MT4531/MT5731: (Advanced) Bayesian Inference

Prediction

Nicolo Margaritella

School of Mathematics and Statistics, University of St Andrews

–

Outline

1. Prior predictive distribution
2. Posterior predictive distribution

Outline

1. Prior predictive distribution

2. Posterior predictive distribution

Prior predictive distribution

- Suppose that we wish to describe our beliefs about a random vector of future data $\boldsymbol{X}$, with likelihood $f(\boldsymbol{x}|\theta)$, but unknown parameter $\theta \in \Theta$.
- If uncertainty for θ is represented by $p(\theta)$, then the pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta\in\Theta} f(\boldsymbol{x},\theta)d\theta = \int_{\theta\in\Theta} f(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

- The term $f(\boldsymbol{x})$ is called the *prior predictive distribution*, in addition to the 'marginal likelihood' term we encountered earlier.
- Essentially, we are weighting the likelihood $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ.
- Since we have not observed any data, that is the prior distribution for θ.

Prior predictive distribution

- Suppose that we wish to describe our beliefs about a random vector of future data $\boldsymbol{X}$, with likelihood $f(\boldsymbol{x}|\theta)$, but unknown parameter $\theta \in \Theta$.
- If uncertainty for θ is represented by $p(\theta)$, then the pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta \in \Theta} f(\boldsymbol{x}, \theta) d\theta = \int_{\theta \in \Theta} f(\boldsymbol{x}|\theta) p(\theta) d\theta.$$

- The term $f(\boldsymbol{x})$ is called the *prior predictive distribution*, in addition to the 'marginal likelihood' term we encountered earlier.
- Essentially, we are weighting the likelihood $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ.
- Since we have not observed any data, that is the prior distribution for θ.

Prior predictive distribution

- Suppose that we wish to describe our beliefs about a random vector of future data $\boldsymbol{X}$, with likelihood $f(\boldsymbol{x}|\theta)$, but unknown parameter $\theta \in \Theta$.
- If uncertainty for θ is represented by $p(\theta)$, then the pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta\in\Theta} f(\boldsymbol{x}, \theta) d\theta = \int_{\theta\in\Theta} f(\boldsymbol{x}|\theta) p(\theta) d\theta.$$

- The term $f(\boldsymbol{x})$ is called the *prior predictive distribution*, in addition to the 'marginal likelihood' term we encountered earlier.
- Essentially, we are weighting the likelihood $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ.
- Since we have not observed any data, that is the prior distribution for θ.

Prior predictive distribution

- Suppose that we wish to describe our beliefs about a random vector of future data $\boldsymbol{X}$, with likelihood $f(\boldsymbol{x}|\theta)$, but unknown parameter $\theta \in \Theta$.
- If uncertainty for θ is represented by $p(\theta)$, then the pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta \in \Theta} f(\boldsymbol{x}, \theta) d\theta = \int_{\theta \in \Theta} f(\boldsymbol{x}|\theta) p(\theta) d\theta.$$

- The term $f(\boldsymbol{x})$ is called the *prior predictive distribution*, in addition to the 'marginal likelihood' term we encountered earlier.
- Essentially, we are weighting the likelihood $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ.
- Since we have not observed any data, that is the prior distribution for θ.

Prior predictive distribution

- Suppose that we wish to describe our beliefs about a random vector of future data $\boldsymbol{X}$, with likelihood $f(\boldsymbol{x}|\theta)$, but unknown parameter $\theta \in \Theta$.
- If uncertainty for θ is represented by $p(\theta)$, then the pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta\in\Theta} f(\boldsymbol{x},\theta)d\theta = \int_{\theta\in\Theta} f(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

- The term $f(\boldsymbol{x})$ is called the *prior predictive distribution*, in addition to the 'marginal likelihood' term we encountered earlier.
- Essentially, we are weighting the likelihood $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ.
- Since we have not observed any data, that is the prior distribution for θ.

Example

- Consider X such that, $X \sim Exp(\lambda)$, $\lambda > 0$.
- Assume also prior beliefs on λ described by a $\Gamma(\alpha, \beta)$ distribution $(\alpha, \beta > 0)$.
- The prior predictive distribution $f(\boldsymbol{x})$ was calculated back in lecture 4, as
$$f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^{\alpha}}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.$$
- The calculation is shown again in the next slide as a reminder.

Example

- Consider X such that, $X \sim Exp(\lambda)$, $\lambda > 0$.
- Assume also prior beliefs on λ described by a $\Gamma(\alpha, \beta)$ distribution $(\alpha, \beta > 0)$.
- The prior predictive distribution $f(\boldsymbol{x})$ was calculated back in lecture 4, as
$$f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^{\alpha}}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.$$
- The calculation is shown again in the next slide as a reminder.

Example

- Consider X such that, $X \sim Exp(\lambda)$, $\lambda > 0$.
- Assume also prior beliefs on λ described by a $\Gamma(\alpha, \beta)$ distribution $(\alpha, \beta > 0)$.
- The prior predictive distribution $f(\boldsymbol{x})$ was calculated back in lecture 4, as

$$f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^{\alpha}}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.$$

- The calculation is shown again in the next slide as a reminder.

Example

- Consider X such that, $X \sim Exp(\lambda)$, $\lambda > 0$.
- Assume also prior beliefs on λ described by a $\Gamma(\alpha, \beta)$ distribution $(\alpha, \beta > 0)$.
- The prior predictive distribution $f(\boldsymbol{x})$ was calculated back in lecture 4, as

$$f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^{\alpha}}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.$$

- The calculation is shown again in the next slide as a reminder.

Example

$$
\begin{aligned}
\pi(\lambda|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\lambda)p(\lambda) = f(x_1|\lambda) \times \ldots f(x_n|\lambda)p(\lambda) \\
&= \prod_{i=1}^{n} \lambda \exp(-x_i\lambda) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\lambda\beta) \\
&\propto \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \times \lambda^{\alpha-1}\exp(-\lambda\beta) \\
&= \lambda^{n+\alpha-1}\exp(-\lambda[n\bar{x}+\beta]) \\
&\propto \frac{(n\bar{x}+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}\lambda^{n+\alpha-1}\exp(-\lambda[n\bar{x}+\beta]) \\
\Rightarrow \lambda|\boldsymbol{x} &\sim \Gamma(n+\alpha, n\bar{x}+\beta).
\end{aligned}
$$

Given this, we can state that the constant of proportionality (the constant we multiply with to obtain a density that integrates to one) is equal to, $\frac{(n\bar{x}+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}$. Or, by inspection, we can write that,

$$
f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^\alpha}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.
$$

Note on the prior predictive distribution

- Note that the prior predictive distribution is in fact the denominator in the expression for Bayes' Theorem, when the data $\boldsymbol{x}$ have **not** been substituted by numerical values.
- So, all examples in the lecture notes or tutorial sheets where the expression for $f(\boldsymbol{x})$ is calculated are also examples of a prior predictive distribution.

Note on the prior predictive distribution

- Note that the prior predictive distribution is in fact the denominator in the expression for Bayes' Theorem, when the data $\boldsymbol{x}$ have **not** been substituted by numerical values.
- So, all examples in the lecture notes or tutorial sheets where the expression for $f(\boldsymbol{x})$ is calculated are also examples of a prior predictive distribution.

Outline

1. Prior predictive distribution

2. Posterior predictive distribution

Posterior predictive distribution

- We observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process.
- Assume that conditional on the parameter θ in the process, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.
- Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \int_\Theta f(\boldsymbol{y},\theta|\boldsymbol{x})d\theta \\ &= \int_\Theta f(\boldsymbol{y}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})d\theta \\ &= \int_\Theta f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x})d\theta, \end{aligned}$$

- Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Posterior predictive distribution

- We observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process.
- Assume that conditional on the parameter θ in the process, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.
- Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{x}) &= \int_{\Theta} f(\boldsymbol{y}, \theta|\boldsymbol{x}) d\theta \\
&= \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{x}, \theta)\pi(\theta|\boldsymbol{x}) d\theta \\
&= \int_{\Theta} f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x}) d\theta,
\end{aligned}
$$

- Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Posterior predictive distribution

- We observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process.
- Assume that conditional on the parameter θ in the process, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.
- Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \int_{\Theta} f(\boldsymbol{y},\theta|\boldsymbol{x})d\theta \\ &= \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})d\theta \\ &= \int_{\Theta} f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x})d\theta, \end{aligned}$$

- Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Posterior predictive distribution

- We observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process.
- Assume that conditional on the parameter θ in the process, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.
- Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \int_{\Theta} f(\boldsymbol{y},\theta|\boldsymbol{x})d\theta \\ &= \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})d\theta \\ &= \int_{\Theta} f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x})d\theta, \end{aligned}$$

- Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Example (set up)

- Suppose that the number of calls X to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown.
- (So, for a period of 10 minutes, $X \sim Poisson(\lambda)$.)
- Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

- So, the prior expectation for λ is $E(\lambda) = 1/10$.

Example (set up)

- Suppose that the number of calls X to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown.
- (So, for a period of 10 minutes, $X \sim Poisson(\lambda)$.)
- Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

- So, the prior expectation for λ is $E(\lambda) = 1/10$.

Example (set up)

- Suppose that the number of calls X to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown.
- (So, for a period of 10 minutes, $X \sim Poisson(\lambda)$.)
- Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

- So, the prior expectation for λ is $E(\lambda) = 1/10$.

Example (set up)

- Suppose that the number of calls X to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown.
- (So, for a period of 10 minutes, $X \sim Poisson(\lambda)$.)
- Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

- So, the prior expectation for λ is $E(\lambda) = 1/10$.

Example (prior predictive distribution)

- The prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X, is

Example (prior predictive distribution)

- The prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X, is
- $$\begin{aligned} P(X=x)=f(x) &= \int_0^\infty f(x|\lambda)p(\lambda)d\lambda \\ &= \int_0^\infty \frac{\lambda^x}{x!}\exp(-\lambda)\times 10\exp(-10\lambda)d\lambda \\ &= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}}\int_0^\infty \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda \\ &= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}} \\ &= \frac{10}{11^{x+1}}, \end{aligned}$$

  since $\Gamma(x+1)=x!$, as x is a positive integer.
- Does this make sense? What if $\lambda \sim Exp(1)$?

Example (prior predictive distribution)

- The prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X, is
-

$$
\begin{aligned}
P(X=x)=f(x) &= \int_0^\infty f(x|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^x}{x!}\exp(-\lambda)\times 10\exp(-10\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}}\int_0^\infty \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}} \\
&= \frac{10}{11^{x+1}},
\end{aligned}
$$

since $\Gamma(x+1)=x!$, as x is a positive integer.

- Does this make sense? What if $\lambda \sim Exp(1)$?

Example (prior predictive distribution)

- The prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X, is
- $$\begin{aligned} P(X=x)=f(x) &= \int_0^\infty f(x|\lambda)p(\lambda)d\lambda \\ &= \int_0^\infty \frac{\lambda^x}{x!}\exp(-\lambda)\times 10\exp(-10\lambda)d\lambda \\ &= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}}\int_0^\infty \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda \\ &= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}} \\ &= \frac{10}{11^{x+1}}, \end{aligned}$$

  since $\Gamma(x+1)=x!$, as x is a positive integer.
- Does this make sense? What if $\lambda \sim Exp(1)$?

Example (prior predictive distribution)

- **Task**: Complete the example in the lecture notes.