Graphical Models

parameter learning in Bayesian networks

Siamak Ravanbakhsh

Winter 2018

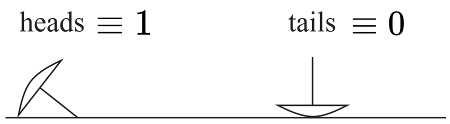
Learning objectives

- Likelihood function and MLE
- Role of the sufficient statistics
- MLE for parameter learning in directed models
  - why is it easy?
- Conjugate priors and Bayesian parameter learning

Likelihood function through an example

a thumbtack with unknown prob. of heads & tails

Bernoulli dist. $p(x;\theta) = \theta^x (1-\theta)^{(1-x)}$



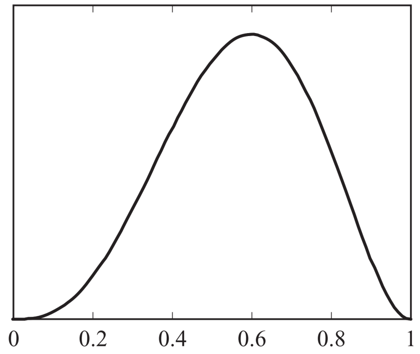
Likelihood function through an example

a thumbtack with unknown prob. of heads & tails

Bernoulli dist. $p(x;\theta) = \theta^x(1-\theta)^{(1-x)}$

IID observations $\mathcal{D} = \{1,0,0,1,1\}$

likelihood of θ is $L(\theta;\mathcal{D}) = \prod_{x\in\mathcal{D}} P(x;\theta) = \theta^3(1-\theta)^2$



likelihood function θ

not a pdf (it does not integrate to 1)

Likelihood function through an example

a thumbtack with unknown prob. of heads & tails

Bernoulli dist. $p(x;\theta) = \theta^x(1-\theta)^{(1-x)}$

IID observations $\mathcal{D} = \{1,0,0,1,1\}$

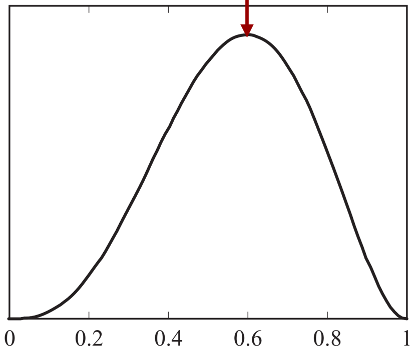
likelihood of θ is $L(\theta;\mathcal{D}) = \prod_{x\in\mathcal{D}} P(x;\theta) = \theta^3(1-\theta)^2$

log-likelihood: $\log L(\theta;\mathcal{D}) = 3\log\theta + 2\log(1-\theta)$

maximizing the log-likelihood (M-projection of $P_{\mathcal{D}}$)

$$\frac{\partial}{\partial\theta}\left(3\log\theta + 2\log(1-\theta)\right) = \frac{3}{\theta} - \frac{2}{1-\theta} = \frac{3-5\theta}{\theta(1-\theta)} = 0 \Rightarrow \hat{\theta} = \frac{3}{5}$$

max-likelihood estimate (MLE)



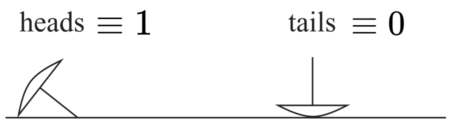
likelihood function θ

not a pdf (it does not integrate to 1)

Sufficient statistics through an example

IID observations $\mathcal{D} = \{1, 0, 0, 1, 1\}$

likelihood of θ is $L(\theta, \mathcal{D}) = \prod_{x \in \mathcal{D}} P(x; \theta) = \theta^3 (1 - \theta)^2$



all we needed to know about the data:

- number of heads and tails

given a distribution $P(x; \theta)$

- its sufficient statistics is function $\phi = [\phi_1, \ldots, \phi_K]$ such that

$$\mathbb{E}_{\mathcal{D}}[\phi(x)] = \mathbb{E}_{\mathcal{D}'}[\phi(x')] \quad \Rightarrow \quad \frac{1}{|\mathcal{D}|} L(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}'|} L(\theta, \mathcal{D}') \quad \forall \mathcal{D}, \mathcal{D}', \theta$$

sufficient statistics of the dataset is all that matters about the data

Revisiting exponential family

given a distribution $P(x;\theta)$

- its sufficient statistics is function $\phi = [\phi_1, \ldots, \phi_K]$ such that

$$\mathbb{E}_{\mathcal{D}}[\phi(x)] = \mathbb{E}_{\mathcal{D}'}[\phi(x')] \quad \Rightarrow \quad \frac{1}{|\mathcal{D}|}L(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}'|}L(\theta, \mathcal{D}') \quad \forall \mathcal{D}, \mathcal{D}', \theta$$

$L(\theta, \mathcal{D}) = \prod_{x \in \mathcal{D}} p(x;\theta)$

the (linear) exponential family: $p(x) \propto \exp(\langle \theta, \phi(x) \rangle)$

- max-entropy distribution subject to $\mathbb{E}_p[\phi(x)] = \mu$

Revisiting exponential family

given a distribution $P(x;\theta)$

- its sufficient statistics is function $\phi = [\phi_1, \ldots, \phi_K]$ such that

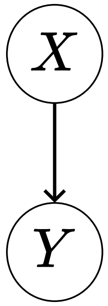
$$\mathbb{E}_{\mathcal{D}}[\phi(x)] = \mathbb{E}_{\mathcal{D}'}[\phi(x')] \quad \Rightarrow \quad \frac{1}{|\mathcal{D}|}L(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}'|}L(\theta, \mathcal{D}') \quad \forall \mathcal{D}, \mathcal{D}', \theta$$

$L(\theta, \mathcal{D}) = \prod_{x \in \mathcal{D}} p(x;\theta)$

the (linear) exponential family: $p(x) \propto \exp(\langle \theta, \phi(x) \rangle)$

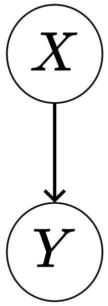
- max-entropy distribution subject to $\mathbb{E}_p[\phi(x)] = \mu$
- if $\phi_1, \ldots, \phi_k$ are linearly independent, then $\theta \leftrightarrow \mu$

MLE for Bayesian networks an example



a simple network $p(x, y; \theta) = p(x; \theta_X) p(y|x; \theta_{Y|X})$

MLE for Bayesian networks an example



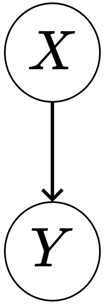
a simple network $p(x, y; \theta) = p(x; \theta_X) p(y|x; \theta_{Y|X})$

likelihood

$$L(\mathcal{D}; \theta) = \prod_{(x,y) \in \mathcal{D}} p(x; \theta_X) p(y|x; \theta_{Y|X})$$

$$= \underbrace{\left(\prod_{(x) \in \mathcal{D}} p(x; \theta_X) \right)}_{\text{likelihood of x}} \underbrace{\left(\prod_{(x,y) \in \mathcal{D}} p(y|x; \theta_{Y|X}) \right)}_{\text{cond. likelihood of y}}$$

MLE for Bayesian networks an example



a simple network $p(x, y; \theta) = p(x; \theta_X) p(y|x; \theta_{Y|X})$

likelihood $L(\mathcal{D}; \theta) = \prod_{(x,y) \in \mathcal{D}} p(x; \theta_X) p(y|x; \theta_{Y|X})$

$$= \underbrace{\left(\prod_{(x) \in \mathcal{D}} p(x; \theta_X) \right)}_{\text{likelihood of x}} \underbrace{\left(\prod_{(x,y) \in \mathcal{D}} p(y|x; \theta_{Y|X}) \right)}_{\text{cond. likelihood of y}}$$

for discrete vars.

number of times $x = \ell$ in the dataset

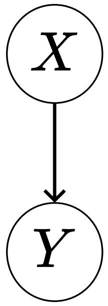
number of times $x = \ell, y = \ell'$ in the dataset

$$L(\mathcal{D}; \theta) = \left(\prod_{\ell \in Val(X)} \theta_{X,\ell}{}^{N(x=\ell)} \right) \left(\prod_{\ell, \ell' \in Val(X) \times Val(Y)} \theta_{Y|X,\ell,\ell'}{}^{N(x=\ell, y=\ell')} \right)$$

$\theta_{X,\ell} \rightarrow p(X = \ell)$

$\theta_{Y|X,\ell,\ell'} \rightarrow p(X = \ell \mid Y = \ell')$

MLE for Bayesian networks an example



a simple network $p(x, y; \theta) = p(x; \theta_X) p(y|x; \theta_{Y|X})$

likelihood $L(\mathcal{D}; \theta) = \prod_{(x,y)\in\mathcal{D}} p(x; \theta_X) p(y|x; \theta_{Y|X})$

$$= \underbrace{\left(\prod_{(x)\in\mathcal{D}} p(x; \theta_X)\right)}_{\text{likelihood of x}} \underbrace{\left(\prod_{(x,y)\in\mathcal{D}} p(y|x; \theta_{Y|X})\right)}_{\text{cond. likelihood of y}}$$

for discrete vars.

number of times $x = \ell$ in the dataset

number of times $x = \ell, y = \ell'$ in the dataset

$$L(\mathcal{D}; \theta) = \left(\prod_{\ell \in Val(X)} \theta_{X,\ell}^{N(x=\ell)}\right) \left(\prod_{\ell,\ell' \in Val(X)\times Val(Y)} \theta_{Y|X,\ell,\ell'}^{N(x=\ell, y=\ell')}\right)$$

$\theta_{X,\ell} \to p(X = \ell)$

$\theta_{Y|X,\ell,\ell'} \to p(X = \ell \mid Y = \ell')$

MLE : maximize local likelihood terms individually

$$\theta_{X,\ell} = \frac{N(x=\ell)}{|\mathcal{D}|} \quad \theta_{Y|X,\ell,\ell'} = \frac{N(x=\ell, y=\ell')}{|\mathcal{D}|}$$

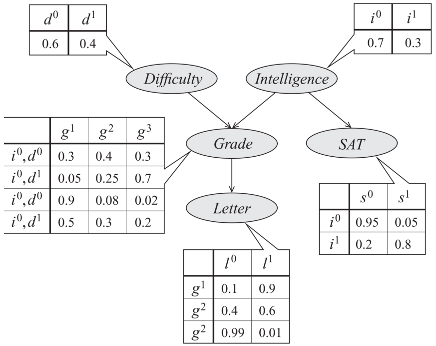
MLE for Bayesian networks general case

Bayes-net $p(x;\theta) = \prod_i p(x_i \mid Pa_{x_i}; \theta_{X_i|Pa_{X_i}})$

likelihood $L(\mathcal{D};\theta) = \prod_{x\in\mathcal{D}} \prod_i p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$$= \prod_i \underline{\prod_{(x_i, Pa_{x_i})\in\mathcal{D}} p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})}$$

local likelihood terms

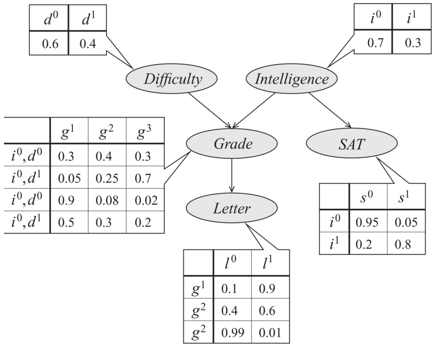


MLE for Bayesian networks general case

Bayes-net $p(x;\theta) = \prod_i p(x_i \mid Pa_{x_i}; \theta_{X_i|Pa_{X_i}})$

likelihood $L(\mathcal{D};\theta) = \prod_{x \in \mathcal{D}} \prod_i p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$$= \prod_i \underbrace{\prod_{(x_i, Pa_{x_i}) \in \mathcal{D}} p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})}_{\text{local likelihood terms}}$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^0,d^0	0.9	0.08	0.02
i^0,d^1	0.5	0.3	0.2

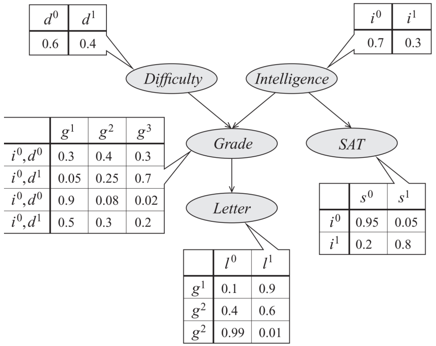
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

maximizing the conditional likelihood for each node

- similar to solving individual prediction problems

MLE for Bayesian networks general case



Bayes-net $p(x;\theta) = \prod_i p(x_i \mid Pa_{x_i}; \theta_{X_i|Pa_{X_i}})$

likelihood $L(\mathcal{D};\theta) = \prod_{x\in\mathcal{D}} \prod_i p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

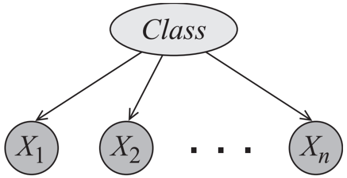
$$= \prod_i \underline{\prod_{(x_i, Pa_{x_i})\in\mathcal{D}} p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})}$$

local likelihood terms

maximizing the conditional likelihood for each node

- similar to solving individual prediction problems

Example how to learn a naive Bayes model?



Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y)$$

↓

conditional entropy $\sum_x p(x) H(p(y|x))$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

$$I(X,Y) = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

$$I(X,Y) = \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

$$= D_{KL}(p(x,y)\|p(x)p(y))$$ positive

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x \in \mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

MLE in Bayes-nets mutual information form

log-likelihood

$$\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

MLE in Bayes-nets mutual information form

log-likelihood

$$\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

using the empirical distribution

$$= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

MLE in Bayes-nets mutual information form

log-likelihood $$\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

using the empirical distribution $$= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

use MLE estimate $$\ell(\mathcal{D}, \theta^*) = N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$$

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

using the empirical distribution $= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

use MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$

using the definition of mutual information $= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \left(\log \frac{p_{\mathcal{D}}(x_i, Pa_{x_i})}{p_{\mathcal{D}}(x_i) p_{\mathcal{D}}(Pa_{x_i})} + \log p_{\mathcal{D}}(x_i)\right)$

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$

$$= \sum_i \sum_{(x_i,Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$$

using the empirical distribution $= N \sum_i \sum_{x_i,Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$

use MLE estimate $\ell(\mathcal{D},\theta^*) = N \sum_i \sum_{x_i,Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$

using the definition of mutual information $= N \sum_i \sum_{x_i,Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \left(\log \frac{p_{\mathcal{D}}(x_i,Pa_{x_i})}{p_{\mathcal{D}}(x_i)p_{\mathcal{D}}(Pa_{x_i})} + \log p_{\mathcal{D}}(x_i)\right)$

$$= N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$$

Mutual information form & structure search

Mutual information form & structure search

for MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

Mutual information form & structure search

for MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

structure learning algorithms use mutual information in the structure search:

- **Chow-Liu algorithm**: find the max-spanning **tree:**
  - edge-weights = mutual information
  - each node has at most one parent
  - add direction to edges later

Mutual information form & structure search

for MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

structure learning algorithms use mutual information in the structure search:

- **Chow-Liu algorithm**: find the max-spanning **tree:**
  - edge-weights = mutual information
  - each node has at most one parent
  - add direction to edges later
- adding more edges always increases the likelihood!
  - have to regularize the likelihood score

Bayesian parameter estimation

max-likelihood is the same $\hat{\theta} = \frac{1}{3}$ for

case 1. $N(x=1) = 1, N(x=0) = 2$

case 2. $N(x=1) = 100, N(x=0) = 200$

Example

Bayesian parameter estimation

max-likelihood is the same $\hat{\theta} = \frac{1}{3}$ for

case 1. $N(x=1) = 1, N(x=0) = 2$

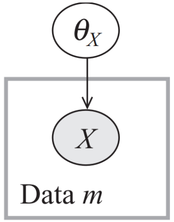
case 2. $N(x=1) = 100, N(x=0) = 200$

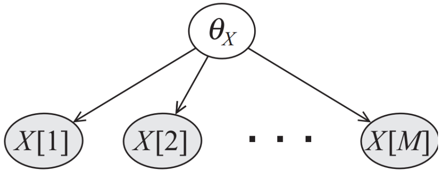
Example

need to model our uncertainty

Bayesian approach:

- define a prior $p(\theta)$
- obtain a posterior





Bayesian parameter estimation

max-likelihood is the same $\hat{\theta} = \frac{1}{3}$ for

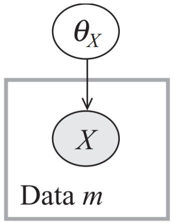
case 1. $N(x=1) = 1, N(x=0) = 2$
case 2. $N(x=1) = 100, N(x=0) = 200$

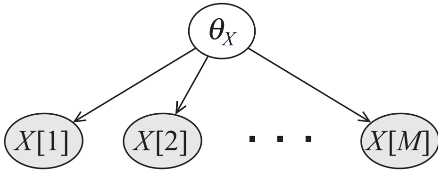
Example

need to model our uncertainty

Bayesian approach:

- define a prior $p(\theta)$
- obtain a posterior





likelihood

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \propto p(\theta)p(\mathcal{D} \mid \theta)$$

marginal likelihood

$\prod_{x \in \mathcal{D}} p(x|\theta)$

Bayesian parameter estimation

assuming a uniform prior $p(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & o.w. \end{cases}$

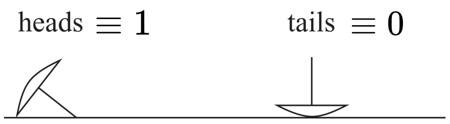
posterior $p(\theta \mid \mathcal{D}) \propto p(\theta)p(\mathcal{D} \mid \theta) \propto p(\mathcal{D} \mid \theta)$

$\propto$

(and normalize)

Bayesian parameter estimation

assuming a uniform prior $p(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & o.w. \end{cases}$



posterior $p(\theta \mid \mathcal{D}) \propto p(\theta)p(\mathcal{D} \mid \theta) \propto p(\mathcal{D} \mid \theta)$

posterior predictive: predicting heads/tails using the posterior

rather than a single MLE value

$\propto \theta^{N(1)}(1-\theta)^{N(0)} \quad \theta^{x}(1-\theta)^{(1-x)}$

in this case the posterior $\propto$ likelihood (the only difference is integration vs using the MLE)

(and normalize)

Bayesian parameter estimation

assuming a uniform prior $p(\theta) = \begin{cases} 1 & 0 \le \theta \le 1 \\ 0 & o.w. \end{cases}$

posterior $p(\theta \mid \mathcal{D}) \propto p(\theta)p(\mathcal{D} \mid \theta) \propto p(\mathcal{D} \mid \theta)$

posterior predictive: predicting heads/tails using the posterior

rather than a single MLE value

$$p(x \mid \mathcal{D}) = \int_0^1 p(\theta|\mathcal{D})p(x|\theta)\mathrm{d}\theta$$

$$\propto \theta^{N(1)}(1-\theta)^{N(0)} \quad \theta^{x}(1-\theta)^{(1-x)}$$

in this case the posterior $\propto$ likelihood (the only difference is integration vs using the MLE)

if we do the integration above: $p(x=1 \mid \mathcal{D}) = \frac{N(1)+1}{N(0)+N(1)+2}$ Laplace correction

(and normalize)

Bayesian parameter estimation

assuming a uniform prior $p(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & o.w. \end{cases}$

posterior $p(\theta \mid \mathcal{D}) \propto p(\theta)p(\mathcal{D} \mid \theta) \propto p(\mathcal{D} \mid \theta)$

posterior predictive: predicting heads/tails using the posterior

rather than a single MLE value

$$p(x \mid \mathcal{D}) = \int_0^1 p(\theta|\mathcal{D})p(x|\theta)\mathrm{d}\theta$$

$\propto \theta^{N(1)}(1-\theta)^{N(0)}$ $\theta^x(1-\theta)^{(1-x)}$

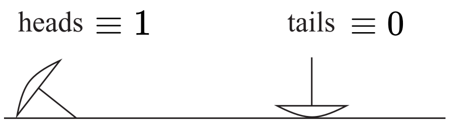
in this case the posterior $\propto$ likelihood (the only difference is integration vs using the MLE)

if we do the integration above: $p(x=1 \mid \mathcal{D}) = \frac{N(1)+1}{N(0)+N(1)+2}$ Laplace correction

(and normalize)

compare with prediction using MLE $p(x=1 \mid \mathcal{D}) = \frac{N(1)}{N(0)+N(1)}$

Conjugate priors



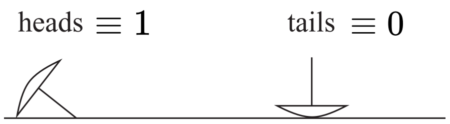
how about non-uniform priors? E.g., more likely to see heads

need an efficient way to get the posterior $p(\theta \mid \mathcal{D}) \propto p(\theta)p(\mathcal{D} \mid \theta)$

ideally the prior $p(\theta)$ & the posterior $p(\theta|\mathcal{D})$ should have the same form

$p(\theta)$ is a conjugate prior to the likelihood $p(\mathcal{D}|\theta)$

Conjugate priors



how about non-uniform priors? E.g., more likely to see heads

need an efficient way to get the posterior $p(\theta \mid \mathcal{D}) \propto p(\theta) p(\mathcal{D} \mid \theta)$

ideally the prior $p(\theta)$ & the posterior $p(\theta|\mathcal{D})$ should have the same form

$p(\theta)$ is a conjugate prior to the likelihood $p(\mathcal{D}|\theta)$

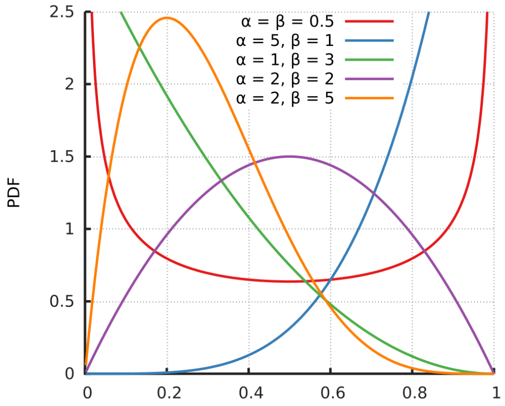
conjugate prior to the Bernoulli likelihood is the Beta distribution

$$p(\mathcal{D}|\theta) \propto \theta^{N(1)}(1-\theta)^{N(0)}$$

$$p(\theta; \alpha, \beta) = \gamma \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\gamma = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Conjugate priors: Beta-Bernoulli



conjugate prior to the Bernoulli likelihood is the Beta distribution

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

extension of factorial function to reals $\Gamma(n+1) = n!$

hyper-parameters: can be interpreted as # **imaginary** heads & tails

$$p(x=1 \mid \mathcal{D}=\emptyset) = \int_\theta p(x=1 \mid \theta)p(\theta;\alpha,\beta)\mathrm{d}\theta = \frac{\alpha}{\alpha+\beta}$$

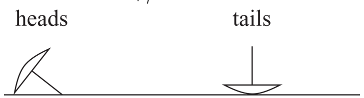
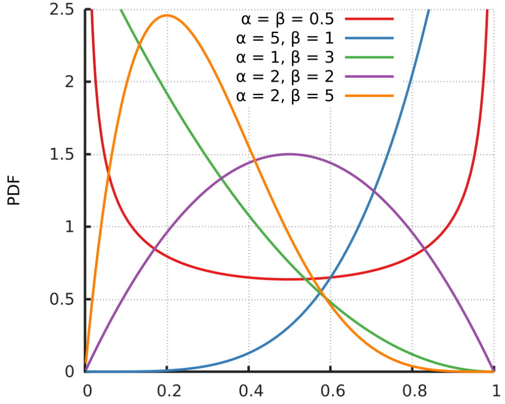


image: wikipedia

Conjugate priors: Beta-Bernoulli



conjugate prior to the Bernoulli likelihood is the Beta distribution

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

extension of factorial function to reals $\Gamma(n+1) = n!$

hyper-parameters: can be interpreted as # **imaginary** heads & tails

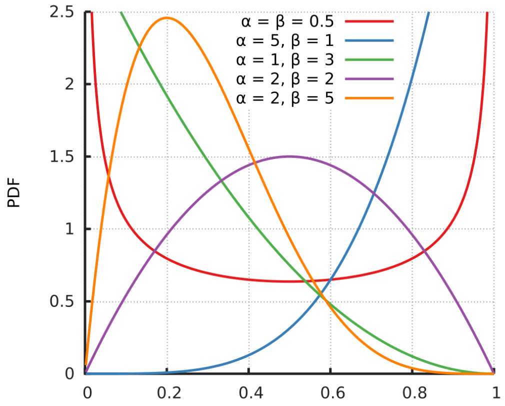
$$p(x=1 \mid \mathcal{D}=\emptyset) = \int_\theta p(x=1 \mid \theta) p(\theta; \alpha, \beta) \mathrm{d}\theta = \frac{\alpha}{\alpha+\beta}$$

posterior: $p(\theta \mid \mathcal{D}) \propto p(\theta) P(\mathcal{D} \mid \theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N(1)}(1-\theta)^{N(0)} = \theta^{\alpha-1+N(1)}(1-\theta)^{\beta-1+N(0)}$

if the prior is $p(\theta; \alpha, \beta)$, the posterior is $p(\theta; \alpha + N(1), \beta + N(0))$

image: wikipedia

Conjugate priors: Beta-Bernoulli



conjugate prior to the Bernoulli likelihood is the Beta distribution

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

extension of factorial function to reals $\Gamma(n+1) = n!$

hyper-parameters: can be interpreted as # **imaginary** heads & tails

$$p(x=1 \mid \mathcal{D}=\emptyset) = \int_\theta p(x=1 \mid \theta) p(\theta; \alpha, \beta) \mathrm{d}\theta = \frac{\alpha}{\alpha+\beta}$$

posterior: $p(\theta \mid \mathcal{D}) \propto p(\theta) P(\mathcal{D} \mid \theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \theta^{N(1)}(1-\theta)^{N(0)} = \theta^{\alpha-1+N(1)}(1-\theta)^{\beta-1+N(0)}$

if the prior is $p(\theta; \alpha, \beta)$, the posterior is $p(\theta; \alpha+N(1), \beta+N(0))$

marginal likelihood: $p(\mathcal{D}) = \int_\theta p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\theta|\mathcal{D})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\frac{\Gamma(\alpha+\beta+N(1)+N(0))}{\Gamma(\alpha+N(1))\Gamma(\beta)+N(0)}}$

image: wikipedia

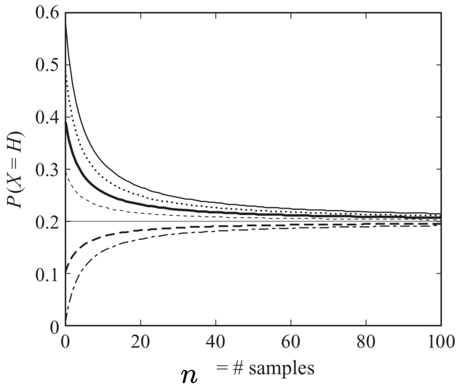
Conjugate priors: Beta-Bernoulli

posterior:

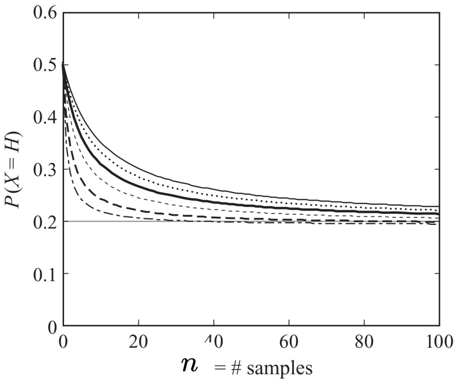
- for each n the dataset is balanced

$p(x = 1) = .2$
ground truth

different prior means $\frac{\alpha}{\alpha+\beta}$



different prior strength $\alpha + \beta$



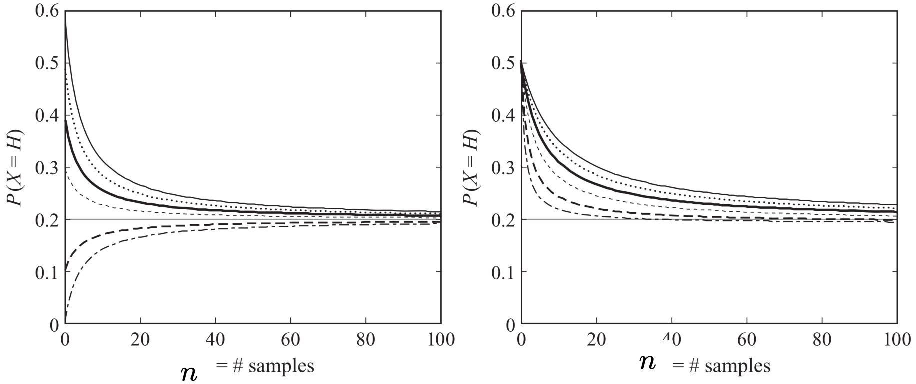
Conjugate priors: Beta-Bernoulli

posterior:

- for each n the dataset is balanced

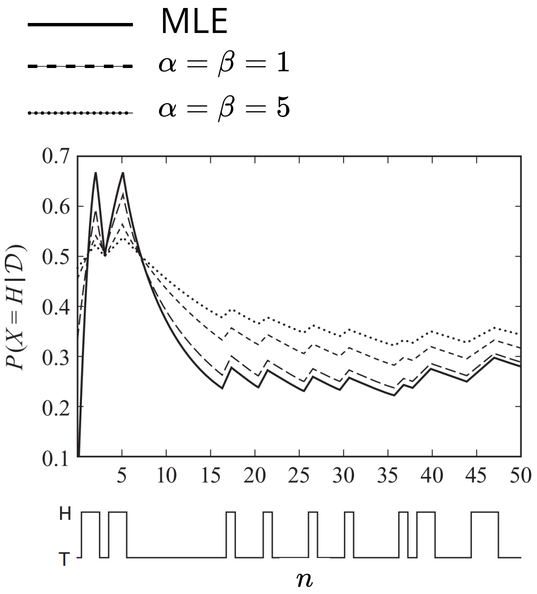
$p(x=1) = .2$
ground truth

different prior means $\frac{\alpha}{\alpha+\beta}$ different prior strength $\alpha+\beta$

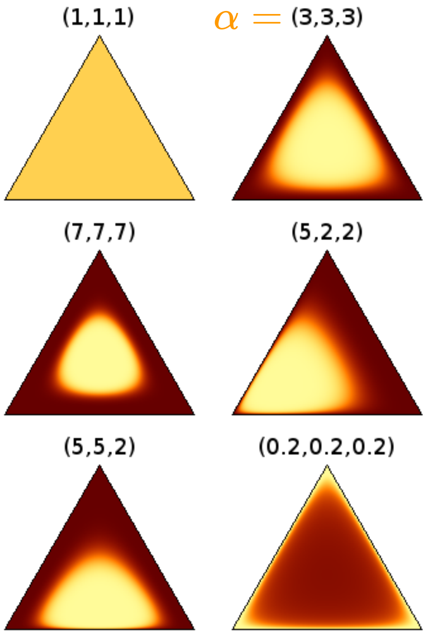


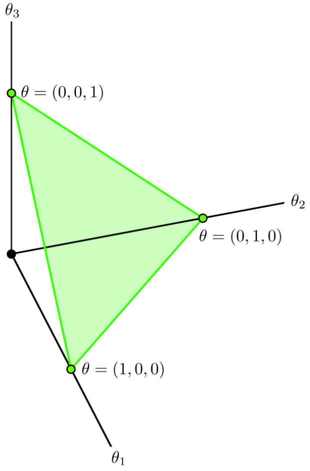
posterior predictive:

- online setting



Conjugate priors: Dirichlet-categorical





prior: $p(\theta;\alpha)$

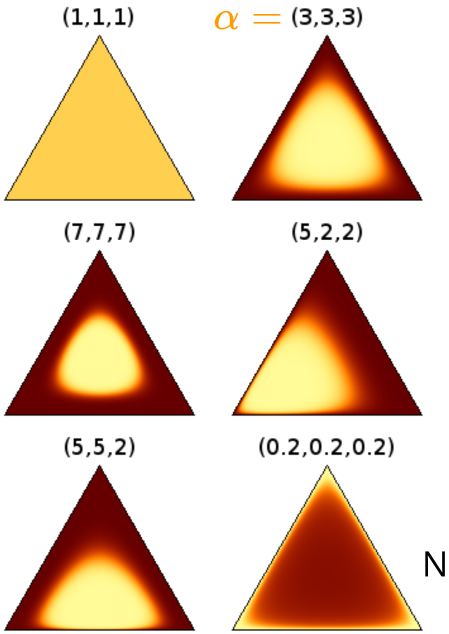
Bernoulli $\Rightarrow$ Beta

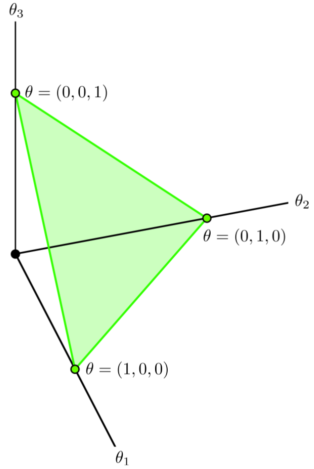
Categorical $\Rightarrow$ Dirichlet

$$p(\theta;\alpha) = \frac{\Gamma(\sum_d \alpha_d)}{\prod_d \Gamma(\alpha_d)} \prod_d \theta_d^{\alpha_d - 1}$$

$\alpha \in \Re^D$ *pseudo-counts for different categories*

Conjugate priors: Dirichlet-categorical





Bernoulli $\Rightarrow$ Beta

Categorical $\Rightarrow$ Dirichlet

$$p(\theta; \alpha) = \frac{\Gamma(\sum_d \alpha_d)}{\prod_d \Gamma(\alpha_d)} \prod_d \theta_d^{\alpha_d - 1}$$

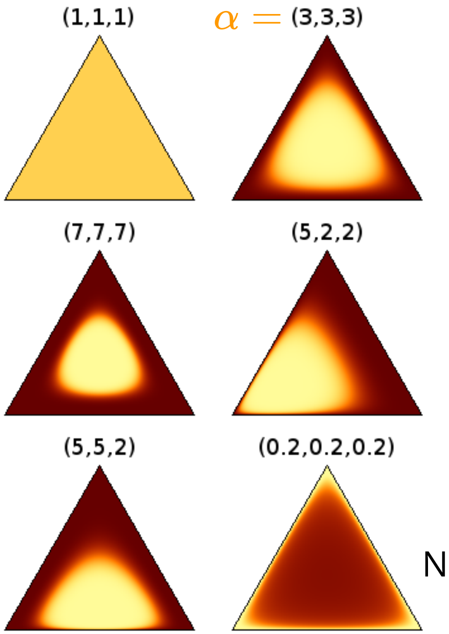
$\alpha \in \Re^D$ *pseudo-counts for different categories*

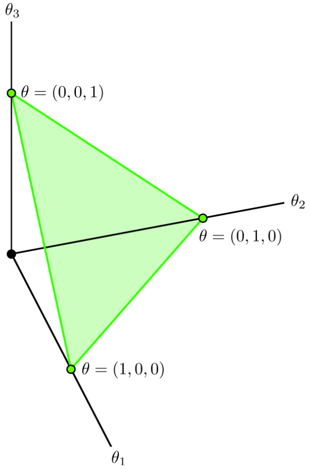
prior: $p(\theta; \alpha)$

N observations: $p(\mathcal{D} \mid \theta) \propto \prod_{x \in \mathcal{D}} \prod_d \theta_d^{\mathbb{I}(x=d)} = \prod_d \theta_d^{N(d)}$

posterior: $p(\theta \mid \mathcal{D}) \propto p(\theta) p(\mathcal{D} \mid \theta) \propto \prod_d \theta_d^{N(d)} \theta_d^{\alpha_d - 1} = \prod_d \theta_d^{\alpha_d + N(d) - 1}$

Conjugate priors: Dirichlet-categorical





Bernoulli $\Rightarrow$ Beta

Categorical $\Rightarrow$ Dirichlet

$$p(\theta;\alpha) = \frac{\Gamma(\sum_d \alpha_d)}{\prod_d \Gamma(\alpha_d)} \prod_d \theta_d^{\alpha_d - 1}$$

$\alpha \in \Re^D$ *pseudo-counts for different categories*

prior: $p(\theta;\alpha)$

N observations: $p(\mathcal{D} \mid \theta) \propto \prod_{x \in \mathcal{D}} \prod_d \theta_d^{\mathbb{I}(x=d)} = \prod_d \theta_d^{N(d)}$

posterior: $p(\theta \mid \mathcal{D}) \propto p(\theta) p(\mathcal{D} \mid \theta) \propto \prod_d \theta_d^{N(d)} \theta_d^{\alpha_d - 1} = \prod_d \theta_d^{\alpha_d + N(d) - 1}$

posterior predictive: $p(x = \bar{x} \mid \mathcal{D}) = \int_\theta p(\theta \mid \mathcal{D}) p(x = \bar{x} \mid \theta) \mathrm{d}\theta = \frac{\alpha_{\bar{x}} + N(\bar{x})}{N + \sum_d \alpha_d}$

Conjugate priors: exponential family

for the likelihood function: $p(x \mid \theta) = \exp(\langle \phi(x), \theta \rangle - A(\theta))$

Conjugate priors: exponential family

for the likelihood function: $p(x \mid \theta) = \exp(\langle \phi(x), \theta \rangle - A(\theta))$

suppose we observe N instances $p(\mathcal{D} \mid \theta) = \exp(\langle \sum_{x \in \mathcal{D}} \phi(x), \theta \rangle - N A(\theta))$

Conjugate priors: exponential family

for the likelihood function: $p(x \mid \theta) = \exp(\langle \phi(x), \theta \rangle - A(\theta))$

suppose we observe N instances $p(\mathcal{D} \mid \theta) = \exp(\langle \sum_{x \in \mathcal{D}} \phi(x), \theta \rangle - NA(\theta))$

conjugate prior $p(\theta; \eta, \nu) = \exp(\langle \nu\eta, \theta \rangle - \nu A(\theta))$

Conjugate priors: exponential family

for the likelihood function: $p(x \mid \theta) = \exp(\langle \phi(x), \theta \rangle - A(\theta))$

suppose we observe N instances $p(\mathcal{D} \mid \theta) = \exp(\langle \sum_{x \in \mathcal{D}} \phi(x), \theta \rangle - N A(\theta))$

conjugate prior $p(\theta; \eta, \nu) = \exp(\langle \nu\eta, \theta \rangle - \nu A(\theta))$

imaginary expected sufficient statistics

imaginary counts

posterior: $p(\theta \mid \mathcal{D}; \eta, \nu) = \exp\left(\langle \nu\eta + \sum_{x \in \mathcal{D}} \phi(x), \theta \rangle - (\nu + N) A(\theta)\right)$

Bayesian learning for Bayes-nets

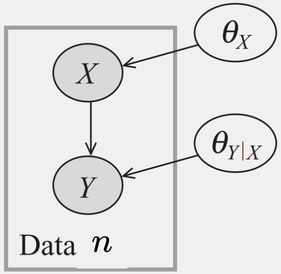
assumption

- global parameter independence: prior decomposes $p(\theta) = \prod_i p(\theta_{X_i|Pa_{X_i}})$

conclusion

- posterior is also decomposes $p(\theta \mid \mathcal{D}) = \prod_i p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})$

example



$\theta_X \perp \theta_{Y|X} \mid X, Y$

Bayesian learning for Bayes-nets

assumption

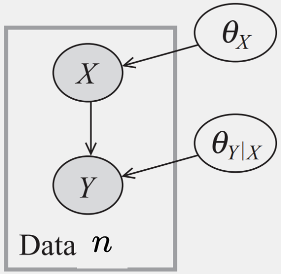
- global parameter independence: prior decomposes $p(\theta) = \prod_i p(\theta_{X_i|Pa_{X_i}})$

conclusion

- posterior is also decomposes $p(\theta \mid \mathcal{D}) = \prod_i p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})$

$$p(\theta \mid \mathcal{D}) = \underbrace{\prod_i p(\theta_{X_i|Pa_{X_i}})}_{\text{prior}} \underbrace{\prod_{x \in \mathcal{D}} \prod_i p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{\text{likelihood}}$$

example



$\theta_X \perp \theta_{Y|X} \mid X, Y$

Bayesian learning for Bayes-nets

assumption

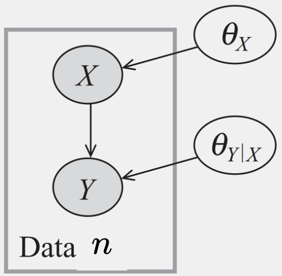
- global parameter independence: prior decomposes $p(\theta) = \prod_i p(\theta_{X_i|Pa_{X_i}})$

conclusion

- posterior is also decomposes $p(\theta \mid \mathcal{D}) = \prod_i p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})$

$$p(\theta \mid \mathcal{D}) = \underbrace{\prod_i p(\theta_{X_i|Pa_{X_i}})}_{\text{prior}} \underbrace{\prod_{x \in \mathcal{D}} \prod_i p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{\text{likelihood}} = \prod_i \underbrace{p(\theta_{X_i|Pa_{X_i}}) \prod_{x \in \mathcal{D}} p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})}$$

example



$\theta_X \perp \theta_{Y|X} \mid X, Y$

Bayesian learning for Bayes-nets

assumption

- global parameter independence: prior decomposes $p(\theta) = \prod_i p(\theta_{X_i|Pa_{X_i}})$

conclusion

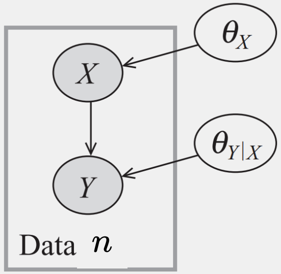
- posterior is also decomposes $p(\theta \mid \mathcal{D}) = \prod_i p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})$

$$p(\theta \mid \mathcal{D}) = \underbrace{\prod_i p(\theta_{X_i|Pa_{X_i}})}_{\text{prior}} \underbrace{\prod_{x \in \mathcal{D}} \prod_i p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{\text{likelihood}} = \prod_i \underbrace{p(\theta_{X_i|Pa_{X_i}}) \prod_{x \in \mathcal{D}} p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})}$$

individual posteriors

- we can apply Bayesian learning to individual conditional distributions

example



$\theta_X \perp \theta_{Y|X} \mid X, Y$

Bayesian learning for Bayes-nets

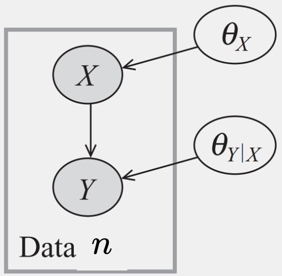
assumption

- global parameter independence: prior decomposes $p(\theta) = \prod_i p(\theta_{X_i|Pa_{X_i}})$

conclusion

- posterior is also decomposes $p(\theta \mid \mathcal{D}) = \prod_i p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})$

example



$\theta_X \perp \theta_{Y|X} \mid X, Y$

$$p(\theta \mid \mathcal{D}) = \underbrace{\prod_i p(\theta_{X_i|Pa_{X_i}})}_{\text{prior}} \underbrace{\prod_{x \in \mathcal{D}} \prod_i p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{\text{likelihood}} = \prod_i \underbrace{p(\theta_{X_i|Pa_{X_i}}) \prod_{x \in \mathcal{D}} p(x_i|Pa_{x_i}; \theta_{X_i|Pa_{X_i}})}_{p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D})}$$

individual posteriors

- we can apply Bayesian learning to individual conditional distributions
- posterior predictive also decomposes: $p(x' \mid \mathcal{D}) = \prod_i p(x'_i \mid \mathcal{D})$

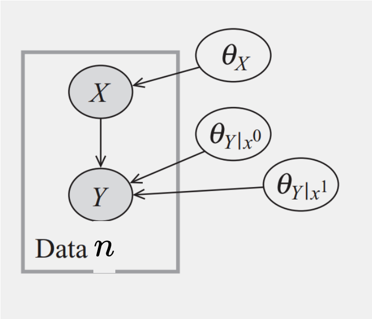
$$\downarrow$$

$$\int_\theta p(\theta_{X_i|Pa_{X_i}} \mid \mathcal{D}) p(x'_i|Pa_{x'_i}; \theta_{X_i|Pa_{X_i}}) \mathrm{d}\theta_{X_i|Pa_{X_i}}$$

Bayesian learning for Bayes-nets

discrete case: conditional probability tables (CPTs)

we can further decompose the prior & posterior



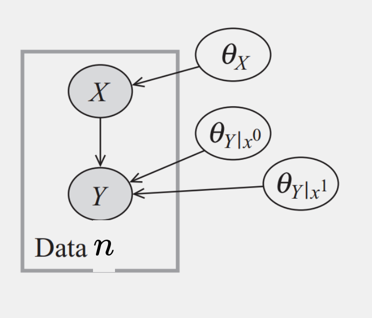
local parameter independence

- assume a decomposed prior $p(\theta_{Y|X}) = p(\theta_{Y|x^0})p(\theta_{Y|x^1})$ for binary X
  - one for each assignment to the parent *(e.g., cols of the table)*

Bayesian learning for Bayes-nets

discrete case: conditional probability tables (CPTs)

we can further decompose the prior & posterior



local parameter independence

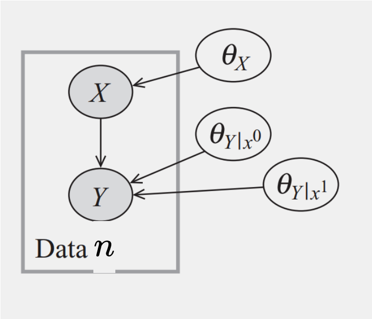
- assume a decomposed prior $p(\theta_{Y|X}) = p(\theta_{Y|x^0})p(\theta_{Y|x^1})$ for binary X
  - one for each assignment to the parent *(e.g., cols of the table)*
- posterior is also decomposed $p(\theta_{Y|X} \mid \mathcal{D}) = p(\theta_{Y|x^0} \mid \mathcal{D})p(\theta_{Y|x^1} \mid \mathcal{D})$

$$p(\theta_{Y|x^0}) \prod_{(x^0,y)\in\mathcal{D}} p(y|x^0; \theta_{Y|x^0})$$

Bayesian learning for Bayes-nets

discrete case: conditional probability tables (CPTs)

we can further decompose the prior & posterior



local parameter independence

- assume a decomposed prior $p(\theta_{Y|X}) = p(\theta_{Y|x^0})p(\theta_{Y|x^1})$ for binary X
  - one for each assignment to the parent *(e.g., cols of the table)*
- posterior is also decomposed $p(\theta_{Y|X} \mid \mathcal{D}) = p(\theta_{Y|x^0} \mid \mathcal{D})p(\theta_{Y|x^1} \mid \mathcal{D})$

$$p(\theta_{Y|x^0}) \prod_{(x^0,y)\in\mathcal{D}} p(y|x^0;\theta_{Y|x^0})$$

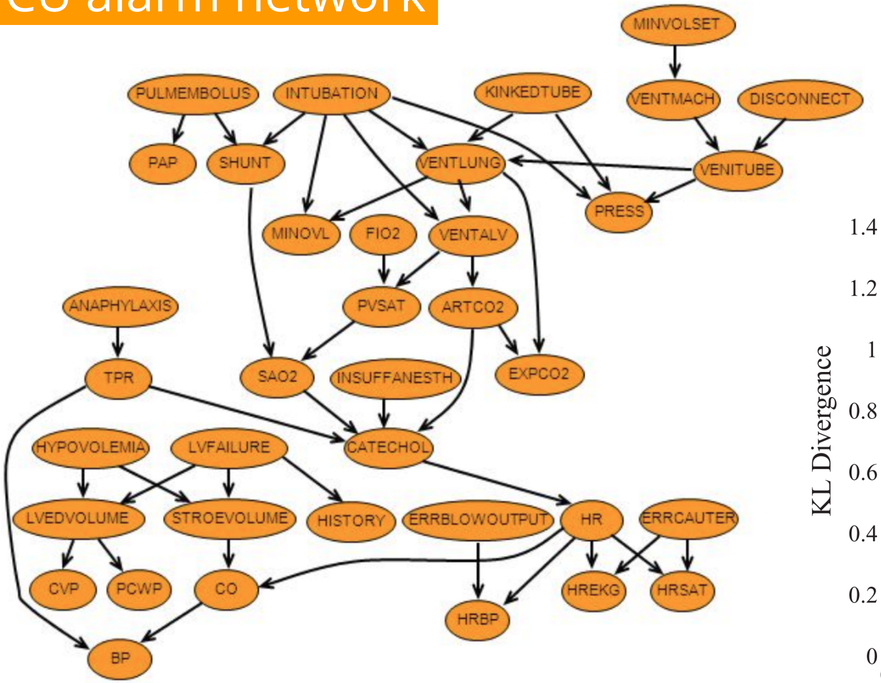
How do I use this?

- keep a vector of psudo-counts for each node $\alpha_{Y|x^0}, \alpha_{Y|x^1}$
  - E.g., K2 prior $\alpha_{Y|x^0} = \alpha_{Y|x^1} = [1,\dots,1]$ (similar to Laplace smoothing)
- after observing N samples: update these based on the frequency of different (x,y) values

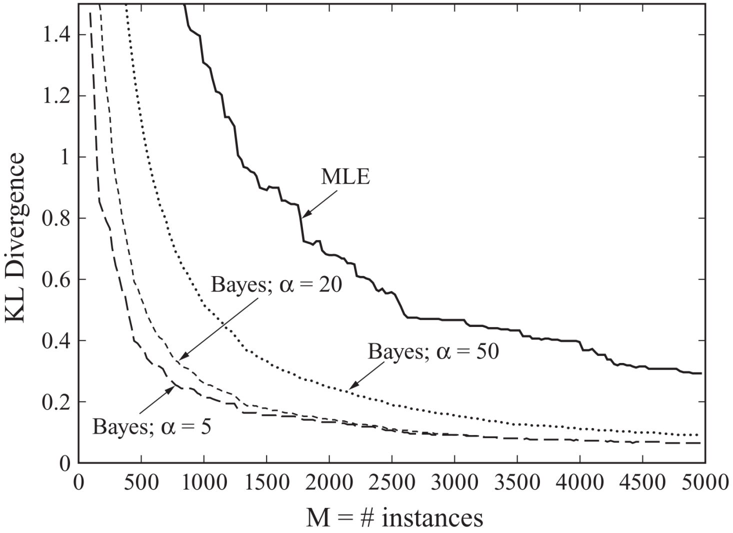
Bayesian learning for Bayes-nets

example

ICU alarm network



Bayesian learning vs MLE



Summary

learn the parameter by maximizing the likelihood

it does not reflect uncertainty:

- maintain a distribution over the parameters
- for conjugate pairs (prior-likelihood), this maintenance is easy

In Bayes-nets:

- both MLE and Bayesian learning is easy
  - they have a decomposed form