HW4: due 2 weeks
Quiz: MCMC

COMPSCI 688: Probabilistic Graphical Models

Lecture 17: (Conjugate) Bayesian Inference

Dan Sheldon

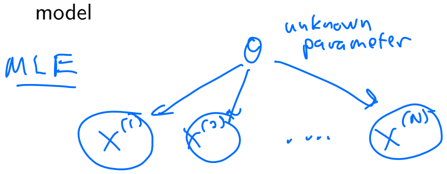
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Bayesian Inference

Bayesian Inference Example

Suppose we observe data $x^{(1)}, \ldots, x^{(N)}$ which we assume to come from a Bernoulli model



$$p(x^{(n)}|\theta) = \begin{cases} \theta & x^{(n)} = 1 \\ 1-\theta & x^{(n)} = 0 \end{cases}$$

- Maximum-likelihood says to find θ by solving $\max_\theta \frac{1}{n} \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$

When might we want something different?

Example: you go on a three-day trip to Australia and want to learn about the weather

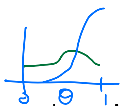
$$X = \begin{cases} 1 & \text{rain} \\ 0 & \text{no rain} \end{cases}$$

Observe $x^{(1)} = 1, x^{(2)} = 1, x^{(3)} = 1$

MLE learning $\rightarrow \hat{\theta} = 1$

It rains every day in Australia. What went wrong?

Being Bayesian



A Bayesian says: give me the **probability** of θ given the data. What does this mean?

$$p(\theta|\text{Data}) = \frac{p(\theta)p(\text{Data}|\theta)}{p(\text{Data})}$$

$p(\theta, \text{Data})$

← "Z"

- $p(\theta)$ is the **prior**. It encodes beliefs (either subjective or objective) about θ **prior** to seeing any evidence. We need one!
- $p(\text{Data}|\theta) = \prod_{n=1}^{N} p(x^{(n)}|\theta)$ is the **likelihood**. It incorporates evidence.
- $p(\text{Data}) = \int p(\theta)p(\text{Data}|\theta)d\theta$ is the **marginal likelihood** or **evidence**. We usually don't need to compute it.
- $p(\theta|\text{Data})$ is the **posterior**. What we believe about θ after observing data.

Why Be Bayesian?

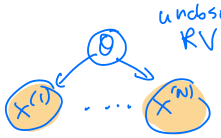
- Philosophy: Update subjective prior beliefs based on evidence.
- Practical: deal with small samples
- Practical: excellent tools exist (MCMC, stan)

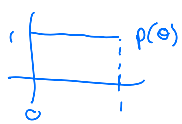
Numpyro

↑ ↑

PPL

Making our Model Bayesian





$$p(\theta) = \begin{cases} 1 & \theta \in [0,1] \\ 0 & \text{o.w.} \end{cases}$$

$$\boxed{\theta \sim \text{Uniform}([0,1])}$$
$$x^{(n)} \sim \text{Bernoulli}(\theta)$$

$$p(x^{(n)}|\theta) = \begin{cases} \theta & x^{(n)} = 1 \\ 1-\theta & x^{(n)} = 0 \end{cases}$$

Bayesian Modeling: Implications

$p(x;\theta)$

↓

- We now have a **joint probability model** $p(\theta, x)$

$$p(\theta, x) = p(\theta)p(x|\theta)$$

- θ is now a **random variable** instead of a fixed but unknown parameter
- Learning is replaced by **posterior inference**
  - Learning: $\max_\theta \mathcal{L}(\theta|x^{(1)}, \ldots, x^{(N)})$
  - Posterior inference: compute $p(\theta|x^{(1)}, \ldots, x^{(N)})$

e.g. draw samples

Posterior Inference

$x^{(1)}, \dots, x^{(N)}$

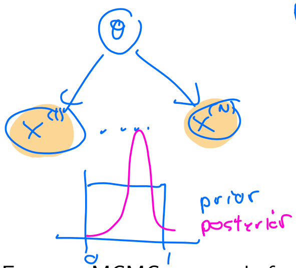
$$p(\theta \mid x^{(1:N)}) = \frac{p(\theta)\, p(x^{(1:N)} \mid \theta)}{p(x^{(1:N)})}$$

(assume $\theta \in [0,1]$)

$$\propto p(\theta)\, p(x^{(1:N)} \mid \theta)$$

$$= 1 \cdot \prod_{i=1}^{N} \theta^{\mathbb{I}[x^{(n)}=1]} (1-\theta)^{\mathbb{I}[x^{(n)}=0]}$$

$$= \theta^{\#(x=1)} (1-\theta)^{\#(x=0)}$$



E.g., use MCMC to sample from density on $[0,1]$ proportional to this

General inference strategy: use MCMC to sample from density proportional to $p(\theta)p(\text{Data} \mid \theta)$

But in some *special* cases the problem is easy to solve without MCMC...

Conjugate Bayesian Inference

The Easy Case: Conjugacy

Beta ~~Uniform~~ Bernoulli

$p(\theta)$ $p(x \mid \theta)$

Some prior-likelihood pairs have a special relationship that makes computing the posterior easy

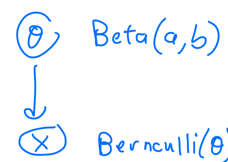
This relationship is called **conjugacy**. It means the posterior $p(\theta|x)$ will be in the same parametric family as the prior $p(\theta)$. E.g.

$$p(\theta) = \text{Beta}(\theta|a,b) \qquad \Longrightarrow \qquad p(\theta|x) = \text{Beta}(\theta|a',b')$$

We say:

- $p(\theta)$ is a conjugate prior for $p(x|\theta)$
- $p(\theta)$ and $p(x|\theta)$ are a conjugate pair

Example: Beta-Bernoulli Model



$x \sim \text{Bernoulli}(\theta)$

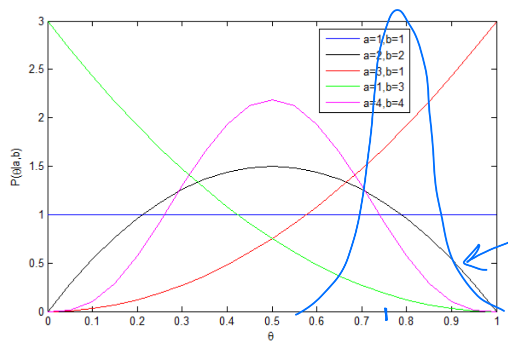
Likelihood: $p(x|\theta) = \text{Bernoulli}(x|\theta)$

Prior: $p(\theta) = \text{Beta}(\theta|a,b)$

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0,1]$$

Beta Density

θ is a prob.



mean $\frac{a}{a+b}$

$a, b \to \infty$, more concentrated

$a=150,\ b=50$

Beta Density

$\frac{(a+b-1)!}{(a-1)!(b-1)!}$

$$\text{Beta}(\theta|a,b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{normalization const.}} \underbrace{\theta^{a-1}(1-\theta)^{b-1}}_{\text{unnormalized density}}, \quad \theta \in [0,1]$$

Discuss

- Unnormalized density!
- Gamma function
  - $\Gamma(t) = \int_0^\infty z^{t-1} e^{-z} dz$
  - $\Gamma(n) = (n-1)!$ for integer n

$\Gamma(100) = 99!$

$\log \Gamma(100)$

"lngamma"

Beta Density

$$\text{Beta}(\theta|a,b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{normalization const.}} \underbrace{\theta^{a-1}(1-\theta)^{b-1}}_{\text{unnormalized density}}, \quad \theta \in [0,1]$$

Question: $p(\theta) \propto \theta^2(1-\theta)^4$ on $\theta \in [0,1]$. What is normalized density?

$= \theta^{a-1}(1-\theta)^{b-1} \qquad a=3, b=5$

$\propto \text{Beta}(\theta \mid 3,5)$

$= \frac{\Gamma(8)}{\Gamma(3)\Gamma(5)} \theta^2 (1-\theta)^4$

The point: recognize unnormalized density, get normalization constant for free

Beta-Bernoulli Posterior

Observe x. **Easy way**: drop all terms that don't involve θ

$p(\theta|x) = p(\theta)p(x|\theta)/p(x)$

$\propto p(\theta)p(x|\theta)$

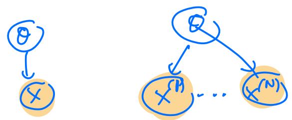
$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^{I[x=1]}(1-\theta)^{I[x=0]}$

$\propto \theta^{\overbrace{a-1+I[x=1]}^{a'}}(1-\theta)^{\overbrace{b-1+I[x=0]}^{b'}}$

$p(\theta \mid x) = \text{Beta}(\theta \mid a + \mathbb{I}[x=1], b + \mathbb{I}[x=0])$

Result: posterior is also Beta (**conjugate**!). Add one to either a or b depending on value of x.

Beta-Bernoulli Belief Updating



Observe $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$, want to compute $p(\theta | x^{(1)}, \ldots, x^{(N)})$

By applying the simple posterior update we just saw sequentially, we get

$$p(\theta | x^{(1)}, \ldots x^{(N)}) = \text{Beta}(\theta \mid \; a + \textstyle\sum_{n=1}^{N} \mathbb{I}[x^{(n)} = 1], \; b + \sum_{n=1}^{N} \mathbb{I}[x^{(n)} = 0])$$

$$= \text{Beta}(\theta \mid \; \underbrace{a + \#(X = 1)}_{a'}, \; \underbrace{b + \#(X = 0)}_{b'})$$

Simple updates based on counting

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

Beta-Bernoulli Belief Updating

Demo

Mixture Model

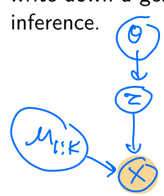
Mixture Model

"Mechanistic": e.g. COVID

"Statistical": Beta-Bernoulli

Bayesian Modeling with generic inference techniques like MCMC is powerful. We can write down a generative model that we think is a good match to our data and perform inference.

"flexible"



$$p(z) = \begin{cases} \theta_1 & z=1 \\ \vdots & \vdots \\ \theta_K & z=K \end{cases}$$

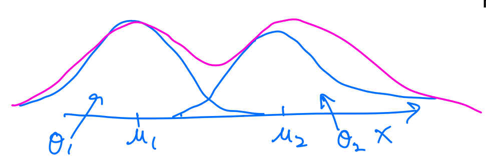
Likelihood:

$$z \sim \text{Categorical}(\theta_1, \ldots, \theta_K)$$
$$x \sim \mathcal{N}(\mu_z, 10)$$

Prior:

$$\theta \sim \text{Dirichlet}(1)$$
$$\mu_z \sim \mathcal{N}(100, 20), \quad z \in \{1, \ldots, K\}$$

dist on vectors θ s.t. $\theta_k \geq 0$, $\sum_k \theta_k = 1$



Mixture with Many Observations

Suppose we draw many $(z^{(n)}, x^{(n)})$ pairs and observe only $x^{(n)}$ (i.e., $z^{(n)}$ is a *latent* variables). Here's what the graphical model looks like:

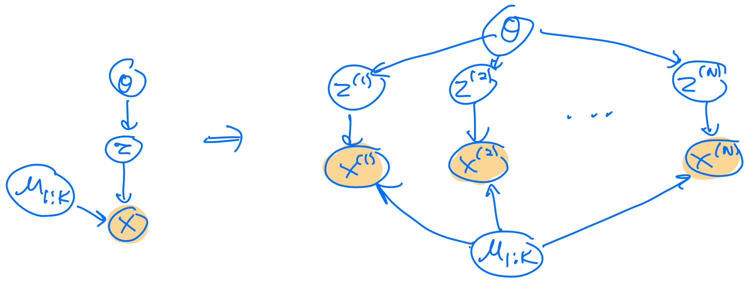
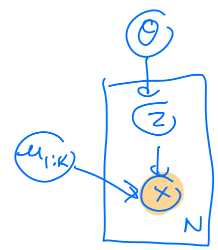


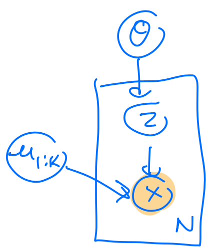
Plate Notation

We can draw the same thing compactly in plate notation to indicate repetition



same graphical model as prev. slide

Computing the Posterior



The posterior in this model looks like this:

$$p(\theta, \mu_{1:K}, z^{(1:N)} \mid x^{(1:N)})$$

$$= p(\theta, \mu_{1:K}, z^{(1:N)}, x^{(1:N)}) / p(x^{(1:N)})$$

$$\propto p(\theta)\, p(\mu_{1:K})\, p(z^{(1:N)} \mid \theta)\, p(x^{(1:N)} \mid z^{(1:N)}, \mu_{1:K})$$

We could sample from this unnormalized distribution using MCMC.