Conjugate Priors, Uninformative Priors

Nasim Zolaktaf

UBC Machine Learning Reading Group

January 2016

Outline

- Exponential Families
- Conjugacy
  - Conjugate priors
  - Mixture of conjugate prior
- Uninformative priors
  - Jeffreys prior

The Exponential Family

- A probability mass function (pmf) or probability distribution function (pdf) $p(X|\theta)$, for $X = (X_1, ..., X_m) \in \mathcal{X}^m$ and $\theta \subseteq R^d$, is said to be in the exponential family if it is the form:

$$p(X|\theta) = \frac{1}{Z(\theta)} h(X) \exp[\theta^T \phi(X)] \tag{1}$$

The Exponential Family

- A probability mass function (pmf) or probability distribution function (pdf) $p(X|\theta)$, for $X = (X_1, ..., X_m) \in \mathcal{X}^m$ and $\theta \subseteq R^d$, is said to be in the exponential family if it is the form:

$$p(X|\theta) = \frac{1}{Z(\theta)} h(X) \exp[\theta^T \phi(X)] \quad (1)$$

$$= h(X) \exp[\theta^T \phi(X) - A(\theta)] \quad (2)$$

The Exponential Family

- A probability mass function (pmf) or probability distribution function (pdf) $p(X|\theta)$, for $X = (X_1, ..., X_m) \in \mathcal{X}^m$ and $\theta \subseteq R^d$, is said to be in the exponential family if it is the form:

$$p(X|\theta) = \frac{1}{Z(\theta)} h(X) \exp[\theta^T \phi(X)] \quad (1)$$

$$= h(X) \exp[\theta^T \phi(X) - A(\theta)] \quad (2)$$

where

$$Z(\theta) = \int_{\mathcal{X}^m} h(X) \exp[\theta^T \phi(X)] dx \quad (3)$$

$$A(\theta) = \log Z(\theta) \quad (4)$$

The Exponential Family

- A probability mass function (pmf) or probability distribution function (pdf) $p(X|\theta)$, for $X = (X_1, ..., X_m) \in \mathcal{X}^m$ and $\theta \subseteq R^d$, is said to be in the exponential family if it is the form:

$$p(X|\theta) = \frac{1}{Z(\theta)} h(X) \exp[\theta^T \phi(X)] \tag{1}$$

$$= h(X) \exp[\theta^T \phi(X) - A(\theta)] \tag{2}$$

where

$$Z(\theta) = \int_{\mathcal{X}^m} h(X) \exp[\theta^T \phi(X)] dx \tag{3}$$

$$A(\theta) = \log Z(\theta) \tag{4}$$

- $Z(\theta)$ is called the partition function, $A(\theta)$ is called the log partition function or cumulant function, and $h(X)$ is a scaling constant.

The Exponential Family

- A probability mass function (pmf) or probability distribution function (pdf) $p(X|\theta)$, for $X = (X_1, ..., X_m) \in \mathcal{X}^m$ and $\theta \subseteq R^d$, is said to be in the exponential family if it is the form:

$$p(X|\theta) = \frac{1}{Z(\theta)} h(X) \exp[\theta^T \phi(X)] \tag{1}$$

$$= h(X) \exp[\theta^T \phi(X) - A(\theta)] \tag{2}$$

where

$$Z(\theta) = \int_{\mathcal{X}^m} h(X) \exp[\theta^T \phi(X)] dx \tag{3}$$

$$A(\theta) = \log Z(\theta) \tag{4}$$

- $Z(\theta)$ is called the partition function, $A(\theta)$ is called the log partition function or cumulant function, and $h(X)$ is a scaling constant.
- Equation 2 can be generalized by writing

$$p(X|\theta) = h(X) \exp[\eta(\theta)^T \phi(X) - A(\eta(\theta))] \tag{5}$$

Binomial Distribution

- As an example of a discrete exponential family, consider the Binomial distribution with known number of trials n. The pmf for this distribution is

$$p(x|\theta) = \text{Binomial}(n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x \in \{0, 1, .., n\} \quad (6)$$

Binomial Distribution

- As an example of a discrete exponential family, consider the Binomial distribution with known number of trials n. The pmf for this distribution is

$$p(x|\theta) = \text{Binomial}(n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, x \in \{0, 1, .., n\} \quad (6)$$

- This can equivalently be written as

$$p(x|\theta) = \binom{n}{x} \exp(x \log(\frac{\theta}{1-\theta}) + n \log(1-\theta)) \quad (7)$$

which shows that the Binomial distribution is an exponential family, whose natural parameter is

$$\eta = \log \frac{\theta}{1-\theta} \quad (8)$$

Conjugacy

- Consider the posterior distribution $p(\theta|X)$ with prior $p(\theta)$ and likelihood function $p(x|\theta)$, where $p(\theta|X) \propto p(X|\theta)p(\theta)$.

Conjugacy

- Consider the posterior distribution $p(\theta|X)$ with prior $p(\theta)$ and likelihood function $p(x|\theta)$, where $p(\theta|X) \propto p(X|\theta)p(\theta)$.
- If the posterior distribution $p(\theta|X)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(X|\theta)$.

Conjugacy

- Consider the posterior distribution $p(\theta|X)$ with prior $p(\theta)$ and likelihood function $p(x|\theta)$, where $p(\theta|X) \propto p(X|\theta)p(\theta)$.
- If the posterior distribution $p(\theta|X)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(X|\theta)$.
- All members of the exponential family have conjugate priors.

Brief List of Conjugate Models

Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative Binomial	Beta	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (mean unknown)	Normal	Normal
Normal (variance unknown)	Inverse Gamma	Inverse Gamma
Normal (mean and variance unknown)	Normal/Gamma	Normal/Gamma
Multinomial	Dirichlet	Dirichlet

The Conjugate Beta Prior

- The Beta distribution is conjugate to the Binomial distribution.

$$
\begin{aligned}
& p(\theta|x) = p(x|\theta)p(\theta) = \text{Binomial}(n,\theta) * \text{Beta}(a,b) = \\
& \binom{n}{x}\theta^{x}(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{b-1} \\
& \propto \theta^{x}(1-\theta)^{n-x}\theta^{(a-1)}(1-\theta)^{b-1}
\end{aligned} \tag{9}
$$

The Conjugate Beta Prior

- The Beta distribution is conjugate to the Binomial distribution.

$$p(\theta|x) = p(x|\theta)p(\theta) = \text{Binomial}(n,\theta) * \text{Beta}(a,b) =$$
$$\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{b-1} \tag{9}$$
$$\propto \theta^{x}(1-\theta)^{n-x}\theta^{(a-1)}(1-\theta)^{b-1}$$

$$p(\theta|x) \propto \theta^{(x+a-1)}(1-\theta)^{n-x+b-1} \tag{10}$$

The Conjugate Beta Prior

- The Beta distribution is conjugate to the Binomial distribution.

$$p(\theta|x) = p(x|\theta)p(\theta) = \text{Binomial}(n,\theta) * \text{Beta}(a,b) =$$
$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{b-1} \tag{9}$$
$$\propto \theta^x(1-\theta)^{n-x}\theta^{(a-1)}(1-\theta)^{b-1}$$

$$p(\theta|x) \propto \theta^{(x+a-1)}(1-\theta)^{n-x+b-1} \tag{10}$$

- The posterior distribution is simply a Beta$(x+a, n-x+b)$ distribution.
- Effectively, our prior is just adding $a-1$ successes and $b-1$ failures to the dataset.

Coin Flipping Example: Model

- Use a Bernoulli likelihood for coin X landing 'heads',

$$p(X = `H'|\theta) = \theta, \quad p(X = `T'|\theta) = 1 - \theta,$$
$$p(X|\theta) = \theta^{\mathcal{I}(X = `H')}(1 - \theta)^{\mathcal{I}(X = `T')}.$$

Coin Flipping Example: Model

- Use a Bernoulli likelihood for coin X landing 'heads',

$$p(X = \text{`}H' | \theta) = \theta, \quad p(X = \text{`}T' | \theta) = 1 - \theta,$$
$$p(X|\theta) = \theta^{\mathcal{I}(X=\text{`}H')}(1-\theta)^{\mathcal{I}(X=\text{`}T')}.$$

- Use a Beta prior for probability θ of 'heads', $\theta \sim \mathrm{Beta}\,(a, b)$,

$$p(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$$

with $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Coin Flipping Example: Model

- Use a Bernoulli likelihood for coin X landing 'heads',

$$p(X = \text{`}H'|\theta) = \theta, \quad p(X = \text{`}T'|\theta) = 1 - \theta,$$
$$p(X|\theta) = \theta^{\mathcal{I}(X=\text{`}H')}(1-\theta)^{\mathcal{I}(X=\text{`}T')}.$$

- Use a Beta prior for probability θ of 'heads', $\theta \sim \text{Beta}\,(a, b)$,

$$p(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$$

with $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

- Remember that probabilities sum to one so we have

$$1 = \int_0^1 p(\theta|a,b)d\theta = \int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}d\theta = \frac{1}{B(a,b)}\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$$

Coin Flipping Example: Model

- Use a Bernoulli likelihood for coin X landing 'heads',

$$p(X = \text{`}H' | \theta) = \theta, \quad p(X = \text{`}T' | \theta) = 1 - \theta,$$

$$p(X|\theta) = \theta^{\mathcal{I}(X = \text{`}H')}(1-\theta)^{\mathcal{I}(X = \text{`}T')}.$$

- Use a Beta prior for probability θ of 'heads', $\theta \sim \text{Beta}\,(a, b)$,

$$p(\theta|a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a, b)} \propto \theta^{a-1}(1-\theta)^{b-1}$$

with $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

- Remember that probabilities sum to one so we have

$$1 = \int_0^1 p(\theta|a, b)d\theta = \int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a, b)}d\theta = \frac{1}{B(a, b)}\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$$

which helps us compute integrals since we have

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = B(a, b).$$

Coin Flipping Example: Posterior

- Our model is:

$$X \sim \mathrm{Ber}(\theta), \quad \theta \sim \mathrm{Beta}\,(a, b).$$

Coin Flipping Example: Posterior

- Our model is:

$$X \sim \text{Ber}(\theta), \quad \theta \sim \text{Beta}\,(a, b).$$

- If we observe 'HHH' then our posterior distribution is

$$\begin{aligned}
p(\theta|HHH) &= \frac{p(HHH|\theta)p(\theta)}{p(HHH)} && \text{(Bayes' rule)}\\
&\propto p(HHH|\theta)p(\theta) && (p(HHH) \text{ is constant})\\
&= \theta^3(1-\theta)^0 p(\theta) && \text{(likelihood def'n)}\\
&= \theta^3(1-\theta)^0 \theta^{a-1}(1-\theta)^{b-1} && \text{(prior def'n)}\\
&= \theta^{(3+a)-1}(1-\theta)^{b-1}.
\end{aligned}$$

Coin Flipping Example: Posterior

- Our model is:

$$X \sim \text{Ber}(\theta), \quad \theta \sim \text{Beta}\,(a, b).$$

- If we observe 'HHH' then our posterior distribution is

$$\begin{aligned}
p(\theta|HHH) &= \frac{p(HHH|\theta)p(\theta)}{p(HHH)} && \text{(Bayes' rule)} \\
&\propto p(HHH|\theta)p(\theta) && \text{($p(HHH)$ is constant)} \\
&= \theta^3(1-\theta)^0 p(\theta) && \text{(likelihood def'n)} \\
&= \theta^3(1-\theta)^0 \theta^{a-1}(1-\theta)^{b-1} && \text{(prior def'n)} \\
&= \theta^{(3+a)-1}(1-\theta)^{b-1}.
\end{aligned}$$

- Which we've written in the form of a Beta distribution,

$$\theta \mid HHH \sim \text{Beta}(3+a, b),$$

which let's us skip computing the integral $p(HHH)$.

Coin Flipping Example: Estimates

If we observe ‘HHH’ with Beta$(1, 1)$ prior, then

$\theta \mid HHH \sim \text{Beta}\,(3 + a, b)$ and our different estimates are:

Coin Flipping Example: Estimates

If we observe 'HHH' with Beta$(1, 1)$ prior, then

$\theta \mid HHH \sim \text{Beta}\,(3 + a, b)$ and our different estimates are:

- Maximum likelihood:

$$\hat{\theta} = \frac{n_H}{n} = \frac{3}{3} = 1.$$

Coin Flipping Example: Estimates

If we observe 'HHH' with Beta$(1, 1)$ prior, then

$\theta \mid HHH \sim \text{Beta}\,(3 + a, b)$ and our different estimates are:

- Maximum likelihood:

$$\hat{\theta} = \frac{n_H}{n} = \frac{3}{3} = 1.$$

- MAP with uniform,

$$\hat{\theta} = \frac{(3 + a) - 1}{(3 + a) + b - 2} = \frac{3}{3} = 1.$$

Coin Flipping Example: Estimates

If we observe 'HHH' with Beta$(1,1)$ prior, then

$\theta \mid HHH \sim \text{Beta}\,(3+a, b)$ and our different estimates are:

- Maximum likelihood:

$$\hat{\theta} = \frac{n_H}{n} = \frac{3}{3} = 1.$$

- MAP with uniform,

$$\hat{\theta} = \frac{(3+a)-1}{(3+a)+b-2} = \frac{3}{3} = 1.$$

- Posterior predictive,

$$\begin{aligned} p(H|HHH) &= \int_0^1 p(H|\theta)p(\theta|HHH)d\theta \\ &= \int_0^1 \text{Ber}(H|\theta)\text{Beta}\,(\theta|3+a,b)d\theta \\ &= \int_0^1 \theta\text{Beta}(\theta|3+a,b)d\theta = \mathbb{E}[\theta] \\ &= \frac{(3+a)}{(3+a)+b} = \frac{4}{5} = 0.8. \end{aligned}$$

Coin Flipping Example: Effect of Prior

- We assume all θ equally likely and saw HHH,
    - ML/MAP predict it will always land heads.
    - Bayes predict probability of landing heads is only 80%.
        - Takes into account other ways that HHH could happen.

Coin Flipping Example: Effect of Prior

- We assume all θ equally likely and saw HHH,
  - ML/MAP predict it will always land heads.
  - Bayes predict probability of landing heads is only 80%.
    - Takes into account other ways that HHH could happen.
- Beta$(1, 1)$ prior is like seeing HT or TH in our mind before we flip,
  - Posterior predictive would be $\frac{3+1}{3+1+1} = 0.80$.

Coin Flipping Example: Effect of Prior

- We assume all θ equally likely and saw HHH,
  - ML/MAP predict it will always land heads.
  - Bayes predict probability of landing heads is only 80%.
    - Takes into account other ways that HHH could happen.
- Beta$(1, 1)$ prior is like seeing HT or TH in our mind <u>before we flip</u>,
  - Posterior predictive would be $\frac{3+1}{3+1+1} = 0.80$.
- Beta$(3, 3)$ prior is like seeing 3 heads and 3 tails (stronger uniform prior),
  - Posterior predictive would be $\frac{3+3}{3+3+3} = 0.667$.
- Beta$(100, 1)$ prior is like seeing 100 heads and 1 tail (biased),
  - Posterior predictive would be $\frac{3+100}{3+100+1} = 0.990$.
- Beta$(0.01, 0.01)$ biases towards having unfair coin (head or tail),
  - Posterior predictive would be $\frac{3+0.01}{3+0.01+0.01} = 0.997$.

Coin Flipping Example: Effect of Prior

- We assume all θ equally likely and saw HHH,
  - ML/MAP predict it will always land heads.
  - Bayes predict probability of landing heads is only 80%.
    - Takes into account other ways that HHH could happen.
- Beta$(1, 1)$ prior is like seeing HT or TH in our mind before we flip,
  - Posterior predictive would be $\frac{3+1}{3+1+1} = 0.80$.
- Beta$(3, 3)$ prior is like seeing 3 heads and 3 tails (stronger uniform prior),
  - Posterior predictive would be $\frac{3+3}{3+3+3} = 0.667$.
- Beta$(100, 1)$ prior is like seeing 100 heads and 1 tail (biased),
  - Posterior predictive would be $\frac{3+100}{3+100+1} = 0.990$.
- Beta$(0.01, 0.01)$ biases towards having unfair coin (head or tail),
  - Posterior predictive would be $\frac{3+0.01}{3+0.01+0.01} = 0.997$.
- Dependence on (a, b) is where people get uncomfortable:
  - But basically the same as choosing regularization parameter λ.
  - If your prior knowledge isn't misleading, you will not overfit.

Mixtures of Conjugate Priors

- A mixture of conjugate priors is also conjugate.
- We can use a mixture of conjugate priors as a prior.

Mixtures of Conjugate Priors

- A mixture of conjugate priors is also conjugate.
- We can use a mixture of conjugate priors as a prior.
- For example, suppose we are modelling coin tosses, and we think the coin is either fair, or is biased towards heads. This cannot be represented by a Beta distribution. However, we can model it using a mixture of two Beta distributions. For example, we might use:

$$p(\theta) = 0.5\,\text{Beta}(\theta|20, 20) + 0.5\,\text{Beta}(\theta|30, 10) \tag{11}$$

Mixtures of Conjugate Priors

- A mixture of conjugate priors is also conjugate.
- We can use a mixture of conjugate priors as a prior.
- For example, suppose we are modelling coin tosses, and we think the coin is either fair, or is biased towards heads. This cannot be represented by a Beta distribution. However, we can model it using a mixture of two Beta distributions. For example, we might use:

$$p(\theta) = 0.5\,\text{Beta}(\theta|20,20) + 0.5\,\text{Beta}(\theta|30,10) \tag{11}$$

- If θ comes from the first distribution, the coin is fair, but if it comes from the second it is biased towards heads.

Mixtures of Conjugate Priors (Cont.)

- The prior has the form

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k) \tag{12}$$

where $z = k$ means that θ comes from mixture component k, $p(z = k)$ are called the prior mixing weights, and each $p(\theta|z = k)$ is conjugate.

Mixtures of Conjugate Priors (Cont.)

- The prior has the form

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k) \tag{12}$$

where $z = k$ means that θ comes from mixture component k, $p(z=k)$ are called the prior mixing weights, and each $p(\theta|z=k)$ is conjugate.

- Posterior can also be written as a mixture of conjugate distributions as follows:

$$p(\theta|X) = \sum_k p(z=k|X)p(\theta|X, z=k) \tag{13}$$

where $p(z=k|X)$ are the posterior mixing weights given by

$$p(z=k|X) = \frac{p(z=k)p(X|z=k)}{\sum_{k'} p(z=k'|X)p(\theta|X, z=k')} \tag{14}$$

Uninformative Priors

- If we don't have strong beliefs about what θ should be, it is common to use an uninformative or non-informative prior, and to let the data speak for itself.
- Designing uninformative priors is tricky.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0, 1)?$.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0, 1)?$.
  - This is equivalent to Beta$(1, 1)$ on θ.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0, 1)?$.
  - This is equivalent to Beta$(1, 1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0,1)?$.
  - This is equivalent to Beta$(1,1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0,1)?$.
  - This is equivalent to Beta$(1,1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!
- By decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By this argument, the most uninformative prior is Beta$(0,0)$, which is called Haldane prior.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0, 1)?$.
  - This is equivalent to Beta$(1, 1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!
- By decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By this argument, the most uninformative prior is Beta$(0, 0)$, which is called Haldane prior.
  - Haldane prior is an improper prior; it does not integrate to 1.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0, 1)?$.
  - This is equivalent to Beta$(1, 1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!
- By decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By this argument, the most uninformative prior is Beta$(0, 0)$, which is called Haldane prior.
  - Haldane prior is an improper prior; it does not integrate to 1.
  - Haldane prior results in the posterior Beta$(x, n-x)$ which will be proper as long as $n - x \neq 0$ and $x \neq 0$.

Uninformative Prior for the Bernoulli

- Consider the Bernoulli parameter

$$p(x|\theta) = \theta^x (1-\theta)^{n-x} \tag{15}$$

- What is the most uninformative prior for this distribution?
- The uniform distribution $Uniform(0,1)?$.
  - This is equivalent to Beta$(1,1)$ on θ.
  - We can predict the MLE is $\frac{N_1}{N_1+N_0}$, whereas the posterior mean is $E[\theta|X] = \frac{N_1+1}{N_1+N_0+2}$. Prior isn't completely uninformative!
- By decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By this argument, the most uninformative prior is Beta$(0,0)$, which is called Haldane prior.
  - Haldane prior is an improper prior; it does not integrate to 1.
  - Haldane prior results in the posterior Beta$(x, n-x)$ which will be proper as long as $n-x \neq 0$ and $x \neq 0$.
- We will see that the "right" uninformative prior is Beta$(\frac{1}{2}, \frac{1}{2})$.

Jeffreys Prior

- Jeffrey argued that a uninformative prior should be invariant to the parametrization used. The key observation is that if $p(\theta)$ is uninformative, then any reparametrization of the prior, such as $\theta = h(\phi)$ for some function h, should also be uninformative.

Jeffreys Prior

- Jeffrey argued that a uninformative prior should be invariant to the parametrization used. The key observation is that if $p(\theta)$ is uninformative, then any reparametrization of the prior, such as $\theta = h(\phi)$ for some function h, should also be uninformative.
- Jeffreys prior is the prior that satisfies $p(\theta) \propto \sqrt{det I(\theta)}$, where $I(\theta)$ is the Fisher information for θ, and is invariant under reparametrization of the parameter vector θ.

Jeffreys Prior

- Jeffrey argued that a uninformative prior should be invariant to the parametrization used. The key observation is that if $p(\theta)$ is uninformative, then any reparametrization of the prior, such as $\theta = h(\phi)$ for some function h, should also be uninformative.
- Jeffreys prior is the prior that satisfies $p(\theta) \propto \sqrt{det I(\theta)}$, where $I(\theta)$ is the Fisher information for θ, and is invariant under reparametrization of the parameter vector θ.
- The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ which the probability of X depends.

$$I(\theta) = E[(\frac{\partial}{\partial\theta} \log f(X;\theta))^2 |\theta] \tag{16}$$

Jeffreys Prior

- Jeffrey argued that a uninformative prior should be invariant to the parametrization used. The key observation is that if $p(\theta)$ is uninformative, then any reparametrization of the prior, such as $\theta = h(\phi)$ for some function h, should also be uninformative.
- Jeffreys prior is the prior that satisfies $p(\theta) \propto \sqrt{det I(\theta)}$, where $I(\theta)$ is the Fisher information for θ, and is invariant under reparametrization of the parameter vector θ.
- The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ which the probability of X depends.

$$I(\theta) = E[(\frac{\partial}{\partial\theta} \log f(X;\theta))^2 |\theta] \tag{16}$$

- If $\log f(X;\theta)$ is twice differentiable with respect to θ, then

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)|\theta] \tag{17}$$

Reparametrization for Jeffreys Prior: One parameter case

- For an alternative parametrization ϕ we can derive $p_(\phi) = \sqrt{I(\phi)}$ from $p(\theta) = \sqrt{I(\theta)}$, using the change of variables theorem and the definition of Fisher information:

$$
\begin{aligned}
p(\phi) = p(\theta)|\frac{d\theta}{d\phi}| \propto \sqrt{I(\theta)(\frac{d\theta}{d\phi})^2} = \sqrt{E[(\frac{d\ln L}{d\theta})^2](\frac{d\theta}{d\phi})^2} \\
= \sqrt{E[(\frac{d\ln L}{d\theta}\frac{d\theta}{d\phi})^2]} = \sqrt{E[(\frac{d\ln L}{d\phi})^2]} = \sqrt{I(\phi)}
\end{aligned}
\tag{18}
$$

Jeffreys Prior for the Bernoulli

- Suppose $X \sim Ber(\theta)$. The log-likelihood for a single sample is

$$\log p(X|\theta) = X\ log\theta + (1 - X)\log(1 - \theta) \tag{19}$$

Jeffreys Prior for the Bernoulli

- Suppose $X \sim Ber(\theta)$. The log-likelihood for a single sample is

$$\log p(X|\theta) = X\ log\theta + (1 - X)\log(1 - \theta) \tag{19}$$

- The score function is gradient of log-likelihood

$$s(\theta) = \frac{d}{d\theta}\log(pX|\theta) = \frac{X}{\theta} - \frac{1 - X}{1 - \theta} \tag{20}$$

Jeffreys Prior for the Bernoulli

- Suppose $X \sim Ber(\theta)$. The log-likelihood for a single sample is

$$\log p(X|\theta) = X\ log\theta + (1 - X)\log(1 - \theta) \tag{19}$$

- The score function is gradient of log-likelihood

$$s(\theta) = \frac{d}{d\theta}\log(pX|\theta) = \frac{X}{\theta} - \frac{1 - X}{1 - \theta} \tag{20}$$

- The observed information is

$$J(\theta) = -\frac{d^2}{d\theta^2}\log p(X|\theta) = -s'(\theta|X) = \frac{X}{\theta^2} + \frac{1 - X}{(1 - \theta)^2} \tag{21}$$

Jeffreys Prior for the Bernoulli

- Suppose $X \sim Ber(\theta)$. The log-likelihood for a single sample is

$$\log p(X|\theta) = X\ log\theta + (1 - X)\log(1 - \theta) \tag{19}$$

- The score function is gradient of log-likelihood

$$s(\theta) = \frac{d}{d\theta}\log(pX|\theta) = \frac{X}{\theta} - \frac{1 - X}{1 - \theta} \tag{20}$$

- The observed information is

$$J(\theta) = -\frac{d^2}{d\theta^2}\log p(X|\theta) = -s'(\theta|X) = \frac{X}{\theta^2} + \frac{1 - X}{(1 - \theta)^2} \tag{21}$$

- The Fisher information is the expected information

$$I(\theta) = E[J(\theta|X)|X \sim \theta] = \frac{\theta}{\theta^2} + \frac{1 - \theta}{(1 - \theta)^2} = \frac{1}{\theta(1 - \theta)} \tag{22}$$

- Hence Jeffreys prior is

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1 - \theta)^{-\frac{1}{2}} \propto \text{Beta}(\frac{1}{2}, \frac{1}{2}). \tag{23}$$

Selected Related Work

[1] Kevin P Murphy(2012)

Machine learning: a probabilistic perspective

MIT press

[2] Jarad Niemi

Conjugacy of prior distributions:

https://www.youtube.com/watch?v=yhewYFqGjFA

[3] Jarad Niemi

Noninformative prior distributions:

https://www.youtube.com/watch?v=25-PpMSrAGM

[4]

Fisher information:

https://en.wikipedia.org/wiki/Fisher_information