STA 602 – Intro to Bayesian Statistics

Lecture 4

Li Ma

The general applicability of Bayes' Theorem

Recall the two pieces we need to apply Bayes' Theorem

1. The distribution of the data *conditional* on the state of nature $p(x|\theta)$. (The sampling model.)
2. The *prior* distribution $p(\theta)$ on the state of nature θ.

Bayes' Theorem provides a probabilistic recipe for inference through learning the *posterior* distribution of the state of nature given the data, $p(\theta|x)$.

- We have seen a particular example where $p(\theta)$ is Beta(α, β) and $p(x|\theta)$ is Binomial.
- This scheme for inference applies generally to *all* choices of $p(x|\theta)$ and $p(\theta)$.
- Let us now look at another example.

Example: Measuring air quality with an imperfect device

Suppose we want to measure the amount (in terms of density) of a certain pollutant in the air.

- We have a device that is accurate "on average".
- However it may in any single reading be off by an unpredictable amount, due to factors such as temperature, humidity, and the way the device is held, etc.

In formal terms, let

- θ be the actual amount of the pollutant in the air.
- X be a reading of the device, which given θ is a random quantity that is close but not exactly θ.

Goal of inference: What is θ?

The Bayesian procedure

Let us carry out a Bayesian inference just as we did for the political poll example. Again, we need to specify the two pieces:

1. The *sampling model*—that is, the conditional distribution of the data given the parameter (or state of nature)—$p(x|\theta)$.
2. A *prior* distribution $p(\theta)$ for the state of nature that summarizes our knowledge about θ before observing data.

What may be a good model for the reading of the device given θ?

- We can *model* the error made by the device on each reading as normally distributed with standard deviation σ. For simplicity, let us assume that this σ is known. The user's manual gives the technical specs of the device including its "precision". For example, $\sigma = 2$. Under this model, the conditional density for $X = x$ given θ is

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \quad \text{for } -\infty < x < \infty.$$

Question: What if the value of σ is unknown? How can we carry out a Bayesian inference? To be exact here our model is $p(x|\theta,\sigma)$. (More on this later.)

Choosing the prior distribution for θ

$p(x|\theta) \sim N(\theta, \sigma^2)$
$p(\theta) \sim N(\mu_0, \tau_0^2)$

We can *model* our prior knowledge of the actual amount θ also as a Normal(μ_0, τ_0^2) distribution.

- The mean of the prior, μ_0, can be chosen as the historical average amount, e.g. $\mu_0 = 100$.
- We can choose τ_0 so that the historical records of the amount fall in the range of $\mu_0 \pm \tau_0$ about 2/3 of the times.

For example, if historically the amount of the pollutant falls in the range 100 ± 10 about 2/3 of the time, then we may choose

$$\mu_0 = 100 \quad \text{and} \quad \tau_0 = 10.$$

Now that we have the two pieces, inference again is an application of Bayes' theorem.

$$
\begin{aligned}
p(\theta|x) &\propto p(\theta)p(x|\theta) \\
&= \boxed{\frac{1}{\sqrt{2\pi}\tau_0}e^{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}} \times \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}}} \quad \text{as a function of } \theta \\
&= \frac{1}{2\cancel{\pi}\tau_0\sigma}e^{-[\frac{(\theta-\mu_0)^2}{2\tau_0^2}+\frac{(x-\theta)^2}{2\sigma^2}]} \\
&\propto e^{-\frac{1}{2}[\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{(x-\theta)^2}{\sigma^2}]}. \quad \exp\left\{-\frac{1}{2}(A\theta^2-2B\theta+C)\right\}
\end{aligned}
$$

What is this distribution?

Note that

$$
\begin{aligned}
&\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{(x-\theta)^2}{\sigma^2}\\
&=\theta^2\underbrace{\left(\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}\right)}_{A}-2\theta\underbrace{\left(\frac{\mu_0}{\tau_0^2}+\frac{x}{\sigma^2}\right)}_{B}+\underbrace{\frac{\mu_0^2}{\tau_0^2}+\frac{x^2}{\sigma^2}}_{C}\\
&=A\theta^2-2B\theta+C\\
&=A\left(\theta^2-2B/A\cdot\theta\right)+C\\
&=A(\theta-B/A)^2+C'\\
&=\frac{(\theta-\mu_1)^2}{\tau_1^2}+C'
\end{aligned}
$$

precision = prior precision + sample precision.

↓

where $\omega_1 = \frac{1}{\tau_0^2}$, $\omega_2 = \frac{1}{\sigma^2}$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

$$\mu_1=\frac{B}{A}=\frac{\boxed{\mu_0}/\tau_0^2+\boxed{x}/\sigma^2}{1/\tau_0^2+1/\sigma^2} \quad \text{and} \quad \tau_1^2=\frac{1}{A}=\frac{1}{1/\tau_0^2+1/\sigma^2}.$$

and C' is a term that doesn't involve θ.

Question: If x is far away from prior (way off scale), then $\tau_1 \downarrow$.
however counterintuitive (which is $\tau_1 \uparrow$).

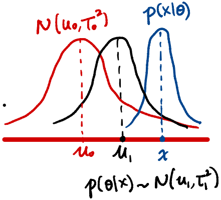
$\Rightarrow$ math correct,
assumption wrong.

Thus

$$p(\theta|x) \propto e^{-\frac{(\theta-\mu_1)^2}{2\tau_1^2}} \quad \text{for } -\infty < \theta < \infty.$$

This is the same as the p.d.f of a Normal(μ_1, τ_1^2) distribution up to a normalizing constant. Therefore we must have

$$p(\theta|x) = \frac{1}{\sqrt{2\pi}\tau_1} e^{-\frac{(\theta-\mu_1)^2}{2\tau_1^2}} \quad \text{for } -\infty < \theta < \infty.$$



and the posterior expectation and variance are

$$\mathrm{E}(\theta|X=x) = \mu_1 = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + 1/\sigma^2}\right)\mu_0 + \left(\frac{1/\sigma^2}{1/\tau_0^2 + 1/\sigma^2}\right)x$$

and

$$\mathrm{Var}(\theta|X=x) = \tau_1^2 = \frac{1}{1/\tau_0^2 + 1/\sigma^2}.$$

$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$, $\quad \tau_1^2 \le \min\{\tau_0^2, \sigma^2\}$

fisher information.
average curvature of likelihood

Note that the posterior expectation is again a weighted average of the prior expectation μ and the observed data x.

$$\mathrm{E}(\theta|X=x) = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + 1/\sigma^2}\right)\mu + \left(\frac{1/\sigma^2}{1/\tau_0^2 + 1/\sigma^2}\right)x.$$

The weights are determined by the relative sizes of $1/\tau_0^2$ (which measures the amount of information about θ in the prior) and $1/\sigma^2$ (which measures the amount of information about θ in the data).

- For fixed σ^2, as $\tau_0^2 \downarrow 0$, we have $\mathrm{E}(\theta|X=x) \to \mu$ and $\mathrm{Var}(\theta|X=x) \downarrow 0$. We are *a priori* “certain” that $\theta = \mu$.
- For fixed τ_0^2, as $\sigma^2 \downarrow 0$, we have $\mathrm{E}(\theta|X=x) \to x$ and $\mathrm{Var}(\theta|X=x) \downarrow 0$. We have a “perfect” device.

The posterior variance

$$\operatorname{Var}(\theta|X=x)=\tau_1^2=\frac{1}{1/\tau_0^2+1/\sigma^2}$$

is smaller than both τ_0^2 and σ^2. We can think of $1/\tau_1^2$ as a measure of the information we have about θ. The total informaion is the sum of prior information and the information from data:

$$\frac{1}{\tau_1^2}=\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}.$$

Update iteratively, $\frac{1}{\tau_n^2}=\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}.$

For our ongoing example

Suppose we observe $X = 105$ and we know that $\sigma = 2$, then

$$\mathrm{E}(\theta|X=105) = \frac{1/10^2}{1/10^2+1/2^2}\cdot 100 + \frac{1/2^2}{1/10^2+1/2^2}\cdot 105 = 104.8$$

$$\mathrm{Var}(\theta|X=105) = 1/(1/10^2+1/2^2) = 3.85.$$

If instead $\tau_0 = 1$, (so we are a lot more certain *a priori*),

$$\mathrm{E}(\theta|X=105) = \frac{1/1^2}{1/1^2+1/2^2}\cdot 100 + \frac{1/2^2}{1/1^2+1/2^2}\cdot 105 = 101$$

$$\mathrm{Var}(\theta|X=105) = 1/(1/1^2+1/2^2) = 0.89.$$

Question: If we observe an observation that's very far, e.g., 110 or even 120. Why doesn't our posterior variance increase but still decrease?

Check.

Food-for-thought: What's the predictive distribution for an observation? Does it support such values?

Bayes' Theorem for multiple independent observations

- Now suppose instead of taking one reading from the device, we take n *independent* readings $X_1, X_2, \ldots, X_n$. That is, conditional on θ, the X_i's are *independent identitically distributed* (i.i.d.) $N(\theta, \sigma^2)$ random variables. That is

$$p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}}.$$

- Then Bayes' Theorem applies just as before:

$$\begin{aligned} p(\theta|x_1, x_2, \ldots, x_n) &= \frac{p(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{\int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n|\theta)p(\theta)d\theta} \\ &\propto p(x_1, x_2, \ldots, x_n|\theta)p(\theta) \\ &= p(\theta)\prod_{i=1}^{n} p(x_i|\theta). \end{aligned}$$

Exercise

If we still use $N(\mu_0, \tau_0^2)$ as the prior for the parameter θ, show that given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the posterior distribution for θ is $N(\mu_n, \tau_n^2)$ with

$$\mu_n = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} \right) \mu_0 + \left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \right) \bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

is the sample average, and

$$\tau_n^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}.$$

More examples of inference tasks

- What is the probability that $\theta > 102$?
- What is the 95% HPD credible interval for θ?
- What is the predictive distribution for a future reading?

$$
\begin{aligned}
p(x_{n+1} \mid \mathbf{x}_n \cancel{, \theta}) &= \int \frac{1}{\sigma} \phi \left(\frac{x_{n+1} - \theta}{\sigma} \right) p(\theta \mid \mathbf{x}_n) d\theta \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{n+1}-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau_n} e^{-\frac{(\theta-\mu_n)^2}{2\tau_n^2}} d\theta \\
&= \cdots .
\end{aligned}
$$

$$= \int p(x_{n+1} | x_n, \theta) \cdot p(\theta | x_n) \, d\theta$$

$$= \int p(x_{n+1} | \theta) \cdot p(\theta | x_n) \, d\theta$$

- Exercise: Compute this integral. Hint: use the following fact

$$\int \frac{1}{\sqrt{2\pi B}} e^{-\frac{(u-A)^2}{2B}} du = 1 \qquad \text{for any real number } A \text{ and } B > 0.$$

A shortcut

$X_1,\dots,X_n, X_{n+1}|\theta \overset{i.i.d}{\sim} N(\theta,\sigma^2)$.

$X_{n+1}=\theta+\varepsilon_{n+1}$

- Note that we can write each observation

$$X_i = \theta + \varepsilon_i$$

where $\varepsilon_i \sim_{iid} \mathrm{N}(0,\sigma^2)$.

- Then given $\mathbf{X}_n$,

$$\theta \,|\, \mathbf{X}_n = \mathbf{x}_n \sim \mathrm{N}(\mu_n, \tau_n^2)$$

and $\varepsilon_{n+1} \,|\, \mathbf{X}_n \sim \mathrm{N}(0,\sigma^2)$. ε_{n+1} has nothing to do with $\underline{X_n}$, $\varepsilon_1,\dots,\varepsilon_n$.

- Therefore, given $\mathbf{X}_n$, the posterior distribution of $X_{n+1} = \theta + \varepsilon_{n+1}$ is

$\theta \,|\, X_n \perp\!\!\!\perp \varepsilon_{n+1} \,|\, X_n$.

$$\theta + \varepsilon_{n+1} \,|\, \mathbf{X}_n = \mathbf{x}_n \sim \mathrm{N}(\mu_n, \tau_n^2 + \sigma^2)$$

where we have used the fact that for two independent Gaussian variables $X \sim \mathrm{N}(a,b)$ and $Y \sim \mathrm{N}(c,d)$,

$$X + Y \sim \mathrm{N}(a+c, b+d).$$

Quick review

- The three-step procedure of *all* Bayesian inference.
- We have seen two examples—political poll (binomial) and measure pollutant (normal).
- In these analysis, we used a “trick” to figure out what the posterior distribution is without carrying out the integration.

When does this “trick” work?

Political poll revisited

- Instead of a Beta(α, β) prior for θ, what if we want to choose $p(\theta) \propto e^{-(\theta-0.5)^4}\mathbf{1}_{0<\theta<1}$?
- Bayes' Theorem still applies.

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot e^{-(\theta-0.5)^4} \\ &\propto \theta^x(1-\theta)^{n-x}e^{-(\theta-0.5)^4} \quad \text{for } 0<\theta<1. \end{aligned}$$

- This is not a standard distribution, and the normalizing constant must be evaluated through integration.

$$p(\theta|x) = \theta^x(1-\theta)^{n-x}e^{-(\theta-0.5)^4}/p(x)$$

  where $p(x) = \int_0^1 \theta^x(1-\theta)^{n-x}e^{-(\theta-0.5)^4}du$.
- This is *perfectly* valid. In fact if one has good reason to believe that $e^{-(\theta-0.5)^4}$ is the best reflection of prior knowledge then one can certainly choose this density as the prior.

- In the old days, this created a serious computational challenge.
- It is very difficult to calculate the normalizing constant, especially when θ is multi-dimensional.
- So it is desirable to use prior distributions that lead to simple posterior distributions, to give quick solutions.
- This is not as much a problem today due to the development of new methodology and faster computers. For simple problems, however, simple solutions may still be desirable.

So what is so special about our earlier examples that made them particularly neat?

Conjugate families

- Let Ψ be a family of probability distributions.
- Let $p(x|\theta)$ be the p.d.f of each data point conditional on the parameter θ.

If no matter which member in Ψ we choose as the prior distribution $p(\theta)$, the posterior distribution given an i.i.d. sample $X_1, X_2, \ldots, X_n$, $p(\theta|x_1, x_2, \ldots, x_n)$ is also a member of Ψ,

- then the family Ψ is said to be *conjugate* to the distribution $p(x|\theta)$.

Remark: Conjugacy by itself is not a useful concept.

- E.g. let Ψ be the family of all probability distributions.
- It is useful only when Ψ is a small enough family so that its members share interesting distributional properties.

Examples

- In the political poll example, the Beta(α, β) family is conjugate to the binomial distributions.
- In the air pollutant example, the Normal(μ, τ_0^2) family is conjugate to Normal(θ, σ^2) distributions with known σ^2. (Conjugate given σ.)

Two other commonly used conjugate families

- Poisson-Gamma conjugacy.
- Exponential-Gamma (or Gamma-Gamma) conjugacy.

Number of phone calls to a customer service line

Gamma-poisson distribution.

A company is deciding whether it should expand its customer service division and therefore wants to have an estimate of the average number of phone calls, denoted by θ, it receives between 9am-5pm.

- The data are the number of phone calls received between 9am-5pm on n different days—$X_1, X_2, \dots, X_n$.
- *What assummptions are already involved? Are they reasonable?*

How to carry out a Bayesian inference on θ.

Again, we need the two pieces to feed into Bayes' Theorem

1. How do we model the distribution of the number of calls received on each day?
   - Given θ, one can model the number of calls received on different days as i.i.d. Poisson(θ) random variables.

2. What prior distribution $p(\theta)$ can we choose to characterize our *a priori* knowledge about θ.
   - A useful family for this purpose is the Gamma(α, β) family.

The Gamma($\underset{\text{shape}}{\alpha}$, $\underset{\text{rate}}{\beta}$) distribution

sum of independent exponentials.

- Its pdf is

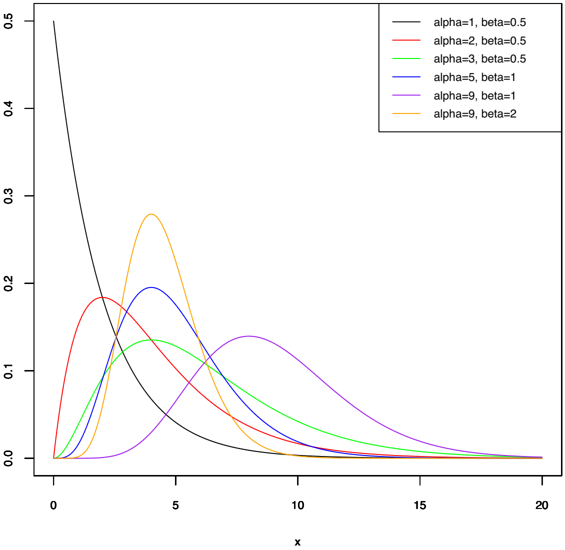
$$p(\theta|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \quad \text{for } \theta > 0$$
$$= 0 \quad \text{otherwise.}$$

- Its mean and variance are

$$\mathrm{E}[\theta] = \frac{\alpha}{\beta} \quad \text{and} \quad \mathrm{Var}[\theta] = \frac{\alpha}{\beta^2} = \frac{\mathrm{E}[\theta]}{\beta}$$

- So for the same mean larger β gives smaller variance.
- $\alpha > 0$ is called the *shape* parameter and $\beta > 0$ the *rate* parameter.
- We can use its relationship to the *exponential distribution* to remember these.

Gamma(α, β) p.d.fs



Applying Bayes' Theorem

$$
\begin{aligned}
p(\theta|x_1,x_2,\ldots,x_n) &\propto p(x_1,x_2,\ldots,x_n|\theta)p(\theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta)\cdot p(\theta) \\
&= \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!}\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\
&\propto \theta^{\alpha+\sum_{i=1}^{n}x_i-1}e^{-(\beta+n)\theta} \\
&= \theta^{\alpha_n-1}e^{-\beta_n\theta}
\end{aligned}
$$

- This is exactly the variable part of a *Gamma*(α_n,β_n) distribution where $\alpha_n=\alpha+\sum_{i=1}^{n}x_i$ and $\beta_n=\beta+n$.
- So $p(\theta|x_1,x_2,\ldots,x_n)$ must be this distribution.
- We can again find this inductively.
- Based on this posterior Bayesian inference can proceed—point summaries, credible intervals, predictive distribution, etc.

The posterior expectation of θ is

$$\begin{aligned} \mathrm{E}(\theta|x_1,x_2,\ldots,x_n) &= \frac{\alpha+\sum_{i=1}^n x_i}{\beta+n} \\ &= \left(\frac{\beta}{\beta+n}\right)\left(\frac{\alpha}{\beta}\right)+\left(\frac{n}{\beta+n}\right)\bar{x} \\ &= \left(\frac{\beta}{\beta+n}\right)\mu_\theta+\left(\frac{n}{\beta+n}\right)\bar{x} \\ &:= \mu_{\theta|x_1,x_2,\ldots,x_n}. \end{aligned}$$

The posterior variance is

$$\mathrm{Var}(\theta|x_1,x_2,\ldots,x_n) = \frac{\mu_{\theta|x_1,x_2,\ldots,x_n}}{\beta+n}.$$

- β measures the amount of information in the prior, while n measures the amount of information in the data.
- Similarly, we can construct 95% HPD CI. Find the predictive pmf of a future observation X_{n+1}. (Trick: Find the predictive distribution of X_1, and simply replace α and β with α_n and β_n.)
- Exercise: Show that the Gamma(α,β) family is also a conjugate prior on the *rate parameter* for Gamma sampling model.

The predictive density

$$p(x_{n+1} \mid \underline{x_n}) = \int p(x_{n+1} \mid \theta) \cdot \underbrace{p(\theta \mid \underline{x_n})}_{\text{math is the same.}} d\theta$$

- The prior predictive density is

$$
\begin{aligned}
\overset{|\underline{x_n}=\phi}{p(x)} &= \int p(x|\theta)\overset{\theta|x_n=\phi.}{p(\theta)}d\theta \\
&= \int \frac{\theta^x e^{-\theta}}{x!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)x!} \int \theta^{\alpha+x-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)x!} \frac{\Gamma(\alpha+x)}{(\beta+1)^{\alpha+x}}.
\end{aligned}
$$