Econ 2120: Section 3

Part II - Bayesian Inference Refresher

Ashesh Rambachan

Fall 2018

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Recall: Inference Problem

The data are a realization of some random vector

$$D = (Y_1, \ldots, Y_n, Z_1, \ldots, Z_n),$$

where Y_i is a scalar outcome and Z_i is a vector of predictors.

Also write $D = (D_1, \ldots, D_n)$ with $D_i = (Y_i, Z_i)$.

Recall: Inference Problem

The data are a realization of some random vector

$$D = (Y_1, \ldots, Y_n, Z_1, \ldots, Z_n),$$

where Y_i is a scalar outcome and Z_i is a vector of predictors.

Also write $D = (D_1, \ldots, D_n)$ with $D_i = (Y_i, Z_i)$.

Assume D is drawn from some distribution (unknown).

Recall: Inference Problem

The data are a realization of some random vector

$$D = (Y_1, \ldots, Y_n, Z_1, \ldots, Z_n),$$

where Y_i is a scalar outcome and Z_i is a vector of predictors.

Also write $D = (D_1, \ldots, D_n)$ with $D_i = (Y_i, Z_i)$.

Assume D is drawn from some distribution (unknown).

Specify set of distributions that contains D

$$D \sim P_\theta, \quad \text{for some} \quad \theta \in \Theta.$$

Θ is the **parameter space**.

Assume D_i i.i.d. $\implies$ can factor the joint distribution of D.

Recall: Probability models

The **probability model** is a map from the parameter space to a set of distributions

$$\theta \to P_\theta$$

Recall: Probability models

The **probability model** is a map from the parameter space to a set of distributions

$$\theta \to P_\theta$$

Last section: We only assumed the data are i.i.d. But paid a price – only able to do inference using asymptotic approximations (which can be very poor in finite samples).

Recall: Probability models

The **probability model** is a map from the parameter space to a set of distributions

$$\theta \to P_\theta$$

Last section: We only assumed the data are i.i.d. But paid a price – only able to do inference using asymptotic approximations (which can be very poor in finite samples).

Bayesian Perspective: Specify a probability measure over Θ and exploit Bayes' Rule to perform inference – inference becomes conditional on the data.

> Let Π be a probability measure over Θ. This is **prior distribution**.

The probability model

So, the **state space** is

$$S = \Theta \times \mathcal{D} = \{(\theta, d) : \theta \in \Theta, d \in \mathcal{D}\}.$$

P_θ is a conditional distributional, $D|\Theta = \theta$.

The probability model

So, the **state space** is

$$S = \Theta \times \mathcal{D} = \{(\theta, d) : \theta \in \Theta, d \in \mathcal{D}\}.$$

P_θ is a conditional distributional, $D|\Theta = \theta$.

We can write the joint distribution.

Notation: Θ is a random variable, D is a random variable.

$$P(\theta \in B, D \in A) = \int_B P_\theta(A)\pi(\theta)d\theta.$$

The probability model

So, the **state space** is

$$S = \Theta \times \mathcal{D} = \{(\theta, d) : \theta \in \Theta, d \in \mathcal{D}\}.$$

P_θ is a conditional distributional, $D|\Theta = \theta$.

We can write the joint distribution.

Notation: Θ is a random variable, D is a random variable.

$$P(\theta \in B, D \in A) = \int_B P_\theta(A)\pi(\theta)d\theta.$$

Bayes' Rule:

$$P(\theta \in B | D \in A) = \int_B P_\theta(A)\pi(\theta)d\theta \Big/ \int_\Theta P_\theta(A)\pi(\theta)d\theta.$$

Bayes' Rule

We will perform inference using the **posterior distribution** of $\theta | D = d$.

> This encodes all our uncertainty about θ given that we observed the data $D = d$.

Typically write

$$\pi(\theta|d) \propto f(d|\theta)\pi(\theta),$$

where we omit a constant that makes the posterior integrate to one ($f(d)$).

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The data

Data are $X = (X_1, \ldots, X_n)$.

Conditional on θ, the X_i are i.i.d with

$$P(X_i = 1|\theta) = \theta, \quad P(X_i = 0|\theta) = 1 - \theta.$$

The parameter space is $\Theta = [0, 1]$.

Observe realizations $x = (x_1, \ldots, x_n)$.

The likelihood

The likelihood function is then

$$\begin{aligned} f_\theta(x) &= f(x|\theta) \\ &= P(X = x|\theta) \\ &= \Pi_{i=1}^{n} P(X_i = x_i|\theta) \\ &= \Pi_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{n_1}(1-\theta)^{n_0} \end{aligned}$$

where $n_1 = \sum_{i=1}^{n} y_i$ and $n_0 = \sum_{i=1}^{n}(1-y_i) = n - n_1$.

The likelihood

The likelihood function is then

$$\begin{aligned} f_\theta(x) &= f(x|\theta) \\ &= P(X = x|\theta) \\ &= \Pi_{i=1}^{n} P(X_i = x_i|\theta) \\ &= \Pi_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{n_1}(1-\theta)^{n_0} \end{aligned}$$

where $n_1 = \sum_{i=1}^{n} y_i$ and $n_0 = \sum_{i=1}^{n}(1-y_i) = n - n_1$.

n_1, n_0 are **sufficient statistics** for the likelihood function.

The prior

The prior distribution is a **beta distribution** with parameters $a, b > 0$.

Support is over $[0, 1]$ with density

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

Prior mean and variance are

$$E[\theta] = \frac{a}{a+b}, \quad V(\theta) = \frac{a}{a+b}\frac{b}{a+b}\frac{1}{a+b+1}.$$

The posterior

The posterior distribution is given by Bayes' rule.

$$\begin{aligned}\pi(\theta|x) &\propto f_\theta(x)\pi(\theta) \\ &\propto \theta^{a+n_1-1}(1-\theta)^{b+n_0-1}\end{aligned}$$

The posterior distribution is also a beta distribution with parameters $a+n_1, b+n_0$.

The posterior

The posterior mean is then

$$E[\theta|x] = \frac{a+n_1}{a+b+n} = \lambda\frac{n_1}{n} + (1-\lambda)\frac{a}{a+b}$$

where $\lambda = \frac{n}{a+b+n}$.

- The posterior mean is a convex combination of the sample mean n_1/n and the prior mean $a/(a+b)$.
- If $a+b$ is small relative to n, then most of the weight is placed on the sample mean.

The posterior

The posterior mean is then

$$E[\theta|x] = \frac{a+n_1}{a+b+n} = \lambda\frac{n_1}{n} + (1-\lambda)\frac{a}{a+b}$$

where $\lambda = \frac{n}{a+b+n}$.

- The posterior mean is a convex combination of the sample mean n_1/n and the prior mean $a/(a+b)$.
- If $a+b$ is small relative to n, then most of the weight is placed on the sample mean.

The posterior variance is

$$V(\theta|x) = \frac{E[\theta|x](1-E[\theta|x])}{n+a+b+1}.$$

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Credible Sets

We can use the posterior distribution to form **credible sets** – the Bayesian "equivalent" of a confidence interval.

Credible Sets

We can use the posterior distribution to form **credible sets** – the Bayesian "equivalent" of a confidence interval.

A $1-\alpha$ **credible set** $\Theta_{1-\alpha}$ satisfies

$$\int_{\Theta_{1-\alpha}} \pi(\theta|x)d\theta = 1-\alpha$$

It covers $1-\alpha\%$ of the mass of the posterior distribution.

- Any set that satisfies this is a credible interval.
- We will typically consder one that is symmetric around the mean.

Interpreting Credible Sets

What is the interpretation of this? How is it different from frequentist confidence intervals?

Interpreting Credible Sets

What is the interpretation of this? How is it different from frequentist confidence intervals?

> **Recall**: A $1-\alpha$ frequentist confidence interval is "if I randomly sampled my data and formed my confidence interval, the true parameter will be contained in the interval 95% of the time."

Interpreting Credible Sets

What is the interpretation of this? How is it different from frequentist confidence intervals?

> **Recall**: A $1 - \alpha$ frequentist confidence interval is "if I randomly sampled my data and formed my confidence interval, the true parameter will be contained in the interval 95% of the time."

Bayesian inference is **conditional on the data**.

> The credible interval states: Given the data I observed, there is a 95% probability that θ falls in this region.
>
> *These are different interpretations.*

Interpreting Credible Sets

What is the interpretation of this? How is it different from frequentist confidence intervals?

> **Recall**: A $1 - \alpha$ frequentist confidence interval is "if I randomly sampled my data and formed my confidence interval, the true parameter will be contained in the interval 95% of the time."

Bayesian inference is **conditional on the data**.

> The credible interval states: Given the data I observed, there is a 95% probability that θ falls in this region.
>
> *These are different interpretations.*

In Frequentist inference, the data are viewed as random and the parameter is fixed.

In Bayesian inference, the data are fixed and the parameter is random.

Improper priors

What happens as $a, b \to 0$? Prior becomes

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}.$$

Not a probability density as it integrates to ∞ over $[0, 1]$. Call this an **improper prior**.

But, the associated posterior distribution is well-defined.

- The posterior distribution is again a beta distribution but with parameters, n_1, n_0.
- Note

$$E[\theta|x] = \frac{n_1}{n} = \bar{x}$$

- That is, the posterior conditional expectation coincides with the sample average

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The data

Data are $D = (D_1, \ldots, D_n)$.

- Each D_i takes on discrete set of values $\{\alpha_j : j = 1, \ldots, J\}$.
- Conditional on θ, the D_i are i.i.d. with

$$P(D_i = \alpha_j | \theta) = \theta_j \quad \text{for} \quad j = 1, \ldots, J.$$

- Parameter space is the unit simplex on $\mathbb{R}^J$ with

$$\Theta = \{\theta \in \mathbb{R}^J : \theta_j \geq 0, \sum_{j=1}^{J} \theta_j = 1\}.$$

Observe realizations $d = (d_1, \ldots, d_n)$.

The data

The values of D_i may be vectors and we will apply these results to inference for the linear predictor.

Think of

$$D_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \quad \alpha_j = \begin{pmatrix} \alpha_{xj} \\ \alpha_{y}j \end{pmatrix}.$$

The likelihood

The likelihood function is

$$\begin{aligned} f_\theta(d) &= f(d|\theta) \\ &= \Pi_{i=1}^{n} P(D_i = d_i|\theta) \\ &= \Pi_{i=1}^{n} \Pi_{j=1}^{J} \theta_j^{1(d_i=\alpha_j)} \\ &= \Pi_{j=1}^{J} \theta_j^{n_j} \end{aligned}$$

where $n_j = \sum_{i=1}^{n} 1(d_i = \alpha_j)$ for $j = 1, \ldots, J$.

The likelihood

The likelihood function is

$$\begin{aligned} f_\theta(d) &= f(d|\theta) \\ &= \Pi_{i=1}^{n} P(D_i = d_i|\theta) \\ &= \Pi_{i=1}^{n} \Pi_{j=1}^{J} \theta_j^{1(d_i=\alpha_j)} \\ &= \Pi_{j=1}^{J} \theta_j^{n_j} \end{aligned}$$

where $n_j = \sum_{i=1}^{n} 1(d_i = \alpha_j)$ for $j = 1, \ldots, J$.

n_j for $j = 1, \ldots, J$ are **sufficient statistics** for the likelihood.

The prior

Prior distribution is a **Dirichlet distribution** with parameters $a_1, \ldots, a_J > 0$.

- Generalizes a generalization of the beta distribution.
- Its support is over the unit simplex in $\mathbb{R}^J$.
- Has density
$$\pi(u_1, \ldots, u_J) \propto \Pi_{j=1}^{J} u_j^{a_j - 1}.$$
for $u_j > 0, \sum_{j=1}^{J} u_j = 1$.

The posterior

The posterior distribution is given by Bayes' rule.

$$\begin{aligned}\pi(\theta|x) &\propto f_\theta(x)\pi(\theta)\\ &\propto \Pi_{j=1}^J \theta_j^{a_j+n_j-1}.\end{aligned}$$

The posterior distribution is also Dirichlet but with parameters $a_j + n_j$ for $j = 1, \ldots, J$.

The posterior

The posterior distribution is given by Bayes' rule.

$$\begin{aligned}\pi(\theta|x) &\propto f_\theta(x)\pi(\theta)\\ &\propto \Pi_{j=1}^{J}\theta_j^{a_j+n_j-1}.\end{aligned}$$

The posterior distribution is also Dirichlet but with parameters $a_j + n_j$ for $j = 1, \ldots, J$.

Can consider the improper prior with $a_j \to 0$ for each $j = 1, \ldots, J$.

With this improper prior, the posterior distribution remains Dirichlet and has parameters $n_1, \ldots, n_J$.

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Representing the Dirichlet Distribution

Recall: We can represent the Dirichlet distribution using Gamma-distributed random variables.

Let $Q_j \sim Gamma(a_j, 1)$. If $Q_1, \ldots, Q_J$ are independent then

$$(Q_1/\sum_{j=1}^{J} Q_j, \ldots, Q_J/\sum_{j=1}^{J} Q_j) \sim Dirichlet(a_1, \ldots, a_J).$$

Representing the Dirichlet Distribution

Recall: We can represent the Dirichlet distribution using Gamma-distributed random variables.

Let $Q_j \sim Gamma(a_j, 1)$. If $Q_1, \ldots, Q_J$ are independent then

$$(Q_1 / \sum_{j=1}^{J} Q_j, \ldots, Q_J / \sum_{j=1}^{J} Q_j) \sim Dirichlet(a_1, \ldots, a_J).$$

For case $J = 2$,

$$(Q_1/(Q_1 + Q_2), Q_2/(Q_1 + Q_2)) \sim Beta(a_1, a_2).$$

Representing the posterior

So, we can represent the posterior for θ as

$$\theta | d \sim (Q_1 / \sum_{j=1}^{J} Q_j, \ldots, Q_J / \sum_{j=1}^{J} Q_j),$$

where $Q_j \sim Gamma(n_j + a_j, 1)$ for $j = 1, \ldots J$.

Representing the posterior

So, we can represent the posterior for θ as

$$\theta | d \sim (Q_1 / \sum_{j=1}^{J} Q_j, \ldots, Q_J / \sum_{j=1}^{J} Q_j),$$

where $Q_j \sim Gamma(n_j + a_j, 1)$ for $j = 1, \ldots J$.

Moreover, each component θ_j has the representation

$$\theta_j | d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} = \beta(n_j + a_j, \sum_{k \neq j} n_k + a_k).$$

Representing the posterior

So, we can represent the posterior for θ as

$$\theta | d \sim (Q_1 / \sum_{j=1}^{J} Q_j, \ldots, Q_J / \sum_{j=1}^{J} Q_j),$$

where $Q_j \sim Gamma(n_j + a_j, 1)$ for $j = 1, \ldots J$.

Moreover, each component θ_j has the representation

$$\theta_j | d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} = \beta(n_j + a_j, \sum_{k \neq j} n_k + a_k).$$

So,

$$E[\theta_j | d] = \frac{n_j + a_j}{\sum_{k=1}^{J} n_k + a_k}.$$

Can similarly write $V(\theta_j | d)$ using formulas from before.

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Predictive distribution

Let's see what we can do with our posterior $\theta | d$.

Let's use it for prediction.

Predictive distribution

Let's see what we can do with our posterior $\theta|d$.

Let's use it for prediction.

Suppose there is a new observation D_{n+1}. We want to predict it.

Our object of interest is

$$\gamma = P(D_{n+1} \in A|\theta) = \sum_{j \in C} \theta_j$$

where $C = \{j : \alpha_j \in A\}$.

Predictive distribution

$P(D_{n+1} \in A|\theta) = \gamma(\theta)$ is just a function of θ. We derive its posterior distribution. That is,

$$\gamma(\theta)|d \sim ?$$

Predictive distribution

$P(D_{n+1} \in A|\theta) = \gamma(\theta)$ is just a function of θ. We derive its posterior distribution. That is,

$$\gamma(\theta)|d \sim ?$$

Turns out to be simple. Use the special case of $J = 2$ from earlier.

$$\theta_j|d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} \implies \sum_{j \in C} \theta_j \sim \frac{\sum_{j \in C} Q_j}{\sum_{k=1}^{J} Q_j}$$

where $\sum_{j \in C} Q_j \sim Gamma(\sum_{j \in C} n_j + a_j)$, $\sum_{j \notin C} Q_j \sim Gamma(\sum_{j \notin C} n_j + a_j)$.

Predictive distribution

So,

$$\gamma(\theta)|d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} \sim Beta(\sum_{j \in C} n_j + a_j, \sum_{j \notin C} n_j + a_j).$$

The conditional distribution of D_{n+1} given $(D_1, \ldots, D_n) = d$ is the **predictive distribution**.

Notice that θ has been integrated out using the posterior distribution.

We can use iterated expectations for this!

Predictive distribution

We have

$$
\begin{aligned}
P(D_{n+1} \in A|d) &= E[1(D_{n+1} \in A)|d] \\
&= E[E[1(D_{n+1} \in A)|\theta, d]|d] \\
&= E[E[1(D_{n+1} \in A)|\theta]|d] \\
&= E[\gamma(\theta)|d] \\
&= \frac{\sum_{j \in C} n_j + a_j}{\sum_{j=1}^{J} n_j + a_j}.
\end{aligned}
$$

Outline

Setting it all up

Beta-Bernoulli Model
- Credible Sets

Discrete-Dirichlet Model
- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

Approximating continuous distributions

We can use the discrete-dirichlet model to approximate continuous distributions by letting $J \to \infty$.

Approximating continuous distributions

We can use the discrete-dirichlet model to approximate continuous distributions by letting $J \to \infty$.

In doing so, we need to be careful that the prior does not become **dogmatic**.

Approximating continuous distributions

We can use the discrete-dirichlet model to approximate continuous distributions by letting $J \to \infty$.

In doing so, we need to be careful that the prior does not become **dogmatic**.

We want to ensure that the prior is "responsive to the data."

We want to ensure that the posterior doesn't just return the prior... that we actually learning from the data.

Approximating continuous distributions

We can use the discrete-dirichlet model to approximate continuous distributions by letting $J \to \infty$.

In doing so, we need to be careful that the prior does not become **dogmatic**.

We want to ensure that the prior is "responsive to the data."

We want to ensure that the posterior doesn't just return the prior... that we actually learning from the data.

Illustrate what can go wrong with the predictive distribution as $J \to \infty$ if we aren't careful.

Predictive distribution: $J \to \infty$

Recall:

$$\gamma = P(D_{n+1} \in A|d) = E[\gamma|w] = \frac{\sum_{j \in C} n_j + a_j}{\sum_{j=1}^{J} n_j + a_j}.$$

Suppose that $a_j = \epsilon > 0$ fixed for all j and let $J \to \infty$ while the data $d = (d_1, \ldots, d_n)$ is fixed.

Predictive distribution: $J \to \infty$

Recall:

$$\gamma = P(D_{n+1} \in A|d) = E[\gamma|w] = \frac{\sum_{j \in C} n_j + a_j}{\sum_{j=1}^{J} n_j + a_j}.$$

Suppose that $a_j = \epsilon > 0$ fixed for all j and let $J \to \infty$ while the data $d = (d_1, \ldots, d_n)$ is fixed.

Assume that

$$\frac{1}{J} \sum_{j \in C} 1 \to r$$

as $J \to \infty$. That is, the fraction of support points in A approaches r in the limit.

Predictive distribution: $J \to \infty$

Claim: Then,

$$P(D_{n+1} \in A|d) \to r.$$

Why? The prior is dogmatic for γ. The prior for γ is

$$\gamma \sim Beta(\sum_{j \in C} a_j, \sum_{j \notin C} a_j).$$

So,

$$E[\gamma] = \frac{\sum_{j \in C} a_j}{\sum_{j \notin C} a_j} = \frac{1}{J} \sum_{j \in C} 1 \to r,$$

$$V(\gamma) = \frac{E[\gamma](1 - E[\gamma])}{1 + \sum_{j=1}^{J} a_j} = \frac{E[\gamma](1 - E[\gamma])}{1 + \epsilon J} \to 0$$

As $J \to \infty$, for fixed ϵ, the prior distribution becomes concentrated around r.

What to do?

To avoid this, we let $a_j \to \infty$ for all j as $J \to \infty$.

In the limit, this produces the improper Dirichlet distribution.

So, if $n_j \geq 1$ for all j, the posterior will be a proper Dirichlet. But if we want J to approximate continuous distributions, we'll allow zero counts.

What to do?

To avoid this, we let $a_j \to \infty$ for all j as $J \to \infty$.

In the limit, this produces the improper Dirichlet distribution.

So, if $n_j \geq 1$ for all j, the posterior will be a proper Dirichlet. But if we want J to approximate continuous distributions, we'll allow zero counts.

So, if the count $n_k = 0$, the limiting posterior of θ_k as $J \to 0$ will become concentrated around 0.

$\implies$ Support points with $n_k = 0$ drop out of the posterior distribution, which is concentrated around support points with $n_j > 0$.

Why?

Gives us a way to take the tools we have (Discrete-Dirichlet) and apply it to continuous data.

- We use a limiting dirichlet prior and the posterior becomes concentrated around only support points on which we observe data.
- We'll next apply this to the linear predictor.