AN EXAMPLE OF CONJUGATE PRIORS

Dinh Huu Nguyen, 10/15/2020

Abstract: an example of conjugate priors in Bayesian modeling.

Contents

1. Overview

- suppose you have a dataset of n observations $x_1, \dots, x_n$
- suppose you have chosen to model them as observations from a distribution $X(\Theta)$ with some parameter Θ

Question: how to choose Θ?

Answer: depends on approach.

1.1. **Randomist approach.** Treat Θ as a number θ and pick a random one.

1.2. **Frequentist approach.** Treat Θ as a number θ and find one that does something. A popular target is θ that maximizes the likelihood of observing $x_1, \ldots, x_n$. Such θ is called the maximum likelihood estimate.

1.3. **Bayesian approach.** Treat Θ as a distribution $\Theta(\alpha)$ with some parameter α and update α to $x_1, \ldots, x_n$ via Bayes' theorem

$$p_{\Theta \mid \alpha, x_1, \ldots, x_n}(\theta) = \frac{p_{X \mid \Theta}(x_1, \ldots, x_n) p_\Theta(\theta)}{p_X(x_1, \ldots, x_n)} \tag{1}$$

- $p_X(x_1, \ldots, x_n)$ is called the evidence
- $p_{X \mid \Theta}(x_1, \ldots, x_n)$ is called the likelihood
- $p_\Theta(\theta)$ is called the prior probability
- $p_{\Theta \mid x_1, \ldots, x_n}(\theta)$ is called the posterior probability

After having chosen model $X(\Theta)$ above, people often choose $\Theta(\alpha)$ in some family $\mathcal{F}$ of distributions so that

1. the evidence $p_X(x) = \int p_{X \mid \Theta}(x) p_\Theta(\theta) d\theta$ is tractable
2. the predictive probability $p_{X \mid \Theta, x_1, \ldots, x_n}(x) = \int p_{X \mid \Theta}(x) p_{\Theta \mid x_1, \ldots, x_n}(\theta) d\theta$ is tractable
3. the posterior distribution $\Theta \mid x_1, \ldots, x_n$ is also in $\mathcal{F}$. Hence the name "conjugate priors".

Goals 1, 2 and goal 3 pose a dilemma:

- if we choose Θ to be in the family $\mathcal{C}$ of all constant distributions then integrals are easy to compute but $\Theta \,|\, x_1, \ldots, x_n$ likely will not be in $\mathcal{C}$.
- if we choose Θ to be in the family $\mathcal{A}$ of all distributions then $\Theta \,|\, x_1, \ldots, x_n$ surely is in $\mathcal{A}$ but integrals are hard to compute

However, if X is an exponential distribution and Θ is chosen to be in the family $\mathcal{E}$ of all exponential distributions then this dilemma is solved:

- product of two exponential distributions is another exponential distribution $e^a e^b = e^{a+b}$
- integrals of exponentials $\int e^a$ are tractable

2. Examples

2.1. Bernoulli likelihood.

- suppose you have a dataset of n observations $x_1 = 0, x_2 = 1, x_3 = 1, \ldots, x_n = 0$
- suppose you have chosen to model them as observations from a Bernoulli distribution $X(\Theta)$ with some parameter Θ and probability mass function

$$\begin{aligned} p_{X\,|\,\Theta}(1) &= \Theta \\ p_{X\,|\,\Theta}(0) &= 1 - \Theta \\ p_{X\,|\,\Theta}(x) &= \Theta^x (1-\Theta)^{1-x} \end{aligned} \tag{2}$$

If we assume $x_1, \ldots, x_n$ are independent then the likelihood in (1) is

$$\begin{aligned} p_{X\,|\,\Theta}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} p_{X\,|\,\Theta}(x_i) \\ &= \prod_{i=1}^{n} \Theta^{x_i} (1-\Theta)^{1-x_i} \end{aligned} \tag{3}$$

2.1.1. *Randomist approach.* Treat Θ as a number θ and pick $\theta = 0.5$.

2.1.2. *Frequentist approach.* Treat Θ as a number θ and find one that maximizes above likelihood

$$p_{X\,|\,\theta}(x_1,\ldots,x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

or equivalently maximizes log of above likelihood

$$\begin{aligned}
L(\theta) &= log(p_{X\,|\,\theta}(x_1,\ldots,x_n)) \\
&= log(\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{n} log(\theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{n} log(\theta^{x_i}) + \sum_{i=1}^{n} log((1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{n} x_i log(\theta) + \sum_{i=1}^{n}(1-x_i) log(1-\theta) \\
&= log(\theta)\sum_{i=1}^{n} x_i + log(1-\theta)n - log(1-\theta)\sum_{i=1}^{n} x_i
\end{aligned}$$

$$\begin{aligned}
L'(\theta) &= \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{1-\theta} + \frac{1}{1-\theta}\sum_{i=1}^{n} x_i \\
&= \frac{\sum_{i=1}^{n} x_i - n\theta}{\theta(1-\theta)}
\end{aligned}$$

It follows that $L'(\theta)$ is 0 and $L(\theta)$ is maximum when θ is $\frac{\sum_{i=1}^{n} x_i}{n}$. This θ is called the maximum likelihood estimate.

Prediction of observations is simple

$$\begin{aligned} p_{X\,|\,\theta}(1) &= \theta \\ &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

$$\begin{aligned} p_{X\,|\,\theta}(0) &= 1 - \theta \\ &= 1 - \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

2.1.3. *Bayesian approach.* Treat $\Theta(\alpha_1, \alpha_2)$ as a distribution with some parameters α_1, α_2 and probability density function

$$p_\Theta(\theta) \propto \theta^{\alpha_1}(1-\theta)^{\alpha_2}$$

and update α_1, α_2 to $x_1, \ldots, x_n$ via Bayes' theorem 1.

That p_Θ looks like $p_{X\,|\,\Theta}$ is by choice.

And after a change of variables $\alpha_1 = \beta_1 - 1, \alpha_2 = \beta_2 - 1$ and normalization

$$p_\Theta(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_2-1}}{B(\beta_1, \beta_2)} \tag{4}$$

where normalizing factor $B(\beta_1, \beta_2) = \frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}$ is the beta function and Γ is the gamma function, we recognize that this choice Θ is the beta distribution $Beta(\beta_1, \beta_2)$ with mean and variance

$$E(\Theta) = \frac{\beta_1}{\beta_1 + \beta_2}$$

$$var(\Theta) = \frac{\beta_1\beta_2}{(\beta_1+\beta_2+1)(\beta_1+\beta_2)^2}$$

Now we are ready to update β_1, β_2 with $x_1, \ldots, x_n$ via (1) using (2) and (4)

$$\begin{aligned} p_{\Theta \mid x_1,\ldots,x_n}(\theta) &\propto p_{X \mid \Theta}(x_1, \ldots, x_n) p_\Theta(\theta) \\ &= \left(\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right) \theta^{\beta_1 - 1} (1-\theta)^{\beta_2 - 1} \\ &= \theta^{\beta_1 - 1 + \sum\limits_{i=1}^{n} x_i} (1-\theta)^{\beta_2 - 1 + n - \sum\limits_{i=1}^{n} x_i} \end{aligned}$$

We recognize that $\Theta \mid x_1, \ldots, x_n$ is another beta distribution $Beta(\beta_1 + \sum\limits_{i=1}^{n} x_i, \beta_2 + n - \sum\limits_{i=1}^{n} x_i)$ with mean and variance

$$\begin{aligned} E(\Theta \mid x_1, \ldots, x_n) &= \frac{\beta_1 + \sum\limits_{i=1}^{n} x_i}{\beta_1 + \sum\limits_{i=1}^{n} x_i + \beta_2 + n - \sum\limits_{i=1}^{n} x_i} \\ &= \frac{\beta_1 + \sum\limits_{i=1}^{n} x_i}{\beta_1 + \beta_2 + n} \\ &= \frac{\beta_1}{\beta_1 + \beta_2 + n} + \frac{\sum\limits_{i=1}^{n} x_i}{\beta_1 + \beta_2 + n} \end{aligned}$$

$$var(\Theta \mid x_1, \ldots, x_n) = \frac{(\beta_1 + \sum\limits_{i=1}^{n} x_i)(\beta_2 + n - \sum\limits_{i=1}^{n} x_i)}{(\beta_1 + \beta_2 + n + 1)(\beta_1 + \beta_2 + n)^2}$$

One can sees two nice things that hint at reconciliation between frequentist approach and Bayesian approach

1. $E(\Theta \,|\, x_1, \ldots, x_n)$ goes to the maximum likelihood estimate $\frac{\sum_{i=1}^{n} x_i}{n}$ as the number of observations n goes to ∞

2. $var(\Theta \,|\, x_1, \ldots, x_n)$ goes to 0 as the number of observations goes to ∞, hence $\Theta \,|\, x_1, \ldots, x_n$ is concentrated around the maximum likelihood estimate

One can see another nice thing when $E(\Theta \mid x_1, \ldots, x_n)$ is written as the following convex sum

$$E(\Theta \mid x_1, \ldots, x_n) = \left(\frac{\beta_1 + \beta_2}{\beta_1 + \beta_2 + n}\right) \frac{\beta_1}{\beta_1 + \beta_2} + \left(1 - \frac{\beta_1 + \beta_2}{\beta_1 + \beta_2 + n}\right) \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= aE(\Theta) + (1 - a)\bar{x}$$

$$= a \text{ prior belief } + (1 - a) \text{ present reality}$$

which goes to present reality as n goes to ∞.

Updating to the next observation x_{n+1} is straightforward

$$
\begin{aligned}
\Theta \,|\, x_{,} \ldots, x_n, x_{n+1} &\text{ is } Beta(\beta_1 + \sum_{i=1}^{n+1} x_i, \beta_2 + n + 1 - \sum_{i=1}^{n+1} x_i) \\
&\text{ is } Beta(\beta_1 + \sum_{i=1}^{n} x_i + x_{n+1}, \beta_2 + n - \sum_{i=1}^{n} x_i + 1 - x_{n+1})
\end{aligned}
$$

Prediction of observations is closed-form

$$p_{X \mid \Theta, x_1, \ldots, x_n}(1) = \frac{\beta_1 + \sum_{i=1}^{n} x_i}{\beta_1 + \beta_2 + n}$$

$$p_{X \mid \Theta, x_1, \ldots, x_n}(0) = 1 - p_{X \mid \Theta, x_1, \ldots, x_n}(1)$$

2.2. **Multinoulli likelihood.** similar holds

2.3. **Poisson likelihood.** similar holds

2.4. **Gaussian likelihood.** similar holds

2.5. **Exponential likelihood.** similar holds for any likelihood X whose probability mass function or probability density function has this canonical form

$$p_X(x) = h(x)e^{\langle \Theta, T(x) \rangle - A(\Theta)}$$

where h, T, A are functions.

- Θ is called canonical parameter
- $T(x)$ is called sufficient statistic
- $A(\Theta)$ is called cumulant function

Exercise 2.1. Show that Bernoulli distribution is an exponential distribution.

Remark 2.2. As we treat $X(\Theta)$ as a distribution with parameter Θ and treat $\Theta(\alpha)$ as a distribution with parameter α, we could continue to treat $\alpha(\beta)$ as a distribution with parameter and so on. But at some point, we have to stop and treat the parameter as a number. Similarly in math, complex theorems are derived from theorems, and theorems are derived from simpler theorems, and so on. But at some point, simplest theorems are derived from axioms, that are taken to be true and serve as building blocks.

References