Bayesian Inference

Conjugate Bayesian Inference

Mixture Model

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 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 \( \hat{\theta} \) by solving

\[
\max \frac{1}{n} \sum_{n=1}^{N} \log p(x^{(n)}|\theta)
\]

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 \( \to \hat{\theta} = 1 \)

It rains every day in Australia. What went wrong?
Bayesian Inference

Conjugate Bayesian Inference

Mixture Model

Being Bayesian

A Bayesian says: give me the probability of \( \theta \) given the data. What does this mean?

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

- \( p(\theta) \) is the prior. It encodes beliefs (either subjective or objective) about \( \theta \) 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 \( \theta \) 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)

Making our Model Bayesian

\( \mathcal{X} \)

\( x^{(1)} \) \ldots \( x^{(N)} \)

\( \theta \sim \text{Uniform}(0, 1) \)

\( x^{(n)} \sim \text{Bernoulli}(\theta) \)

\[
p(\theta) = \begin{cases} 1 & \text{if } \theta \in [0, 1] \\ 0 & \text{otherwise} \end{cases}
\]

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

\[
p(\theta, x) = p(\theta)p(x | \theta)
\]

- \( \theta \) 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
Bayesian Inference

Conjugate Bayesian Inference

Mixture Model

Posterior Inference

\[
p(\theta | x (1:N)) = \frac{p(\theta)p(x (1:N) | \theta)}{p(x (1:N))} \quad \text{(assume } \Theta \in \Theta_0)\]

\[
\theta \sim \text{Beta}(a', b') \quad \text{if} \quad \text{Beta}(\theta | a, b)
\]

\[
p(\theta | x) = \text{Beta}(\theta | a', b')
\]

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 \(p(\theta)p(\text{Data} | \theta)\)

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

The Easy Case: Conjugacy

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.

\[
p(\theta) = \text{Beta}(\theta | a, b) \implies p(\theta | x) = \text{Beta}(\theta | a', b')
\]

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

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

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

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

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 \(p(\theta)p(\text{Data} | \theta)\)

But in some special cases the problem is easy to solve without MCMC.
Bayesian Inference
Conjugate Bayesian Inference
Mixture Model
Beta Density

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

Question: \( p(\theta) \propto \theta^a (1-\theta)^b \) on \( \theta \in [0, 1] \). What is normalized density?
\[ a = 3, \quad b = 5 \]
\[ \propto \text{Beta}(\theta | 3, 5) \]
\[ = \frac{\Gamma(3)}{\Gamma(3) \Gamma(5)} \theta^2 (1-\theta)^4 \]

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

Beta-Bernoulli Posterior

Observe \( x \). Easy way: drop all terms that don’t involve \( \theta \)
\[ p(\theta | x) = \frac{p(\theta) p(x|\theta)}{p(x)} \propto p(\theta) p(x|\theta) \]
\[ = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \cdot \theta^{x-1} (1-\theta)^{1-x} \]

Result: posterior is also Beta (conjugate!). Add one to either \( a \) or \( b \) depending on value of \( x \).
Bayesian Inference
Conjugate Bayesian Inference
Mixture Model
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

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

\[
= \text{Beta}(\theta | a + \#(X = 1), b + \#(X = 0))
\]

Simple updates based on counting

\[
\text{MLE } \hat{\theta} = \frac{\#(X = 1)}{N}
\]

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.

\[
p(z) = \begin{cases} \theta_z z = 1 \\ \vdots \\ \theta_{z=K} z = K \end{cases}
\]

Likelihood:

\[
z \sim \text{Categorical}(\theta_1, \ldots, \theta_K)
\]

\[
x \sim \mathcal{N}(\mu_z, 10)
\]

Prior:

\[
\theta \sim \text{Dirichlet}(1)
\]

\[
\mu_z \sim \mathcal{N}(100, 20), \quad z \in \{1, \ldots, K\}
\]
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 variable). Here’s what the graphical model looks like:

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

Computing the Posterior
The posterior in this model looks like this:

We could sample from this unnormalized distribution using MCMC.