Bayesian Inference: Intro via Conjugacy

June 8, 2021

Bayes' Rule

$$\underbrace{p(\theta|x)}_{posterior} = \frac{\underbrace{p(x|\theta)}_{likelihood}\ \underbrace{p(\theta)}_{prior}}{\underbrace{p(x)}_{evidence}} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,d\theta}$$

Posterior

The posterior distribution is proportional to the prior times the likelihood:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

The posterior distribution *is a distribution* over θ.

Evidence

The evidence, or *marginal likelihood*, can be used for model comparison.

Example: Beta-bernoulli model

Sometimes, we can compute the posterior distribution by hand, given prior and likelihood.

Setup: flipping a coin

Probability that it lands heads is (unknown) θ.
Prior probability over θ assumed to follow a $Beta(3, 3)$ distribution:

$$p(\theta) = \frac{\theta^{3-1}(1-\theta)^{3-1}}{B(3,3)}$$

Note: $\theta \sim Beta(a, b)$ means $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$
Will collect data by flipping coin once. Likelihood of observing heads ($x = 1$) or tails ($x = 0$) is given by a Bernoulli distribution:

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

.

Example: Beta-bernoulli model

Setup: flipping a coin

Probability that it lands heads is (unknown) θ.

Prior probability over θ assumed to follow a $Beta(3,3)$ distribution:

$$p(\theta) = \frac{\theta^{3-1}(1-\theta)^{3-1}}{B(3,3)}$$

Note: $\theta \sim Beta(a,b)$ means $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$

Will collect data by flipping coin once. Likelihood of observing heads ($x=1$) or tails ($x=0$) is given by a Bernoulli distribution:

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

.

Computing the posterior after observing x=1

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \theta^1(1-\theta)^0\theta^2(1-\theta)^2 = \theta^3(1-\theta)^2 \implies \theta|x \sim Beta(4,3)$$

Conjugacy

The idea

We have conjugacy when the prior and the posterior distributions are in the same family (e.g. in the previous example, the prior and posterior are beta distributions).

Definition

Conjugacy can be defined as follows (*gelman2013bayesian*) . If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

Technical note: the condition trivially holds if $\mathcal{P}$ is taken to be the space of all probability distributions!

Posterior predictive distribution

Given

$p(\theta|x)$ - posterior
$p(\theta)$ - prior
$p(x|\theta)$ - likelihood

Posterior predictive distribution

Consider the probability of new data x'. Posterior predictive distribution is:

$$p(x'|x) = \int p(x', \theta|x)\, d\theta = \int p(x'|\theta, x) p(\theta|x)\, d\theta = \int p(x'|\theta) p(\theta|x)\, d\theta$$

Incorporates the knowledge and uncertainty about θ that we still had after seeing data x.

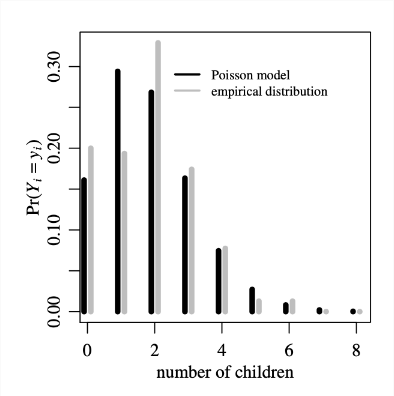
Another example: The Gamma-Poisson model

Some measurements, such as a person's number of children or number of friends, have values that are whole numbers. Perhaps the simplest probability model on whole numbers is the Poisson model.

- **Sample**: X the observed number.
- **Sample space**: $\mathcal{X} = \{0, 1, 2, ...\}$
- **Density:**

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad (1)$$

- **Parameter:** $\lambda = \mathbb{E}[X]$
- **Parameter space**: $\lambda \in (0, \infty)$



A Poisson distribution with mean 1.83, along with the empirical distribution of the number of children of women of age 40 from the GSS during the 1990's

The Gamma-Poisson model

The Gamma distribution is a conjugate prior for the Poisson likelihood. Can you show this?

The Gamma-Poisson model

The Gamma distribution is a conjugate prior for the Poisson likelihood. Can you show this?

$$
\begin{aligned}
p(\lambda) &\propto \lambda^{\alpha-1} e^{-\beta\lambda} \\
p(\boldsymbol{x} \mid \lambda) &\overset{iid}{=} \prod_{i=1}^{n} p(x_i \mid \lambda) \overset{(1)}{\propto} \lambda^{\sum_i x_i} e^{-n\lambda}
\end{aligned}
$$

so

$$
\begin{aligned}
p(\lambda \mid \boldsymbol{x}) &\propto p(\boldsymbol{x} \mid \lambda) p(\lambda) \\
&= \lambda^{(\alpha + \sum_i x_i) - 1} e^{-(\beta + n)\lambda}
\end{aligned}
$$

The Gamma-Poisson model

That is,

$$
\begin{aligned}
\lambda &\sim \mathrm{Gamma}(\alpha, \beta) \\
x_i \mid \lambda &\overset{\text{iid}}{\sim} \mathrm{Poisson}(\lambda) \\
\implies \lambda \mid \boldsymbol{x} &\sim \mathrm{Gamma}(\alpha + \sum_{i=1}^{n} x_i, \beta + n)
\end{aligned}
$$

Interpretation

Can you write the posterior expectation as a compromise between the prior expectation and sample mean (like we did for the Beta-bernoulli model)?

Interpretation

Can you write the posterior expectation as a compromise between the prior expectation and sample mean (like we did for the Beta-bernoulli model)?

$$\mathbb{E}[\lambda] = \frac{\alpha}{\beta} \qquad \text{Gamma dist.}$$

$$\mathbb{E}[\lambda \mid \boldsymbol{x}] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n} \qquad \text{Gamma dist.}$$

$$= \frac{\beta}{\beta + n} \underbrace{\frac{\alpha}{\beta}}_{\text{prior expectation}} + \frac{n}{\beta + n} \underbrace{\frac{\sum_{i=1}^{n} x_i}{n}}_{\text{sample mean}}$$

Interpretation?

Interpretation

Can you write the posterior expectation as a compromise between the prior expectation and sample mean (like we did for the Beta-bernoulli model)?

$$\mathbb{E}[\lambda] = \frac{\alpha}{\beta} \qquad \text{Gamma dist.}$$

$$\mathbb{E}[\lambda \mid \boldsymbol{x}] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n} \qquad \text{Gamma dist.}$$

$$= \frac{\beta}{\beta + n} \underbrace{\frac{\alpha}{\beta}}_{\text{prior expectation}} + \frac{n}{\beta + n} \underbrace{\frac{\sum_{i=1}^{n} x_i}{n}}_{\text{sample mean}}$$

Interpretation?

- β: number of prior observations
- α: sum of counts from β prior observations

Tractability notes

Generally, computing the posterior distribution is much harder than in this example!

Consider the denominator in $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)}\,d\theta$ - integrals are hard

In nonconjugate examples, we need approaches to work with the posterior distribution when we cannot calculate it directly. Stay tuned!