Parameter Estimation with Complete Data

Brandon Malone

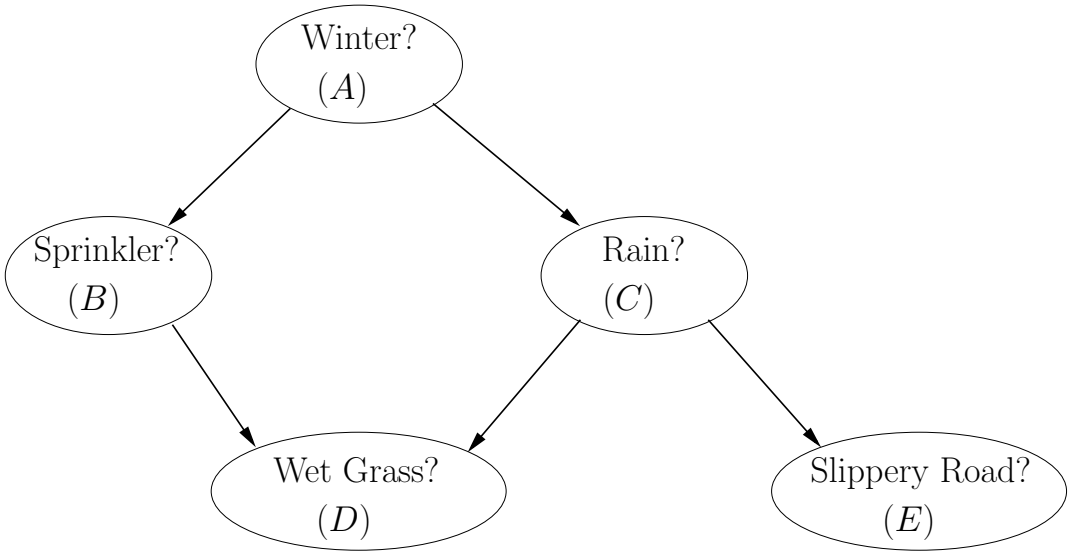
Much of this material is adapted from Chapter 17 in Koller's book

Many of the images were taken from the Internet

February 13, 2014

Parameter Estimation with Complete Data

Suppose we have a general Bayesian network structure $\mathcal{N}$, but do not know the parameters.



How do we estimate the parameters from a complete dataset $\mathcal{D}$?

Maximum Likelihood Estimates

Suppose we want to maximize $P(\mathcal{D}|\mathcal{N})$. How should we set the probabilities?

- Revisiting the multinomial
- Likelihood of data
- Maximum likelihood parameter estimates

Revisiting the multinomial

Previously, we saw that for a set of observations, we should set binomial parameters as follows (ignoring smoothing).

$$p_H = \frac{n_H}{n_H + n_T}$$

Why?

Revisiting the multinomial

Say we observe $\mathcal{D} = H, H, T, H, T$ and set $p_H = \theta$.

$$\begin{aligned} L(\theta : \mathcal{D}) &= P(\mathcal{D} : \theta) \\ &= \theta\theta(1-\theta)\theta(1-\theta) \\ &= \theta^{n_H}(1-\theta)^{n_T} \end{aligned}$$

Revisiting the multinomial

We can simplify things by taking the logarithm.

$$\ell(\theta : \mathcal{D}) = n_H \log \theta + n_T \log (1 - \theta)$$

We can now maximize $\ell(\theta : \mathcal{D})$ w.r.t. θ by taking the derivative and setting equal to 0 and solving for $\hat{\theta}$.

$$\frac{\partial \ell}{\partial \theta} = \frac{n_H}{\theta} - \frac{n_T}{1 - \theta}$$
$$0 = \frac{n_H}{\theta} - \frac{n_T}{1 - \theta}$$
$$\hat{\theta} = \frac{n_H}{n_H + n_T}$$

Notation

- X_i: a variable
- r_i: the number of values of X_i
- PA_i: the parents of X_i
- q_i: the number of instantiations of PA_i
- θ_{ijk}: the parameter when $X_i = k$ and $PA_i = j$
- Θ: all parameters in a network
- n: the number of variables in a network.
- n_{ijk}: the number of times we observe $X_i = k$ and $PA_i = j$ in $\mathcal{D}$
- N: the number of records in $\mathcal{D}$
- $\theta_{ijk:l}$ $X_i = k$ and $PA_i = j$ in record l

X_1	X_2	X_3	$\theta_{X_3 \mid X_1, X_2}$	label
T	T	T	.95	$\theta_{3,1,1}$
T	T	F	.05	$\theta_{3,1,2}$
T	F	T	.9	$\theta_{3,2,1}$
⋮	⋮	⋮	⋮	⋮
F	F	F	1	$\theta_{3,4,2}$

Likelihood of parameters in a Bayesian network

We define likelihood of Θ for a single $D_l \in \mathcal{D}$ based on chain rule.

$$\begin{aligned} L(\theta : D_l) &= P(D_l : \Theta) \\ &= \prod_{i=1}^{n} \theta_{ijk:l}, \end{aligned}$$

We can now define the likelihood of Θ for the entire datasets $\mathcal{D}$.

$$\begin{aligned} L(\theta : \mathcal{D}) &= P(\mathcal{D} : \Theta) \\ &= \prod_{l=1}^{N} P(D_l : \Theta) \\ &= \prod_{l=1}^{N} \prod_{i=1}^{n} \theta_{ijk:l} \\ &= \prod_{i}^{n} \prod_{k}^{r_i} \prod_{j}^{q_i} \theta_{ijk}^{n_{ijk}} \end{aligned}$$

Likelihood of parameters in a Bayesian network

We define likelihood of Θ for a single $D_l \in \mathcal{D}$ based on chain rule.

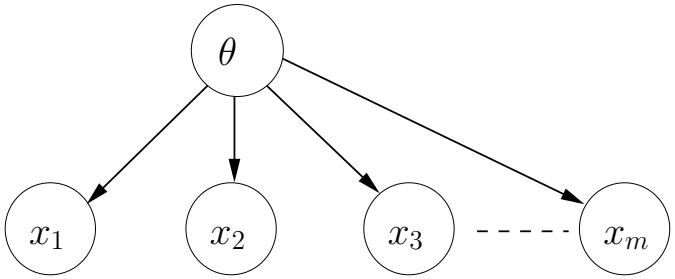
$$\begin{aligned} L(\theta : D_l) &= P(D_l : \Theta) \\ &= \prod_{i=1}^{n} \theta_{ijk:l}, \end{aligned}$$

Note that, just like in the multinomial, the parameters are independent. Similar reasoning shows the maximum likelihood estimates.

$$\hat{\theta}_{ijk}^{ML} = \frac{n_{ijk}}{\sum_k n_{ijk}}$$

Explicit parameters

Consider flipping m iid coins where we observe n_H heads and n_T tails. The set of all results is $\mathcal{D}$.



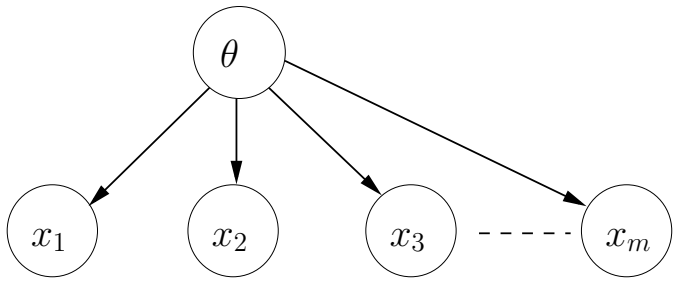
$$P(x_i = \mathrm{H}|\theta) = \theta$$

So then,

$$\begin{aligned} P(\mathcal{D}|\theta) &= \prod_m P(x_i|\theta) \\ &= \theta^{n_H}(1-\theta)^{n_T} \end{aligned}$$

Explicit parameters

Consider flipping m iid coins where we observe n_H heads and n_T tails. The set of all results is $\mathcal{D}$.



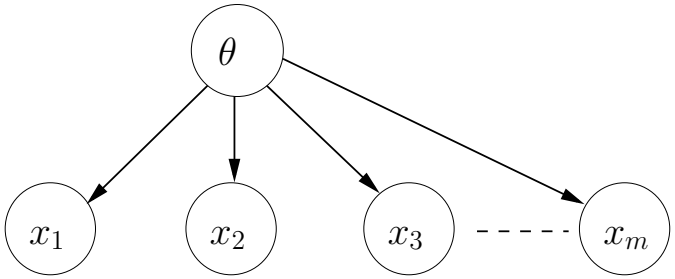
$$P(x_i = \mathrm{H}|\theta) = \theta$$

So then,

$$\begin{aligned} P(\mathcal{D}, \theta) &= P(\theta)P(\mathcal{D}|\theta) \\ &= P(\theta)\theta^{n_H}(1-\theta)^{n_T} \end{aligned}$$

Explicit parameters

Consider flipping m iid coins where we observe n_H heads and n_T tails. The set of all results is $\mathcal{D}$.



$$P(x_i = \mathrm{H}|\theta) = \theta$$

So then,

What do we know about θ?

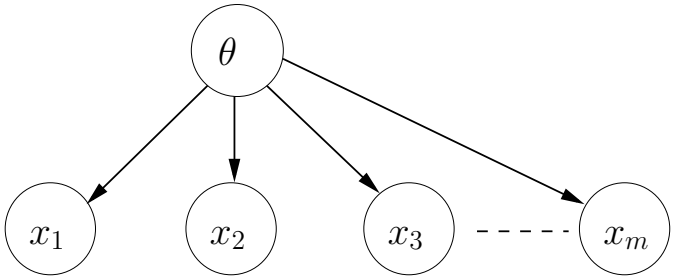
$\theta \in [0, 1]$.

What do we know about $P(\theta)$?

It is a probability distribution, so $\int_0^1 P(\theta)d\theta = 1$.

Explicit parameters

Consider flipping m iid coins where we observe n_H heads and n_T tails. The set of all results is $\mathcal{D}$.



$$P(x_i = \mathrm{H}|\theta) = \theta$$

So then,

Say we flip a single coin. What is $P(x_1 = \mathrm{H})$?

$$\begin{aligned} P(x_1 = \mathrm{H}) &= \int_0^1 P(x_1 = \mathrm{H}|\theta)P(\theta)d\theta \\ &= \int_0^1 \theta P(\theta)d\theta \end{aligned}$$

Bayesian prediction with a uniform prior

After we flip m coins, what is $P(x_{m+1} = \mathrm{H}|\mathcal{D})$? ("prediction")

$$
\begin{aligned}
P(x_{m+1} = \mathrm{H}|\mathcal{D}) &= \int_0^1 P(x_m = \mathrm{H}|\mathcal{D}, \theta) P(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta P(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} d\theta \\
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H} (1-\theta)^{n_T} P(\theta) d\theta \\
&\text{Say that } \theta \sim \text{Uniform}(0, 1) \\
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H} (1-\theta)^{n_T} d\theta \\
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta^{n_H+1} (1-\theta)^{n_T} d\theta \\
&\ldots \text{Calculus} \ldots \\
&= \frac{n_H + 1}{n_H + n_T + 2}
\end{aligned}
$$

How does this relate to "smoothing?"

Bayesian prediction with a uniform prior

After we flip m coins, what is $P(x_{m+1} = \mathrm{H}|\mathcal{D})$? ("prediction")

$$
\begin{aligned}
P(x_{m+1} = \mathrm{H}|\mathcal{D}) &= \int_0^1 P(x_m = \mathrm{H}|\mathcal{D}, \theta) P(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta P(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} d\theta \\
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H} (1-\theta)^{n_T} P(\theta) d\theta
\end{aligned}
$$

Say that $\theta \sim \text{Uniform}(0, 1)$ **Do we really believe this?**

$$
\begin{aligned}
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H} (1-\theta)^{n_T} d\theta \\
&= \frac{1}{P(\mathcal{D})} \int_0^1 \theta^{n_H+1} (1-\theta)^{n_T} d\theta
\end{aligned}
$$

. . . Calculus . . .

$$
= \frac{n_H + 1}{n_H + n_T + 2}
$$

How does this relate to "smoothing?"

The Dirichlet distribution

The **Dirichlet distribution** allows us to express biased preferences.

If $\theta \sim$Dirichlet(α_H, α_T), then

$$\begin{aligned} P(\theta) &= \frac{\Gamma(\alpha_H + \alpha_T)}{\Gamma(\alpha_H)\Gamma(\alpha_T)} \theta^{(\alpha_H - 1)}(1-\theta)^{(\alpha_T - 1)} \\ &\propto \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1} \end{aligned}$$

Say we flip a single coin. What is $P(x_1 = \mathrm{H})$?

$$\begin{aligned} P(x_1 = \mathrm{H}) &= \int_0^1 P(x_1 = \mathrm{H}|\theta) P(\theta) d\theta \\ &= \int_0^1 \theta P(\theta) d\theta \\ &\propto \int_0^1 \theta \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1} d\theta \\ &\ldots \text{Calculus} \ldots \\ &= \frac{\alpha_H}{\alpha_H + \alpha_T} \end{aligned}$$

The αs are "imaginary observations" which bias our beliefs.

Conjugate priors

After we flip m coins, what is $P(x_{m+1} = \mathrm{H}|\mathcal{D})$? ("prediction")

To use the previous derivation, we need $P(\theta|\mathcal{D})$. Say $\theta \sim$Dirichlet(α_H, α_T).

$$\begin{aligned} P(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta)P(\theta) \\ &\propto \theta^{n_H}(1-\theta)^{n_T}\theta^{\alpha_1-1}(1-\theta)^{alpha_0-1} \\ &= \theta^{n_H+\alpha_H-1}(1-\theta)^{n_T+\alpha_T-1} \end{aligned}$$

So $P(\theta|\mathcal{D}) \sim$ Dirichlet$(n_H + \alpha_H, n_T + \alpha_T)$.

The Dirichlet is a **conjugate prior** for the multinomial because θ is from the same family after observing data.

Bayesian prediction with the Dirichlet prior

$$P(x_{m+1} = \mathrm{H}|\mathcal{D}) = \int_0^1 P(x_m = \mathrm{H}|\mathcal{D}, \theta) P(\theta|\mathcal{D}) d\theta$$

$$= \int_0^1 \theta P(\theta|\mathcal{D}) d\theta$$

$$= \int_0^1 \theta \frac{P(\mathcal{D}|\theta) P(\theta)}{P(\mathcal{D})} d\theta$$

$$= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H} (1-\theta)^{n_T} P(\theta) d\theta$$

Say $\theta \sim \mathrm{Dirichlet}(\alpha_H, \alpha_T)$

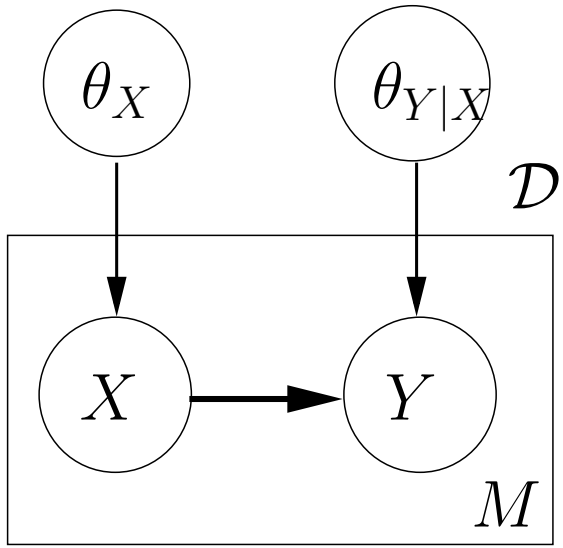
$$= \frac{1}{P(\mathcal{D})} \int_0^1 \theta \theta^{n_H + \alpha_H - 1} (1-\theta)^{n_T + \alpha_T - 1} d\theta$$

. . . Calculus . . .

$$= \frac{\alpha_H + n_H}{\alpha_H + n_H + \alpha_T + n_T}$$

Assumption: Global parameter independence

Global parameter independence: knowing something about the parameters for one variable tells us nothing about the parameters for the other variables



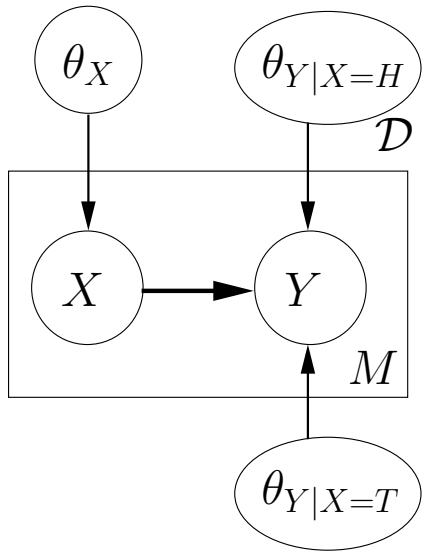
We see that $I(\theta_X, \mathcal{D}, \theta_{Y|X})$.

So then $P(\theta_X, \theta_{Y|X}|\mathcal{D}) = P(\theta_X|\mathcal{D})P(\theta_{Y|X}|\mathcal{D})$.

In general, $P(\Theta|\mathcal{D}) = \prod_{\theta \in \Theta} P(\theta|\mathcal{D})$.

Assumption: Local parameter independence

Local parameter independence: knowing something about the parameters for one parent instantiation tells us nothing about parameters for the others



We see that $I(Y, X = \mathsf{H}, \theta_{Y|X=\mathsf{H}})$.

So then $P(\theta_{Y|X}|\mathcal{D}) = P(\theta_{Y|X=\mathsf{H}}|\mathcal{D})P(\theta_{Y|X=\mathsf{H}}|\mathcal{D})$.

In general, $P(\theta_{X_i|PA_i}|\mathcal{D}) = \prod_j^{q_i} \prod_k^{r_i} P(\theta_{ijk}|\mathcal{D})$.

Bayesian prediction in Bayesian networks

If we assume global and local independence and Dirichlet priors on parameters, then

$$
\begin{aligned}
P(\mathbf{x}_{m+1}|\mathcal{D}) &= \prod_i^n \int_\Theta P(\mathbf{x}_{ijk:m+1}|\theta_{ijk}) P(\theta_{ijk}|\mathcal{D}) d\Theta \\
&= \prod_i^n \prod_j^{q_i} \prod_k^{r_i} \int_{\theta_{ijk}} P(\mathbf{x}_{ijk:m+1}|\theta_{ijk}) P(\theta_{ijk}|\mathcal{D}) d\theta_{ijk} \\
&\ldots \text{Calculus} \ldots \\
&= \prod_i^n \prod_j^{q_i} \prod_k^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\sum_k (\alpha_{ijk} + n_{ijk})}
\end{aligned}
$$

In all of the equations, we only consider combinations of i, j and k compatiable with $\mathbf{x}$.

Long story short ...

$$P(\mathbf{x}_{m+1}|\mathcal{D}) = \prod_i^n \prod_j^{q_i} \prod_k^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\sum_k \alpha_{ijk} + n_{ijk}}$$

As this equation suggests, given global and local parameter independence and a Dirichlet prior on parameters, we can calculate the Bayesian prediction for a new observation by setting

$$\theta_{ijk} = \frac{\alpha_{ijk} + n_{ijk}}{\sum_k (\alpha_{ijk} + n_{ijk})}.$$

These are sometimes called the *maximum a posteriori* (MAP) parameters.

Picking α_{ijk}

Where do the α_{ijk}s come from?

We could solicit them from experts . . .

How many α_{ijk}s are there?

Picking α_{ijk}

Where do the α_{ijk}s come from?

We could solicit them from experts . . .

How many α_{ijk}s are there?

It will be more convenient to select an **equivalent sample size** α.

$$\alpha_{ijk} = \frac{\alpha}{r_i \cdot q_i}$$

This setting has nice properties for structure learning.

Recap

During this part of the course, we have discussed:

- Theoretical justification for MLE parameters
- Bayesian parameter learning
- Prediction

Next in probabilistic models

Structure learning based on

- Minimum description length principle
- Bayesian criterion