MSA101/MVE187 2020 Lecture 2.1

Petter Mostad

Chalmers University

September 1, 2020

Notation

- In Bayesian statistics we talk a lot about probability density functions (for continuous variables) , or probability mass functions (for discrete variables), and *conditional* versions of these.
- Instead of naming a new such function each time, we use *generic notation* such as $\pi(x)$, $\pi(x, y)$, $\pi(x \mid y)$, $\pi(y \mid x)$, ...
- If x is continuous and y is discrete, we get, for example,

$$\int_x \pi(x)\,dx = 1,\ \sum_y \pi(y) = 1,\ \int_x \sum_y \pi(x, y) = 1,\ \int_x \pi(x, y)\,dx = \pi(y)$$

  where the integrals or sums are over all possible values of the variable.
- For conditional distributions we have basic relations such as

$$\pi(x \mid y)\pi(y) = \pi(x, y) \qquad \pi(x, y \mid z, w)\pi(z, w) = \pi(x, y, z, w).$$

- Thus for example $\pi(x, y \mid z, w)$ is interpreted as the density function for all combinations of x and y, when z and w are fixed to given values.

Example: Learning about a proportion

- An experiment is performed n times. We assume there is a probability p for "success" each time, and that the outcomes are independent. Let X be the observed number of successes. We get $X \mid p \sim \text{Binomial}(n, p)$. Given $X = x$, what do we know about p?
- For a Bayesian analysis, we need a joint probability density (or mass function) $\pi(X, p)$. We have defined $\pi(X \mid p)$ (the *likelihood*). We need to define $\pi(p)$ (the *prior*).
- Let us first try with the prior $p \sim \text{Uniform}[0, 1]$.
- The conditional model $\pi(p \mid X = x)$ (the *posterior* for p) can be computed with Bayes formula. We get

$$\pi(p \mid X = x) \propto_p p^x (1 - p)^{n-x}.$$

- We can recognize this as a Beta distribution: $p \mid X = x \sim \text{Beta}(x + 1, n - x + 1)$

Review of definition: The Beta distribution

θ has a Beta distribution on $[0, 1]$, with parameters α and β, if its density has the form

$$\pi(\theta \mid \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta *function* defined by

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where $\Gamma(t)$ is the *Gamma function* defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$$

Recall that for positive integers, $\Gamma(n) = (n-1)! = 0 \cdot 1 \cdot \cdots \cdot (n-1)$. See for example Wikipedia for more properties of the Beta distribution, and the Beta and Gamma functions. We write $\pi(\theta \mid \alpha, \beta) = \mathrm{Beta}(\theta; \alpha, \beta)$ for the Beta density; we then also write $\theta \sim \mathrm{Beta}(\alpha, \beta)$.

Using a Beta distribution as prior

- Assume the prior is $p \sim \text{Beta}(\alpha, \beta)$.
- The posterior becomes

$$p \mid (X = x) \sim \text{Beta}(\alpha + x, \beta + n - x)$$

- DEFINITION: Given a likelihood model $\pi(x \mid \theta)$. A *conjugate family of priors* to this likelihood is a parametric family of distributions so that if the prior for θ is in this family, the posterior $\theta \mid x$ is also in the family.

MSA101/MVE187 2020 Lecture 2.2

Petter Mostad

Chalmers University

September 9, 2020

Using a discrete prior

- What if the prior for p is a discrete distribution, i.e., $\pi(p) = \sum_{i=1}^{k} I(p = p_i) q_i$ where $p_1, \ldots, p_k$ are points in the interval $[0, 1]$ and $q_1, \ldots, q_k$ are their probabilities? Here, we use $I(\cdot)$ as the indicator function, so that $I(p = p_i)$ is a function of p that is equal to 1 when $p = p_i$ and zero otherwise.
- Continuing the example where $x \mid p \sim \text{Binomial}(n, p)$, the conditional model is obtained with Bayes theorem:

$$
\begin{aligned}
P(p = p_i \mid x) &= \frac{\pi(x \mid p = p_i)\pi(p_i)}{\pi(x)} = \frac{\pi(x \mid p = p_i)\pi(p_i)}{\sum_{i=1}^{k} \pi(x \mid p = p_i)\pi(p_i)} \\
&= \frac{\pi(x \mid p = p_i) q_i}{\sum_{i=1}^{k} \pi(x \mid p = p_i) q_i} = \frac{p_i^x (1 - p_i)^{n-x} q_i}{\sum_{j=1}^{k} p_j^x (1 - p_j)^{n-x} q_j}.
\end{aligned}
$$

  Note how the common factor $\binom{n}{x}$ disappears from the numerator and denominator.
- Computationally, you compute the vector of likelihoods, multiply termwise with the vector $(q_1, \ldots, q_k)$ of prior probabilities, and normalize to 1 (i.e., divide by the sum to get a vector summing to 1).

Using discretization

- Assume you have ANY prior, with density $\pi(p)$ on $[0, 1]$. This density can be approximated, generally with reasonable accuracy, with a discrete distribution, a *discretization*.
- The corresponding posterior produced by discretization can be easily produced by computer: Compute the likelihood on a grid over p, compute the prior on the same grid, multiply, and normalize.
- NOTE: This works for ANY likelihood, as long as the parameter p has a prior distribution on a bounded set.

Example: The Poisson-Gamma conjugacy

- Assume $\pi(x \mid \theta) = \text{Poisson}(x; \theta)$, i.e., that
$$\pi(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!}$$
- Then $\pi(\theta \mid \alpha, \beta) = \text{Gamma}(\theta; \alpha, \beta)$ where α, β are positive parameters, is a conjugate family. Recall that
$$\text{Gamma}(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$
- Specifically, we have the posterior
$$\pi(\theta \mid x) = \text{Gamma}\left(\theta; \alpha + x, \beta + 1\right).$$
- See Albert Section 3.3 for a computational example.

Example: The Normal-Gamma conjugacy

- Assume $\pi(x \mid \tau) = \text{Normal}(x; \mu, 1/\tau)$, so that x is normally distributed with known mean μ and unknown precision τ. The likelihood becomes

$$\pi(x \mid \tau) = \frac{1}{\sqrt{2\pi 1/\tau}} \exp\left(-\frac{1}{2/\tau}(x-\mu)^2\right) \propto_\tau \tau^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^2\tau\right)$$

- Then $\pi(\tau \mid \alpha, \beta) = \text{Gamma}(\tau; \alpha, \beta)$ is a conjugate family, in other words,

$$\pi(\tau \mid \alpha, \beta) \propto_\tau \tau^{\alpha-1} \exp(-\beta\tau).$$

- Specifically, we get the posterior below.

$$\pi(\tau \mid x) = \text{Gamma}\left(\tau; \alpha + \frac{1}{2}, \beta + \frac{1}{2}(x-\mu)^2\right).$$

- We can also describe this conjugacy using the variance σ^2 and an inverse Gamma (or inverse Chi-squared) distribution.

Example: the Normal-Normal conjugacy

- Assume $\pi(x \mid \theta) = \text{Normal}(x; \theta, 1/\tau_0)$, where τ_0 is a known and fixed *precision*.
- Then $\pi(\theta \mid \mu, \tau) = \text{Normal}(\theta; \mu, 1/\tau)$, where τ is positive and μ has any real value, is a conjugate family.
- Specifically, we have the posterior

$$\pi(\theta \mid x) = \text{Normal}\left(\theta; \frac{\tau_0 x + \tau\mu}{\tau_0 + \tau}, \frac{1}{\tau_0 + \tau}\right)$$

- PROOF: Use completion of squares.

PROOF

$$
\begin{aligned}
\pi(\theta \mid x) &\propto_\theta \pi(x \mid \theta)\pi(\theta) \\
&\propto_\theta \exp\left(-\frac{\tau_0}{2}(x-\theta)^2\right)\exp\left(-\frac{\tau}{2}(\theta-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\tau_0 x^2 - 2\tau_0 x\theta + \tau_0\theta^2 + \tau\theta^2 - 2\tau\theta\mu + \tau\mu^2\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}\left[(\tau_0+\tau)\theta^2 - 2(\tau_0 x + \tau\mu)\theta\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}(\tau_0+\tau)\left(\theta - \frac{\tau_0 x + \tau\mu}{\tau_0+\tau}\right)^2\right) \\
&\propto_\theta \text{Normal}\left(\theta; \frac{\tau_0 x + \tau\mu}{\tau_0+\tau}, \frac{1}{\tau_0+\tau}\right)
\end{aligned}
$$

MSA101/MVE187 2020 Lecture 2.3

Petter Mostad

Chalmers University

September 9, 2020

Prediction

The Bayesian paradigm implies:

- The usefulness of a model lies in its ability to predict.
- We create a joint probability model for the parameters θ, the observed data x, and data we would like to predict x_{new}. Often on the form $\pi(\theta, x, x_{new}) = \pi(\theta)\pi(x \mid \theta)\pi(x_{new} \mid \theta)$.
- The distribution for x_{new} is given by conditioning on the observed x and marginalizing out θ:

$$\begin{aligned}\pi(x_{new} \mid x) &= \int_\theta \pi(\theta, x_{new} \mid x)\, d\theta = \int_\theta \pi(x_{new} \mid \theta, x)\pi(\theta \mid x)\, d\theta \\ &= \int_\theta \pi(x_{new} \mid \theta)\pi(\theta \mid x)\, d\theta\end{aligned}$$

  This is called the *posterior predictive distribution*.

- It is also possible to look at the predictive distribution for x before it has been observed. This is called the *prior predictive distribution*:

$$\pi(x) = \int_\theta \pi(x, \theta)\, d\theta = \int_\theta \pi(x \mid \theta)\pi(\theta)\, d\theta$$

Predictive distributions when using conjugate priors

- When using a conjugate prior, not only do we have an analytic expression for the posterior density for θ, we also have analytic expressions for the prior predictive density and the posterior predictive density.
- To see this for the prior predictive density, use this formula derived from Bayes formula:

  $$\pi(x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)}$$

  The prior predictive density is on the left and all expressions on the right have analytic formulas.
- Note that, when using the right hand side for computing, θ will necessarily eventually disappear.
- As the posterior predictive distribution is on the same form as the prior predictive, we also get an analytic formula for it. Specifically, we can write

  $$\pi(x_{new} \mid x) = \frac{\pi(x_{new} \mid \theta)\pi(\theta \mid x)}{\pi(\theta \mid x_{new}, x)}.$$

Example: Predictive distribution for the Beta-Binomial conjugacy

- Assume $\pi(x \mid \theta) = \text{Binomial}(x; n, \theta)$ and $\pi(\theta) = \text{Beta}(\theta; \alpha, \beta)$.
- We get for the prior predictive

$$
\begin{aligned}
\pi(x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)} \\
&= \frac{\text{Binomial}(x; n, \theta)\,\text{Beta}(\theta; \alpha, \beta)}{\text{Beta}(\theta; \alpha + x, \beta + n - x)} \\
&= \frac{\begin{pmatrix} n \\ x \end{pmatrix} \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} / \,\mathrm{B}(\alpha, \beta)}{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} / \,\mathrm{B}(\alpha + x, \beta + n - x)} \\
&= \begin{pmatrix} n \\ x \end{pmatrix} \frac{\mathrm{B}(\alpha + x, \beta + n - x)}{\mathrm{B}(\alpha, \beta)}
\end{aligned}
$$

- This is the Beta-Binomial distribution with parameters n, α, and β.

Example: Predictive distribution for the Normal-Normal conjugacy

- Assume $\pi(x \mid \theta) = \text{Normal}(x; \theta, 1/\tau_0)$ and $\pi(\theta) = \text{Normal}(\mu, 1/\tau)$.
- Instead of using the type of computations above, the following is simpler:
  - We know from general theory of the normal distribution that $\pi(x)$ is normal.
  - Using the "law of total expectations", $E(x) = E(E(x \mid \theta)) = E(\theta) = \mu$.
  - Using the "law of total variance", $Var(x) = Var(E(x \mid \theta)) + E(Var(x \mid \theta)) = Var(\theta) + E(1/\tau_0) = 1/\tau + 1/\tau_0$.
- So for the prior predictive we get

$$\pi(x) = \text{Normal}(x; \mu; 1/\tau + 1/\tau_0)$$

MSA101/MVE187 2020 Lecture 2.4

Petter Mostad

Chalmers University

September 1, 2020

The exponential family of distributions

- Many parametric families of distributions can be written in a particular form:

$$\pi(x \mid \eta) = h(x)g(\eta)\exp\left(\eta \cdot u(x)\right)$$

  where η and $u(x)$ are vectors, $\eta \cdot u(x)$ is their dot product, and η is called the "natural parameters" of the family.
- Some examples of exponential families of distributions, corresponding to particular choices of g, h, and u:
  - Normal distributions.
  - Beta distributions.
  - Poisson distributions.
  - Gamma distributions.
  - Bernoulli distributions and Binomial distributions for a fixed N.
  - Multinomial distributions for a fixed N.
  - ....and many more.
- Exponential families of distributions share many properties and can be studied together.

Conjugacies and exponential families

- If $\pi(x \mid \eta) = h(x)g(\eta)\exp(\eta \cdot u(x))$, then a conjugate family of priors for η is given as

$$\pi(\eta \mid \nu, \beta) \propto_\eta g(\eta)^\nu \exp(\eta \cdot \beta) .$$

  The posterior becomes

$$\pi(\eta \mid x) \propto_\eta g(\eta)^{\nu+1} \exp(\eta \cdot (\beta + u(x))) .$$

- Essentially all examples of conjugacy fit into the framework above, so the above describes conjugacy in general.
- Note that the conjugate family of priors is also an exponential family.

Some properties

Assume $\pi(x \mid \eta) = h(x)g(\eta)\exp(\eta \cdot u(x))$.

- Given data $x_1, x_2, \ldots, x_N$ and a prior $\pi(\eta \mid \nu, \beta) \propto_\eta g(\eta)^\nu \exp(\eta \cdot \beta)$ the posterior becomes

$$\pi(\eta \mid x_1, \ldots, x_N) \propto_\eta g(\eta)^{\nu+N} \exp\left(\eta \cdot \left(\beta + \sum_{i=1}^{N} u(x_i)\right)\right).$$

- The expectation (and further moments) of $u(x)$ can be expressed with a differentiation of $g(\eta)$:

$$\mathrm{E}_{x|\eta}[u(x)] = -\nabla_\eta \log g(\eta).$$

- With for example a flat prior ($\nu = 0$, $\beta = 0$), the posterior is $\propto_\eta g(\eta)^N \exp\left(\eta \cdot \sum_{i=1}^{N} u(x_i)\right)$ and
  - The posterior (i.e., likelihood) depends only on $\sum_i u(x_i)$.
  - The maximum posterior (i.e., maximum likelihood) is the $\hat{\eta}$ satisfying

$$-\nabla_\eta \log g(\hat{\eta}) = \frac{1}{N} \sum_{i=1}^{N} u(x_i).$$