Structure Learning: Parameter Estimation

Structure Learning

Learning Bayesian networks from data can be broken down into the following:

1. Known structure, unknown parameters
2. Unknown structure, unknown parameters

The first case, which involves parameter estimation. We will deal with this case today.

Parameter Estimation

There are many techniques for parameter estimation:

Algorithm	Situation
Maximum Likelihood / Maximum a posteriori (MAP)	General
Laplace	2nd order approximation
EM	Missing values, hidden variables
Iterative Proportional Fitting (IPF)	Undirected networks
Mean field	Approximate moments
Gibbs	Approximate moments
MCMC	Approximate moments

Table from Wray Buntine. "A Guide to the Literature on Learning Probabilistic Networks from Data".

Parameter Estimation

There are many techniques for parameter estimation:

Algorithm	Situation
Maximum Likelihood / Maximum a posteriori (MAP)	General
Laplace	2nd order approximation
EM	es, hidden variables
Iterative	etworks
Mean fie	moments
Gibbs	Approximate moments
MCMC	Approximate moments

We will discuss Maximum Likelihood Estimation (MLE), which is part of statistical inference

Table from Wray Buntine. "A Guide to the Literature on Learning Probabilistic Networks from Data".

Statistical Inference

Statistical inference is the process of using data to infer the ditribution that generated the data.

Given a sample $X_1, \ldots, X_n \sim F$ how do we infer F?

This means "drawn from a distribution F"

In some cases, we may want to infer only some feature of F such as its mean

5

Parametric Models

A statistical model F is a set of distributions. A parametric model is a set F that can be parameterized by a finite set of parameters.

In general, a parametric model takes the form:

$$F = \{ f(x;\theta) : \theta \in \Theta \}$$

Value of the random variable

Parameter (or vector of parameters) that can take values in the parameter space Θ.

Side note: A non-parametric model is a set F that cannot be parameterized by a finite number of parameters. It makes no assumptions about the form of the model.

6

Examples of Parametric Models

Discrete Distributions:

1. Bernoulli (Think of this as flipping a coin)

$$f(x;p) = p^x(1-p)^{1-x}$$ for $p \in [0,1]$, $x \in \{0,1\}$

2. Binomial (Think of this as flipping n coins)

$$f(x;n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x=0,\ldots,n,\ p \in [0,1], \text{ n is a positive integer} \\ 0 & \text{otherwise} \end{cases}$$

3. Multinomial (Think of this as flipping a k-sided dice n times)

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

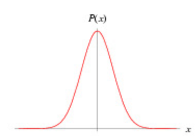
for $\Sigma x_i = n$, $p_i \in [0,1]$, $\Sigma p_i = 1$

Examples of Parametric Models

Continuous Distributions:

1. Normal

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$ for $\mu \in$ Real numbers, $\sigma > 0$



8

Examples of Parametric Models

Continuous Distributions:

2. Beta

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$ for x ε [0,1], α > 0, β > 0

where
$$\Gamma(n) = (n-1)!$$
$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$
$$\Gamma(1) = 1$$

Examples of Parametric Models

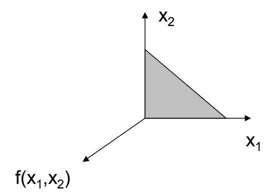
Continuous Distributions:

3. Dirichlet (Generalization of a Beta)

$$f(x_1,\ldots,x_k;\alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}$$

for $x_i \geq 0$, $\Sigma x_i = 1$

For the 2D case:



Frequentist vs Bayesian Inference

- There are two dominant approaches to statistical inference known as the frequentist and Bayesian approaches
- We'll first cover the frequentist approach
- Then we will discuss the Bayesian approach and the differences

Point Estimation

- Point estimation refers to providing a single "best guess" of some quantity of interest eg. a parameter θ
- We denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$
- Note:
  - θ is the true value of the parameter. It is a fixed, unknown quantity
  - $\hat{\theta}$ is an estimate of θ. It depends on the data and is a random variable

Point Estimation (Formally)

Let $X_1, \ldots, X_n$ be n independent, identically distributed data points from some distribution F.

A point estimator $\hat{\theta}$ of a parameter θ is some function of $X_1, .., X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

The bias of an estimator is defined by:

$$bias(\hat{\theta}_n) = E_\theta(\hat{\theta}_n) - \theta$$

$\hat{\theta}$ is unbiased if $bias(\hat{\theta}_n) = 0$

13

Point Estimation

A point estimator $\hat{\theta}$ of a parameter θ is consistent if:

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0 \quad \text{for every } \varepsilon > 0 \text{ as } n \to \infty$$

We say " $\hat{\theta}$ converges to θ in probability" or write:

$$\hat{\theta}_n \xrightarrow{p} \theta$$

14

Maximum Likelihood

Let $X_1, \ldots, X_n$ be independent, identically distributed with pdf $f(x;\theta)$. The likelihood function is defined by:

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

The log-likelihood function is defined by:

$$l_n(\theta) = \log L_n(\theta)$$

15

Maximum Likelihood

The maximum likelihood function is just the joint density of the data. We treat it as a function of θ ie. $L_n : \Theta \to [0,\infty)$

Note: The likelihood function is not a density function. In general, it is not true that $L_n(\theta)$ integrates to 1 with respect to θ

16

Maximum Likelihood

- The Maximum Likelihood Estimator (MLE) denoted $\hat{\theta}$ is the value of θ that maximizes $L_n(\theta)$.
- Maximizing the log-likelihood leads to the same answer as maximizing the likelihood.
- Note: Multiplying $L_n(\theta)$ by any positive constant c does not change the MLE. We tend to drop constants in the likelihood function

Maximum Likelihood

Example: You buy a bag of lime-cherry candy with *n* pieces of candy. You unwrap all n pieces, resulting in data $X_1, \ldots, X_n$ where X_i = {*cherry*, *lime*}.

You want to estimate θ, which is the probability that a randomly chosen candy from the bag is cherry flavored.

The probability function for a single candy is $f(x;\theta) = \theta^x(1-\theta)^{1-x}$ (Bernoulli distribution)

Maximum Likelihood

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i} = \theta^c(1-\theta)^l$$

Let c = # of cherries and n-c = l be the # of limes. Note that $c = \sum_i X_i$

$$l_n(\theta) = c\log\theta + (n-c)\log(1-\theta)$$

$$\frac{\partial}{\partial\theta} l_n(\theta) = \frac{c}{\theta} - \frac{n-c}{1-\theta} = 0$$

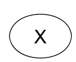
$$\Rightarrow \frac{c(1-\theta)-(n-c)\theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow c - c\theta - n\theta + c\theta = 0$$

$$\Rightarrow c = n\theta$$

$$\Rightarrow \theta = \frac{c}{n} \qquad \therefore \hat{\theta}_n = \frac{c}{n}$$

You've just estimated the parameters for a one node Bayesian network with the following CPT:



X	P(X)
lime	1-θ
cherry	θ

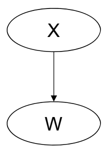
Maximum Likelihood

The recipe for the MLE:

1. Write down $L_n(\theta)$ or $l_n(\theta)$
2. Take derivative with respect to each parameter
3. Find the parameter values such that the derivatives are zero

Maximum Likelihood

Example: suppose the candy wrapper gives a hint as to the flavor. The wrapper can be red or green and is chosen probabilistically given the flavor X.



X	P(X)
lime	1-θ
cherry	θ

X	W	P(W|X)
cherry	red	θ_c
cherry	green	1-θ_c
lime	red	θ_l
lime	green	1-θ_l

P(W,X) = P(W|X)P(X)

21

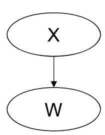
Maximum Likelihood

Notation:

- c = # of cherry flavored candies
- l = # of lime flavored candies
- r_c = # of cherry flavored candies with red wrappers
- g_c = # of cherry flavored candies with green wrappers
- r_l = # of lime flavored candies with red wrappers
- g_l = # of lime flavored candies with green wrappers

22

Maximum Likelihood



X	P(X)
lime	1-θ
cherry	θ

X	W	P(W|X)
cherry	red	θ_c
cherry	green	1-θ_c
lime	red	θ_l
lime	green	1- θ_l

$$L_n(\theta,\theta_c,\theta_l) = (\theta^c(1-\theta)^l)(\theta_c^{r_c}(1-\theta_c)^{g_c})(\theta_l^{r_l}(1-\theta_l)^{g_l})$$

$$l_n(\theta,\theta_c,\theta_l) = [c\log\theta + l\log(1-\theta)] + [r_c\log\theta_c + g_c\log(1-\theta_c)] + [r_l\log\theta_l + g_l\log(1-\theta_l)]$$

$$\frac{\partial}{\partial\theta} l_n(\theta) = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \qquad \Rightarrow \theta = \frac{c}{c+l}$$

$$\frac{\partial}{\partial\theta_c} l_n(\theta_c) = \frac{r_c}{\theta_c} - \frac{g_c}{1-\theta_c} = 0 \qquad \Rightarrow \theta_c = \frac{r_c}{r_c + g_c}$$

$$\frac{\partial}{\partial\theta_l} l_n(\theta_l) = \frac{r_l}{\theta_l} - \frac{g_l}{1-\theta_l} = 0 \qquad \Rightarrow \theta_l = \frac{r_l}{r_l + g_l}$$

23

Maximum Likelihood

- With complete data (ie. no missing values or hidden variables), parameter learning decomposes into separate learning problems, one for each parameter
- If any of the observed counts are 0, the MLE for that parameter is 0
- The MLE is consistent:

$$\hat{\theta}_n \xrightarrow{p} \theta^*$$

Where θ^* is the true value of the parameter θ

24