Parameter Estimation for Bayesian Networks

Sargur Srihari srihari@cedar.buffalo.edu

Topics

- Problem Statement
- Maximum Likelihood Approach
 - Thumb-tack (Bernoulli), Multinomial, Gaussian
 - Application to Bayesian Networks
- Bayesian Approach
 - Thumbtack vs Coin Toss
 - Uniform Prior vs Beta Prior
 - Multinomial
 - Dirichlet Prior

Problem Statement

- BN structure is fixed
- Data set consists of M fully observed instances of the network variables

$$\mathcal{D} = \{\xi[1], ..., \xi[M]\}$$

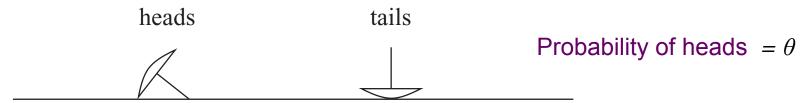
- Arises commonly in practice
 - Since hard to elicit from human experts
- Forms building block for structure learning and learning from incomplete data

Two Main Approaches

- 1. Maximum Likelihood Estimation
- 2. Bayesian
- Discuss general principles of both
- Start with simplest learning problem
 - BN with a single random variable
 - Generalize to arbitrary network structures

Thumbtack Example

- Simple Problem contains interesting issues to be tackled:
 - Thumbtack is tossed to land head/tail



- Tossed several times obtaining heads or tails
- Based on data set wish to estimate probability that next flip will land heads/tails
 - It is controlled by an unknown parameter: the Probability of heads θ

Results of Tosses

- Toss M=100 times
 - We get 35 heads
- What is our estimate for θ ?
 - Intuition suggests 0.35
 - If it was 0.1 chances of 35/100 would be lower
- Law of larger number says
 - as no of tosses grows it is increasingly unlikely that fraction will be far from θ
 - For large M, fraction of heads observed is a good estimate with high probability
- How to formalize this intuition?

Maximum Likelihood Estimator

function for the sequence

HTTHH

0.2

 $\hat{\theta} = 0.6$

0.6

0.8

it is > 0.5

• Since tosses are independent, probability of

sequence is

$$P(H,T,T,H,H:\theta) = \theta (1-\theta)(1-\theta) \theta\theta = \theta^3(1-\theta)^2$$

- Probability depends on value of θ
- Define the likelihood function to be

$$L(\theta : < H, T, T, H, H >) = P(H, T, T, H, H : \theta) = \theta^{3} (1 - \theta)^{2}$$

- Parameter values with higher likelihood are more likely to generate the observed sequences
- Thus likelihood is measure of parameter quality

MLE for General Case

• If M[1] = no. of heads among M tosses and

$$M[0] = no. of tails$$

Likelihood function is

$$L(\theta:D) = \theta^{M[1]} (1-\theta)^{M[0]}$$

Log-likelihood is

$$l(\theta:D) = M[1] \log \theta + M[0] \log(1-\theta)$$

Differentiating and setting equal to zero

$$\hat{\theta} = \frac{M[1]}{M[1] + M[0]}$$

Likelihood can be expressed as counts M[1], M[0]No other aspects of data needed: They are called Sufficient Statistics

Limitations of Max Likelihood

- If we get 3 heads out of 10 tosses, $\hat{\theta} = 0.3$
- Same estimate if 300 heads out of 1,000 tosses
 - Should be more confident with second estimate
- Statistical estimation theory deals with Confidence Intervals
 - E.g., in election polls 61 + 2 percent plan to vote for a certain candidate
 - MLE estimate lies within 0.2 of true parameter with high probability

Maximum Likelihood Principle

- Observe samples from unknown distribution $P^*(\chi)$
- Training sample D consists of M instances of $\chi: \{\xi[1],...,\xi[M]\}$
- For a parametric model $P(\xi;\theta)$, we wish to estimate parameters θ
- For each choice of parameter θ , $P(\xi;\theta)$ is a legal distribution $\sum_{\xi} P(\xi;\theta) = 1$
- We need to define parameter space Θ which is the set of allowable parameters

Common Parameter Spaces

- Thumb-tack
- X takes one of 2 values H, T
- Model
- $P(x:\theta) = \int \theta \text{ if } x = H$ $1-\theta \text{ if } x = T$
- Parameter Space

$$\Theta_{thumbtack} = [0, 1]$$
 which is a probability

- Multinomial
- X takes one of K values $x^{I}...x^{K}$
- Model
- $P(x:\mathbf{\theta}) = \theta_k$ if $x = x^k$
- Parameter Space
- $\Theta_{multinomial} = \{ \mathbf{\theta} \in [0,1]^K : \sum \theta_i = 1 \}$ a set of K probabilities

- Gaussian
- X is continuous on real line
- Model
- $P(x:\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Parameter Space
- $\Theta_{Gaussian} = R \times R^+$
- R: real value of μ
- R^+ : positive value σ

Likelihood Function

$$L(\theta:D) = \prod_{m} P(\xi[m]:\theta)$$

Sufficient Statistics Definition

• A function τ from instances χ to R^l (for some l) is a sufficient statistic if for any two data sets D and D' and any $\theta \in \Theta$ we have that

$$\sum_{\xi[M]\in D}\tau\big(\xi[M]\big)=\sum_{\xi'[M]\in D'}\tau\big(\xi'[M]\big) \Longrightarrow L(\theta:D)=L(\theta:D')$$

• We refer to $\sum_{\xi[M]\in D} \tau(\xi[M])$ as the sufficient statistic of data set D

Sufficient Statistics Examples

- In thumbtack
 - Likelihood is in terms of M[0] and M[1] which are counts of 0s and 1s, $L(\theta:\mathcal{D}) = \theta^{M[1]}(1-\theta)^{M(0)}$
 - No need for other aspects of training data e.g., toss order
 - They are summaries of data for computing likelihood
- For multinomial (with K values instead of 2)
 - Vector of counts M[1],..M[K] which are counts of x^k
 - Number of times x^k appears in training data
 - Sum of instance level statistics
 - obtained using $\tau(x^k)=0,...0,1,0...0$
 - Likelihood function is

$$\begin{cases} 1 \text{-} of\text{-}K \text{ representation} \\ \sum_{i=1}^{N} X_i \text{ yields } K\text{-}dim \text{ vector} \end{cases}$$

$$L(\theta:D) = \prod_{k} \theta_{k}^{M[k]}$$
 13

Sufficient Statistics: Gaussian

From Gaussian model $\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ by expanding $(x-\mu)^2$ we can rewrite the model as

$$P_{Gaussian}(x:\mu,\sigma) = e^{-x^2 \frac{1}{2\sigma^2} + x \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi) - \log\sigma}$$

Likelihood function will involve $\sum x_i^2$, $\sum x_i$ and $\sum I$

Sufficient statistic is the tuple

$$s_{Gaussian}(x) = \langle 1, x, x^2 \rangle$$

I does not depend on value of data item, but serves to count number of sample

Maximum Likelihood Estimates

• Given data set \mathcal{D} choose parameters that satisfy $L(\hat{\theta}:D) = \max_{\theta \in \Theta} L(\theta:D)$

Multinomial

$$\hat{\theta}_k = \frac{M[k]}{M}$$

Gaussian

$$\hat{\mu} = \frac{1}{M} \sum_{m} x[m]$$

$$\hat{\sigma} = \sqrt{\frac{1}{M} \sum_{m} (x[m] - \hat{\mu})^2}$$

MLE for Bayesian Networks

- Structure of Bayesian network allows us to reduce parameter estimation problem into a set of unrelated problems
- Each can be addressed using methods described earlier
- To clarify intuition consider a simple BN and then generalize to more complex networks

Simple Bayesian Network

- Network with two binary variables $X \rightarrow Y$
- Parameter vector θ defines CPD parameters
- Each training sample is a tuple $\langle x[m], y[m] \rangle$

$$L(\theta:D) = \prod_{m=1}^{M} P(x[m], y[m]:\theta)$$

Network model specifies P(X,Y)=P(X)P(Y\X)

$$L(\theta:D) = \prod_{m=1}^{M} P(x[m]:\theta) P(y | [m] | x[m]:\theta)$$

changing order of multiplication

$$L(\theta:D) = \prod_{m=1}^{M} P(x[m]:\theta) \prod_{m=1}^{M} P(y | [m] | x[m]:\theta)$$

Likelihood decomposes into two separate terms, each a local likelihood function Measures how well variable is predicted given its parents

Global Likelihood Decomposition

- A similar decomposition is obtained in general BNs
- Leads to the conclusion:
- We can maximize each local likelihood function independently of the rest of the network, and then combine the solutions to get an MLE solution

Table CPDs

- Simplest parameterization of a CPD is a table
- Suppose we have a variable X with parents U
- If we represent that CPD P(X|U) as a table then we will have a parameters $\theta_{x|u}$ for each combination of $x \in Val(X)$ and $u \in Val(U)$
- M[u,x] is no.of times $\xi[m]=x$ and u[m]=u in \mathcal{D}
- The MLE parameters are

$$\hat{\theta}_{x|u} = \frac{M[u,x]}{M[u]}$$
 where $M[u] = \sum_{x} M[u,x]$

Challenge with BN parameter estimation

- No. of data points used to estimate parameter $\hat{\theta}_{x|u}$ is M[u]
 - estimated from samples with parent value u
- Data points that do not agree with the parent assignment u play no role in this computation
- ullet As the no. of parents U grows, no of parent assignments grows exponentially
 - Therefore no. of data instances that we expect to have for a single parent shrinks exponentially

Data Fragmentation problem

- Data set is partitioned into a large number of small subsets
- When we have a very small no. of data instances from which to estimate parameters
 - Yield noisy estimates leading to over-fitting
 - Likely to get a large no. of zeros leading to poor performance
- Key limiting factor of learning BNs from data

Bayesian Parameter Estimation

- Thumbtack example
 - Toss tack and get 3 heads out of 10
 - Conclude that parameter θ is set to 0.3
- Coin toss example
 - Get 3 heads out of 10, Can we conclude $\theta = 0.3$?
 - No. We have lot more experience and have prior knowledge about their behavior
 - If we observe 1 million tosses and 300,000 came out heads we conclude a trick (unfair) coin with $\theta = 0.3$
- MLE cannot distinguish between
 - 1.Thumbtack and coin (coin is fairer than thumb-tack)
 - 2.Between 10 tosses and 106 tosses

Joint Probabilistic Model

- Encode prior knowledge about θ with a probability distribution
- Represents how likely we are about different choices of parameters
- Create a joint distribution θ involving parameter and samples X[1],... X[M]
- Joint distribution captures our assumptions about experiment

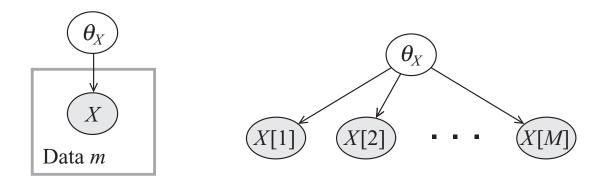
Thumbtack Revisited

- In MLE Tosses were assumed independent
 - Assumption made when θ was fixed
- If we do not know θ then one toss tells us something about the parameter θ
 - Thus about next toss
 - Thus tosses are <u>not</u> marginally independent
- Once we know θ we cannot learn probability of one toss from others
 - So tosses are conditionally independent given θ

Joint distribution of parameter and samples

Tosses are conditionally independent given θ Network for independent identically distributed samples

Plate Model Ground Bayesian network



Joint distribution needs:

- (i) probability of sample given parameter and
- (ii) probability of parameter

Local CPDs for Thumbtack

1. Probability of sample given parameter is $P(X[m] \mid \theta)$

$$P(x[m] \mid \theta) = \begin{cases} \theta & \text{if } x[m] = x^1 \\ 1 - \theta & \text{if } x[m] = x^0 \end{cases}$$

It is equal to θ if sample is heads and 1- θ if sample is tails

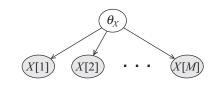
2. Probability distribution of parameter θ

Prior distribution is a continuous density over interval [0,1]

Substituting into joint probability distribution

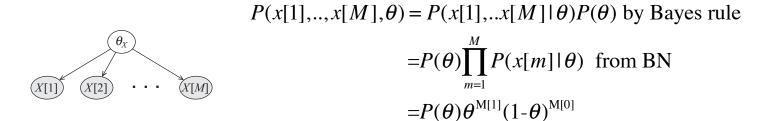
$$P(x[1],...,x[M],\theta) = P(x[1],...x[M]|\theta)P(\theta)$$
 by Bayes rule
= $P(\theta)\prod_{m=1}^{M}P(x[m]|\theta)$ from BN

$$=P(\theta)\theta^{M[1]}(1-\theta)^{M[0]}$$



Using the Joint Probability Distribution

 Network specifies joint probability model over parameters and data



- We can use this model to:
 - 1. Instantiate values of x[1],...,x[M] and compute posterior
 - 2. Predict probability over the next toss

Computing A Parameter's Posterior

- 1. Determine posterior distribution of parameter
- From observed data set \mathcal{D} instantiate values x[1],..x[M] then compute posterior over θ

$$P(\theta \mid x[1], ..x[M]) = \frac{P(x[1], .., x[M] \mid \theta) P(\theta)}{P(x[1], ..x[M])} \qquad P(\theta \mid D]) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

- First term in numerator is likelihood
- Second is the *prior*
- Denominator is a normalizing factor
- If $P(\theta)$ is uniform, i.e., $P(\theta)=1$ for all $\theta \in [0,1]$ posterior is just the normalized likelihood function

Prediction and Uniform Prior

- For predicting probability of heads in next toss
 - In Bayesian approach instead of single value for θ we use its distribution
- Parameter is unknown, so we consider all possible values, with D=x[1],...,x[M]

 $P(x[M+1]|D) = \int P(x[M+1]|\theta,D) P(\theta|D) d\theta$ By Bayes rule & Sum rule = $\int P(x[M+1]|\theta) P(\theta|D) d\theta$ since samples are independent given θ

- Integrating posterior over θ to predict prob heads in next toss
- Thumbtack with Uniform Prior
 - Assume uniform prior over [0,1], i.e., $P(\theta)=1$
 - Substituting $P(\theta \mid D] = \frac{P(D \mid \theta)P(\theta)}{P(D)}$ and $P(\mathcal{D} \mid \theta) = \theta^{M[1]}(1-\theta)^{M[0]}$
 - We get

$$P(X[M+1] = x^1 \mid D) = \frac{M[1]+1}{M[1]+M[0]+2}$$

Similar to MLE estimate except adds an imaginary sample to each count $\hat{\theta} = \frac{M[1]}{M[1] + M[0]}$ Called "Laplace's Correction"

Non-uniform Prior

- How to choose a non-uniform prior?
- Pick a distribution that can be represented
 - 1. Compactly, e.g., analytical formula
 - 2. Updated efficiently as we get new data
- Appropriate prior is Beta Distribution
 - Parameterized by two hyper-parameters
 - α_0 and α_I which are positive reals
 - They are imaginary numbers of heads and tails we have seen before starting the experiment

Details of Beta Distribution

Defined as

$$\theta \sim Beta(\alpha_1, \alpha_0)$$
 if $p(\theta) = \gamma \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_0 - 1}$

γ is a normalizing constant, defined as

$$\gamma = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \text{ where } \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \text{ is the Gamma function}$$

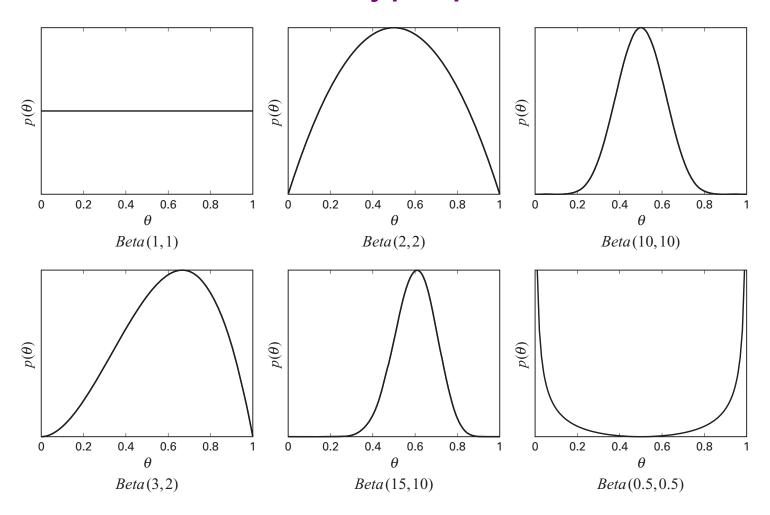
Gamma function is a continuous generalization of factorials

$$\Gamma(1) = 1$$
 and $\Gamma(x+1) = x\Gamma(x)$
Thus $\Gamma(n+1) = n!$ when n is large

 Hyper-parameters correspond to no of heads and tails before experiment

Examples of Beta Distribution

For different hyperparameters



Thumbtack with Beta Prior

• Marginal probability of X based on $P(\theta) \sim Beta(\alpha_0, \alpha_1)$

$$P(X[1] = x^{1}) = \int_{0}^{1} P(X[1] = x^{1} | \theta) \cdot P(\theta) d\theta = \int_{0}^{1} \theta \cdot P(\theta) d\theta = \frac{\alpha_{1}}{\alpha_{1} + \alpha_{0}}$$

Using standard integration techniques

Supports intuition that we have seen α_1 heads and α_0 tails

What happens when we het more samples

$$P(\theta \mid D]) \propto P(D \mid \theta)P(\theta)$$

$$= \theta^{\alpha_1 + M[1] - 1} (1 - \theta)^{\alpha_0 + M[0] - 1}$$

Which is precisely $Beta(\alpha_1 + M/1), \alpha_0 + M/0)$

- Key property of Beta distribution
 - •If prior is a beta distribution, then posterior is also a beta distribution
 - Beta is conjugate with Bernoulli
 - •Immediate consequence is prediction (computing probability over next toss):

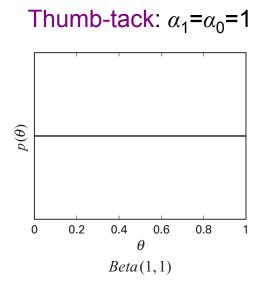
$$P(X[M+1] = x^1 \mid D) = \frac{M[1] + \alpha_1}{M[1] + M[0] + \alpha_1 + \alpha_0}$$
 Tells us that we have seen
$$\alpha_1 + M[1] \text{ heads and } \alpha_0 + M[0] \text{ tails}$$

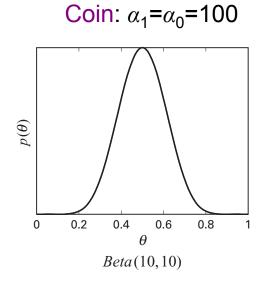
Compare with uniform prior
$$P(X[M+1] = x^1 \mid D) = \frac{M[1]+1}{M[1]+M[0]+2}$$

Machine Learning

Bayesian Approach overcomes both shortcomings of MLE

1. Distinction between thumb-tack and coin captured by choice of hyper-parameters





Beta(10,10) is more Entrenched than Beta(1,1) 3 heads in 10 tosse Gives 0.43 and 0.33 300 heads in 1000 tosses Gives 0.3 in both

2. Distinction between few samples and many is captured by peakedness of posterior

Multinomial Case: Bayesian Approach

- Multinom. Distribn
- X takes one of K values $x^1...x^K$
- $\theta \in [0,1]^K$ i.e, K reals in [0,1]
- $P(x:\theta) = \theta_k if x = x^k$

• <u>Likelihood function</u> (probability of iid samples) $L(\mathbf{\theta}:D) = \prod_{k} \theta_{k}^{M[k]}$

- Since *posterior* = *prior* x *likelihood*
 - Require prior to have form similar to likelihood
 - Dirichlet has that form

$$\mathbf{\theta} \sim Dirichlet(\alpha_1, ..., \alpha_k) \text{ if } P(\mathbf{\theta}) \propto \prod_k \theta_k^{\alpha_k - 1}$$

$$\alpha = \sum_j \alpha_j$$

Dirichlet Distribution

- Generalizes the Beta Distribution
- Specified by a set of hyper-parameters
- If $P(\theta)$ is $Dirichlet(\alpha_1,...\alpha_K)$ then
- 1. $P(\mathbf{\theta}|D)$ is $Dirichlet(\alpha_1 + M[1],...,\alpha_K + M[K])$
- 2. $E[\theta_k] = \alpha_k/\alpha$

Prediction with Dirichlet Prior

• If posterior $P(\theta|D)$ is

Dirichlet
$$(\alpha_1 + M[1],...,\alpha_K + M[K])$$

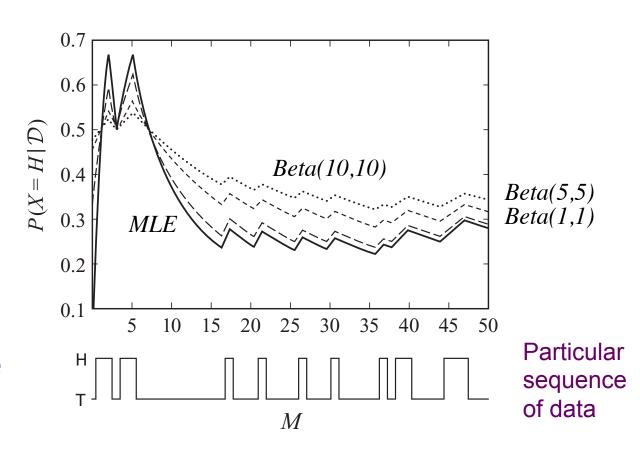
Prediction is

$$P(x[M+1] = x^{k} \mid D) = \frac{M[k] + \alpha_{k}}{M + \alpha}$$

- Dirichlet hyper-parameters are called pseudocounts
- The total α of the pseudo counts reflects how confident we are in our prior
 - Called the equivalent sample size

Effect of different priors

- Changing estimates of θ given data sequence
- Solid line is MLE estimate
- Remaining are Bayesian estimates



Smoother estimate with higher equivalent sample size

Dirichlet Process is about determining cardinality

- A Dirichlet Process is an application of Dirichlet prior to a partial data problem
- Problem of latent variables with unknown cardinality
- A Bayesian approach is used over cardinalities
- A Dirichlet prior over K cardinalities is assumed
- Since K can be infinite, a legal prior is obtained by defining a Chinese restaurant process