CS 5135/6035 Learning Probabilistic Models

Lecture 16: Fisher Information and Jeffreys' Prior

Gowtham Atluri

October 25, 2018

Reading Material

- Ly et al, A Tutorial on Fisher Information
  https://arxiv.org/pdf/1705.01064.pdf
- Surya Tokdar, The Jeffreys Prior
  https://www2.stat.duke.edu/courses/Fall11/sta114/jeffreys.pdf

Topics

- Sufficient Statistic
- Fisher Information
- Issue with noninformative prior
  - Varies with transformations
- Jeffreys Prior
  - Invariant under transformations

Sufficient Statistic

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

- $T(y)$ is a 'sufficient statistic' for θ
  - A sufficient statistic for θ contains all the information in the sample about θ
  - We cannot improve our knowledge about $\boldsymbol{\theta}$ by a detailed analysis of data $y_1, \ldots, y_n$

Questions:

- What is a statistic?
- What is a sufficient statistic?
- What is the amount of information captured in a sufficient statistic?

Statistic

- Statistic is a function of the data

$$T = f(\{y_1, y_2, \ldots, y_n\})$$

- Alternatively, it can be treated as a *data summary*
- Consider the coin toss example:
  - $y = \{$H T H H H H T H H T$\}$
  - $y = \{1\ 0\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 0\}$
- The number of trials that resulted in heads is a summary of the data

$$S(y) = \sum_{i=1}^{n} y_i$$

- Let us introduce a new random variable $s = S(y)$
- In the coin toss example: $s = S(y) = \sum_{i=1}^{n} y_i = 7$

Statistic

- Probability function for the original random variable y

$$p(y|\theta) = \theta^y (1-\theta)^{1-y}$$

- Probability function for the new random variable that is a summary s

$$p(s|\theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

- Notice that both y and s are governed by θ

Statistic

$$p(y|\theta) = \theta^y(1-\theta)^{1-y} \qquad p(s|\theta) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$$

- Number of possible outcomes
  - y has 2^n possible outcomes (coin toss: e.g., $2^{10} = 1024$)
  - s has $n+1$ possible outcomes ($\{0, 1, \dots, n\}$) (coin toss: e.g., $10 + 1 = 11$)
- There is reduction in #outcomes from y to s
  - s ignores the order with which the data are collected
  - $\binom{n}{s}$ accounts for all possible sequences of length n that consist of s heads and $n - s$ tails.
  - Coin toss example: $n = 10$ and $s = 7$ there are $\binom{10}{7} = 120$ possible sequence of 0s and 1s that contain 7 heads.
- Conditional probability of the raw data y given s is

$$p(y|s,\theta) = \frac{1}{\binom{n}{s}}$$

Sufficient Statistic

$$p(y|\theta) = \theta^y(1-\theta)^{1-y} \qquad p(s|\theta) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$$

- Conditional probability of the raw data y given s is

$$p(y|s,\theta) = \frac{1}{\binom{n}{s}}$$

- This does not depend on θ
  - Even though y and s separately depend on θ
  - y is conditionally independent of θ, given s
- There in no information about θ left in data y, after observing summary statistic s
- We say a summary statistic, $s = S(y)$, a sufficient statistic
  - if the expression $p(y|s,\theta)$ does not depend on θ
- Advantage: we can discard non-informative pieces of the dataset.
- How do we quantify the amount of information?

Fisher Information

- Fisher Information is denoted as $\mathcal{I}_y(\theta)$ or simply $\mathcal{I}(\theta)$
  - It is the Fisher Info. of a random variable y about θ

$$\mathcal{I}_y(\theta) = -\sum_y \Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big]p(y|\theta) \quad \text{when } y \text{ is discrete}$$

$$\mathcal{I}_y(\theta) = -\int_y \Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big]p(y|\theta) \quad \text{when } y \text{ is continuous}$$

- Interpretation: It measures the overall sensitivity at each potential outcome y w.r.t. chance defined by $p(y|\theta)$
  - Weighting w.r.t $p(y|\theta)$, implies that Fisher Info. is also an expectation
- It is also expressed as

$$\mathcal{I}_y(\theta) = -\mathbb{E}_{y|\theta}\Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big]$$

More on Fisher Information

$$\mathcal{I}_y(\theta) = -\sum_y \Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big]p(y|\theta)$$

- Fisher information within data $y = \{y_1, y_2, \dots, y_n\}$ about θ
  - is calculated by replacing $p(y|\theta)$ with $p(y_1, y_2, \dots, y_n|\theta)$
- With the assumption that samples in y are drawn i.i.d.,

$$\mathcal{I}_{\{y_1,y_2,\dots,y_n\}}(\theta) = n\mathcal{I}_y(\theta)$$

- For this reason, $\mathcal{I}_y(\theta)$ is also known as the unit Fisher Information
- Intuitively, an experiment consisting of $n = 10$ trails is twice as informative about θ compared to an experiement of only $n = 5$ trails

More on Fisher Information

$$\mathcal{I}_y(\theta) = -\sum_y \Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big]p(y|\theta)$$

- We cannot expect an aribitrary summary statistic s to extract more information about θ than what is available in the data y

$$\mathcal{I}_y(\theta) \geq I_s(\theta)$$

- When s is a sufficient statistic, i.e., $p(y|s,\theta)$ does not depend on theta

$$\mathcal{I}_y(\theta) = I_s(\theta)$$

Example: Fisher Information for y

- In the coin toss experiment, y is a Bernoulli random variable

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

- Fisher information:

$$\begin{aligned}
\mathcal{I}_y(\theta) &= -\mathbb{E}_y\Big[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\Big] \\
&= -\mathbb{E}_{y|\theta}\Big[\frac{\partial^2}{\partial\theta^2}\log(\theta^y(1-\theta)^{1-y})\Big] \\
&= -\mathbb{E}_{y|\theta}\Big[\frac{\partial^2}{\partial\theta^2}(y\log\theta + (1-y)\log(1-\theta))\Big] \\
&= \mathbb{E}_{y|\theta}\Big[\frac{y}{\theta^2} + \frac{1-y}{(1-\theta)^2}\Big] \quad \text{(we known } \mathbb{E}(y) = \theta) \\
&= \Big[\frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2}\Big] = \frac{1}{\theta(1-\theta)}
\end{aligned}$$

For n independent Bernoulli trials $\mathcal{I}_y(\theta) = n/\theta(1-\theta)$

Example: Fisher Information for summary statistic s

- In the coin toss experiment, $s = S(y) = \sum_{i=1}^{n} y_i$

$$p(s|\theta) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$$

- Fisher information:

$$\begin{aligned}
\mathcal{I}_s(\theta) &= -E_{s|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(s|\theta)\right] \\
&= -E_{s|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log\binom{n}{s} + s\log\theta + (n-s)\log(1-\theta)\right] \\
&= -E_{s|\theta}\left[-\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}\right] \quad \text{(we known } \mathbb{E}(s) = n\theta) \\
&= -\left[-\frac{n\theta}{\theta^2} - \frac{n-n\theta}{(1-\theta)^2}\right] = \frac{n}{\theta} + \frac{n}{(1-\theta)} \\
&= \frac{n}{\theta(1-\theta)}
\end{aligned}$$

Revisiting Sufficient Statistic

- y is a random variable that represents result of a coin toss
- s is a summary statistic $s = S(y) = \sum_{i=1}^{n} y_i$

$$\mathcal{I}_y(\theta) = \frac{n}{\theta(1-\theta)} \qquad \mathcal{I}_s(\theta) = \frac{n}{\theta(1-\theta)}$$

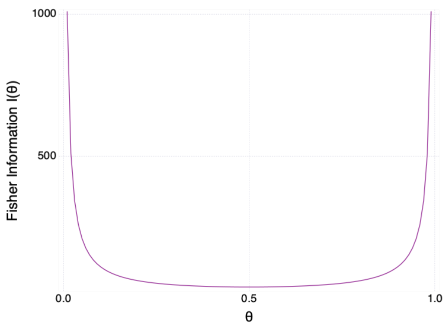
- When s is a sufficient statistic, i.e., $p(y|s,\theta)$ does not depend on theta

$$\mathcal{I}_y(\theta) = I_s(\theta)$$

- Hence, $S(y) = \sum_{i=1}^{n} y_i$ is a sufficient statistic for θ

Visualizing Fisher Information

$$\mathcal{I}_y(\theta) = \frac{n}{\theta(1-\theta)} \qquad \mathcal{I}_s(\theta) = \frac{n}{\theta(1-\theta)}$$

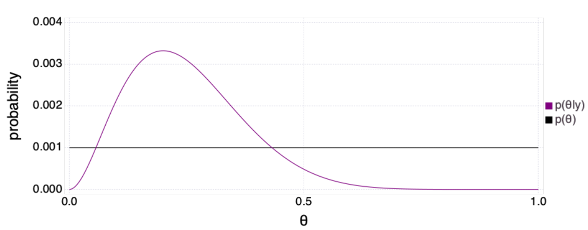


Fisher Information in Bayesian Statistic

- Issues with uniform prior
  - Undesirable consequences
  - Varies with transformation
- Jeffreys' prior
  - Invariant with transformation

Uniform Distribution as a non-informative prior

- In Bayesian parameter estimation
  - when little is known about θ that governs the outcome of y
  - it may seem reasonable to express this ignorance with a uniform prior
- Coin toss example
  - Prior distribution is $p(\theta) = 1$
  - Likelihood is $\theta^{N_H}(1-\theta)^{N_T}$
  - Posterior $p(\theta|y_1,\ldots,y_n) = \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T}$



Different representations, different conclusions

- The value of θ is related to the angle ϕ with which the coin is bent
  - ϕ takes values in the interval $(-\pi, \pi)$
- Assume the relation between angle ϕ and probability of heads θ is given by

$$\theta = h(\phi) = \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3 \qquad \phi = h^{-1}(\theta) = \pi\sqrt[3]{2(\theta - 1/2)}$$

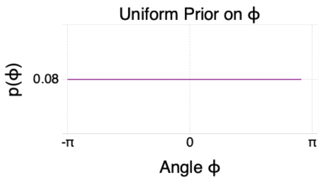
- If we have no prior preference over θ, should we have any preference over ϕ?
  - or vice versa?
- We will show that when we use a uniform prior over ϕ, it suggests a non-uniform prior over θ!

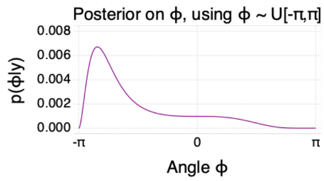
Different representations, different conclusions

$$\theta = h(\phi) = \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3 \qquad \phi = h^{-1}(\theta) = \pi\sqrt[3]{2(\theta - 1/2)}$$

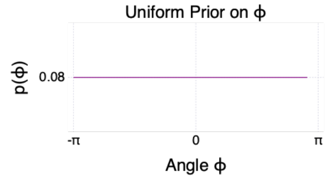
- Let us define a uniform Prior over ϕ: $p(\phi) \propto 1$
- We know the likelihood for θ is $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$
- Posterior $\propto$ likelihood $\times$ prior

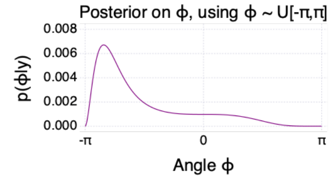
$$p(\phi|y) \propto p(y|\phi)p(\phi) = \left(\frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3\right)^{N_H}\left(1 - \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3\right)^{N_T}$$





Different representations, different conclusions





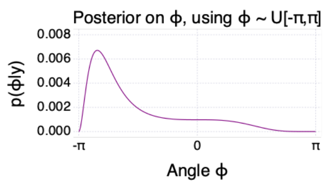
- Determining the prior on θ using the prior on ϕ

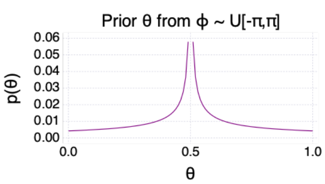
$$p(\theta) = p(\phi)\left|\frac{d\phi}{d\theta}\right| \propto \frac{2\pi}{3(\theta - 1/2)^{2/3}} \quad (\text{as } p(\phi) \propto 1)$$

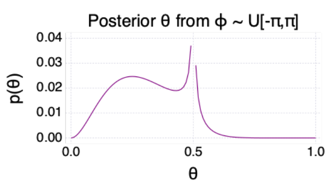
- Writing the posterior on θ: $p(\theta|y) \propto p(y|\theta)p(\theta)$
  - We know $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$

$$p(\theta|y) \propto \theta^{N_H}(1-\theta)^{N_T} \times \frac{2\pi}{3(\theta - 1/2)^{2/3}}$$

Different representations, different conclusions







- A uniform prior on ϕ leads to a highly informative prior on θ.
  - Posterior knowledge gained is also different

A True Non-informative Prior: Jeffreys' Prior

- If we had no preference for values of ϕ, then we should have no preference for values of θ.
  - This was held as a major criticism against Bayesian inference in the early 20^{th} century by R. A. Fisher and others.
  - Until H Jeffreys revived this topic in mid 20th century
- Jeffreys proposed that an acceptable non-informative prior should invariant under monotone transformations of the parameter.
- Jeffreys described how to construct such a prior

Jeffreys' Prior: $p(\theta) \propto \sqrt{\mathcal{I}_y(\theta)}$ (where $\mathcal{I}_y(\theta)$ is Fisher information)

Jeffreys' Prior: Coin toss example

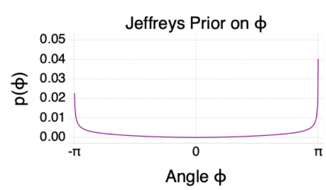
Jeffreys Prior: $p(\theta) \propto \sqrt{\mathcal{I}_y(\theta)}$ (where $\mathcal{I}_y(\theta)$ is Fisher information)

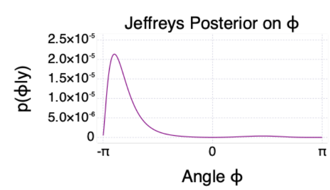
$$\theta = h(\phi) = \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3 \qquad \phi = h^{-1}(\theta) = \pi\sqrt[3]{2(\theta - 1/2)}$$

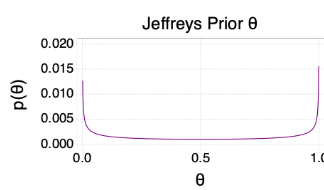
Determining Jeffreys Prior for θ and ϕ.

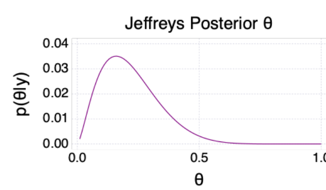
- We know $\mathcal{I}(\theta) = n/\theta(1-\theta)$
- Jeffreys prior $p(\theta) \propto \sqrt{\mathcal{I}_y(\theta)} = \sqrt{n/\theta(1-\theta)} \propto 1/\sqrt{\pi^6 - \phi^6}$
- Jeffreys prior $p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \propto 3\phi^2/\sqrt{\pi^6 - \phi^6}$

Using Jeffreys' Prior









- Prior info. is same for both ϕ and θ
  - Same for Posteriors $p(\phi|y)$ and $p(\theta|y)$

Summary

- Uniform prior for ϕ transforms into highly informative prior for θ
  - Affects the posterior too!
  - Different posterior knowledge gained from ϕ and θ
- Jeffreys' prior has the same form for ϕ and θ
- Same conclusions about θ are draw regardless of whether we
  1. Use Jeffreys' prior on θ and update with the observed data
  2. Use Jeffreys' prior on ϕ and update to a posterior on ϕ and then transform it to a posterior on θ.
- Jeffreys' prior leads to the same posterior knowledge regardless of how we as a researcher present the problem
- Best to use Jeffreys' prior as a non-informative prior

Jeffreys Prior: $p(\theta) \propto \sqrt{\mathcal{I}_y(\theta)}$ (where $\mathcal{I}_y(\theta)$ is Fisher information)