can we select $p_{\theta^*} \in M$ that is as close as possible to p_*?

Parametric Probability Model

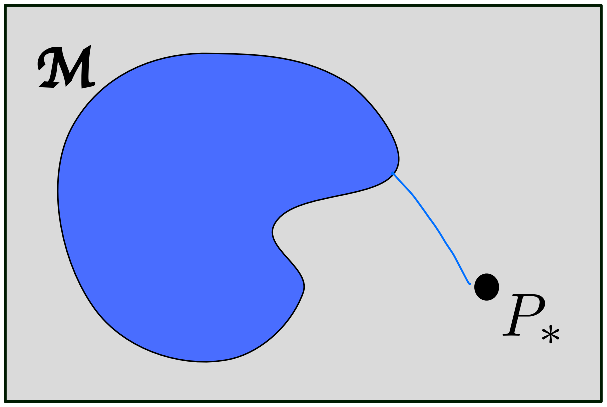
p_θ



Parameter Selection: Case 1



Parameter Selection: Case 2



Kullback-Leibler Divergence KL-divergence

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

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

"distance-like"

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

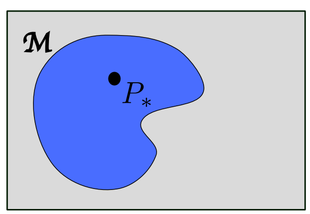
- $KL(p||q) \geq 0$ for all p and q
- $KL(p||q) = 0$ if and only if $p = q$

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

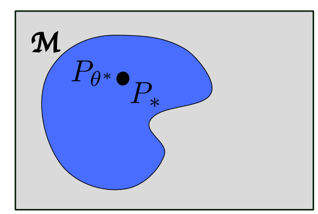
KL Divergence Minimization

- If $p_* \in M$ then there exists a θ^* such that $p_* = p_{\theta^*}$.



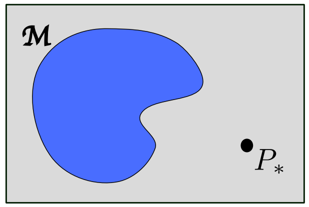
KL Divergence Minimization

- If $p_* \in M$ then there exists a θ^* such that $p_* = p_{\theta^*}$.



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.



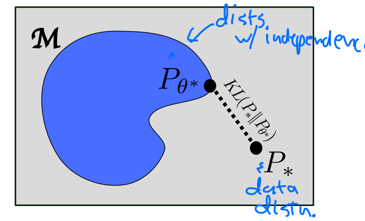
Learning Intro | Estimation | MLE Examples | Learning Bayesian Networks | Estimation Theory

KL Divergence Minimization

$p^*(a,b) =$ generic $\qquad A \not\perp B$

(a) (b) $\qquad p_\theta(a,b) = p_\theta(a) \cdot p_\theta(b) \qquad A \perp B$

- 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.



$$p_0 \xrightarrow{\text{data}} x^{(1)}, \dots, x^{(N)} \to p_* \xrightarrow{\text{MLE}} p_{\theta^*}$$

Learning Intro | Estimation | MLE Examples | Learning Bayesian Networks | Estimation Theory

KL Divergence Minimization Simplification

$$KL(p_* | p_\theta) = \sum_x p_*(x) \log\left(\frac{p_*(x)}{p_\theta(x)}\right)$$

$$= \underbrace{\sum_x p_*(x) \log p_*(x)}_{\text{no } \theta} - \sum_x p_*(x) \log p_\theta(x)$$

$$= -\sum_x p_*(x) \log p_\theta(x) + C$$

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

"log-likelihood" $\to \mathcal{L}(\theta | p_*) = \sum_{\mathbf{x} \in \mathrm{Val}(\mathbf{X})} p_*(\mathbf{x}) \log p_\theta(\mathbf{x}) = \mathbb{E}_{x \sim p_*}[\log p_\theta(x)]$

$\mathcal{L}(\theta | x^{(1:N)})$

Learning Intro | Estimation | MLE Examples | Learning Bayesian Networks | Estimation Theory

Maximum Likelihood = KL Minimization

$$p_*(x) = \begin{cases} \frac{1}{N} & \text{if } x = x^{(n)} \text{ for some } n \\ 0 & \text{ow} \end{cases}$$

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

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

$\implies$ 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.