STAT 339
Bayesian Inference IV

December 3rd, 2021

Colin Reimer Dawson

Outline

Outline

The Predictive Distribution

Model Selection and Bayesian Occam's Razor

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data
- If our goal is understanding that process, this is an end in itself

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data
- If our goal is understanding that process, this is an end in itself
- However, many of our ML models are designed to to make predictions about some $\mathbf{y}_{\text{new}}$.

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data
- If our goal is understanding that process, this is an end in itself
- However, many of our ML models are designed to to make predictions about some $\mathbf{y}_{\text{new}}$.
- When using optimization methods such MLE, we get a single value $\hat{\theta}$ and can then predict using $p(\mathbf{y}_{\text{new}} \mid \hat{\theta})$

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data
- If our goal is understanding that process, this is an end in itself
- However, many of our ML models are designed to to make predictions about some $\mathbf{y}_{\text{new}}$.
- When using optimization methods such MLE, we get a single value $\hat{\theta}$ and can then predict using $p(\mathbf{y}_{\text{new}} \mid \hat{\theta})$
- With Bayesian inference, however, we get a **distribution**, $p(\theta \mid \mathbf{y}_{\text{train}})$, not a single value.

Inference vs Prediction

- The posterior distribution, $p(\theta \mid \mathbf{y}_{\text{train}})$, expresses information about the process that generated the data
- If our goal is understanding that process, this is an end in itself
- However, many of our ML models are designed to to make predictions about some $\mathbf{y}_{\text{new}}$.
- When using optimization methods such MLE, we get a single value $\hat{\theta}$ and can then predict using $p(\mathbf{y}_{\text{new}} \mid \hat{\theta})$
- With Bayesian inference, however, we get a **distribution**, $p(\theta \mid \mathbf{y}_{\text{train}})$, not a single value.
- How do we use this to make predictions?

Option 1: Distribution to Point Estimate

- One option: Find $\hat{\theta}$, a **point estimate** of θ from the posterior (e.g., the mean, or mode) and use $p(\mathbf{y}_{\text{new}} \mid \hat{\theta})$ for prediction

Option 1: Distribution to Point Estimate

- One option: Find $\hat{\theta}$, a **point estimate** of θ from the posterior (e.g., the mean, or mode) and use $p(\mathbf{y}_{\text{new}} \mid \hat{\theta})$ for prediction
- However, **this discards our uncertainty**, and one of the main points of a Bayesian approach is **principled handling of uncertainty**

Option 2: Posterior Predictive Distribution

A more "fully Bayesian" solution: Compute the **posterior predictive distribution**:

$$
\begin{aligned}
p(\mathbf{y}_{\text{new}} \mid \mathbf{y}_{\text{train}}) &= \int p(\mathbf{y}_{\text{new}}, \theta \mid \mathbf{y}_{\text{train}}) \, d\theta \\
&= \int p(\mathbf{y}_{\text{new}} \mid \theta, \mathbf{y}_{\text{train}}) p(\theta \mid \mathbf{y}_{\text{train}}) \, d\theta
\end{aligned}
$$

Option 2: Posterior Predictive Distribution

A more "fully Bayesian" solution: Compute the **posterior predictive distribution**:

$$
\begin{aligned}
p(\mathbf{y}_{\text{new}} \mid \mathbf{y}_{\text{train}}) &= \int p(\mathbf{y}_{\text{new}}, \theta \mid \mathbf{y}_{\text{train}})\; d\theta \\
&= \int p(\mathbf{y}_{\text{new}} \mid \theta, \mathbf{y}_{\text{train}}) p(\theta \mid \mathbf{y}_{\text{train}})\; d\theta
\end{aligned}
$$

If $\mathbf{y}_{\text{new}}$ and $\mathbf{y}_{\text{train}}$ are **conditionally independent** given θ, this simplifies to

$$
p(\mathbf{y}_{\text{new}} \mid \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{new}} \mid \theta) p(\theta \mid \mathbf{y}_{\text{train}})\; d\theta
$$

which is expressed in terms of the data-generating model (likelihood) and the posterior. In fact, it is equivalent to

$$
\mathbb{E}\left[p(\mathbf{y}_{\text{new}} \mid \theta) \mid \mathbf{y}_{\text{train}}\right]
$$

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \dots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

and we observe 12 "successes" out of $N = 40$ observations

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \ldots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

  and we observe 12 "successes" out of $N = 40$ observations
- This yields posterior

$$\theta \mid \mathbf{y}_{\text{train}} \sim \text{Beta}(12 + 1, 28 + 1)$$

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \dots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

  and we observe 12 "successes" out of $N = 40$ observations
- This yields posterior

$$\theta \mid \mathbf{y}_{\text{train}} \sim \text{Beta}(12 + 1, 28 + 1)$$

- The predictive probability that the next observation is a success if we **know** μ, that is, $p(y_{\text{new}} \mid \mu)$, is just μ.

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \ldots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

  and we observe 12 "successes" out of $N = 40$ observations
- This yields posterior

$$\theta \mid \mathbf{y}_{\text{train}} \sim \text{Beta}(12 + 1, 28 + 1)$$

- The predictive probability that the next observation is a success if we **know** μ, that is, $p(y_{\text{new}} \mid \mu)$, is just μ.
- Using point estimation for μ we might get:

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \dots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

  and we observe 12 "successes" out of $N = 40$ observations
- This yields posterior

$$\theta \mid \mathbf{y}_{\text{train}} \sim \text{Beta}(12 + 1, 28 + 1)$$

- The predictive probability that the next observation is a success if we **know** μ, that is, $p(y_{\text{new}} \mid \mu)$, is just μ.
- Using point estimation for μ we might get:
  - MLE: $\hat{\mu} = p(y_{\text{new}} = 1 \mid \hat{\mu}) = \frac{12}{40} = 0.30$.

Example: Beta-Bernoulli Model

- Suppose we have data-generating model and prior

$$y_1, \dots y_N \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu)$$
$$\mu \sim \text{Unif}(0, 1)$$

  and we observe 12 "successes" out of $N = 40$ observations
- This yields posterior

$$\theta \mid \mathbf{y}_{\text{train}} \sim \text{Beta}(12 + 1, 28 + 1)$$

- The predictive probability that the next observation is a success if we **know** μ, that is, $p(y_{\text{new}} \mid \mu)$, is just μ.
- Using point estimation for μ we might get:
  - MLE: $\hat{\mu} = p(y_{\text{new}} = 1 \mid \hat{\mu}) = \frac{12}{40} = 0.30.$
  - Posterior mean: $\mathbb{E}\left[\mu \mid \mathbf{y}_{\text{new}}\right] = \frac{12+1}{40+2} = 0.31$

Beta-Bernoulli: Predictive Distribution

Alternatively, calculating the predictive probability directly:

$$p(y_{new} = 1 \mid \mathbf{y}_{\text{train}}) = \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu$$

Beta-Bernoulli: Predictive Distribution

Alternatively, calculating the predictive probability directly:

$$\begin{aligned} p(y_{new} = 1 \mid \mathbf{y}_{\text{train}}) &= \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \\ &= \int_0^1 \mu \, p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \end{aligned}$$

Beta-Bernoulli: Predictive Distribution

Alternatively, calculating the predictive probability directly:

$$\begin{aligned} p(y_{new} = 1 \mid \mathbf{y}_{\text{train}}) &= \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \\ &= \int_0^1 \mu \, p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \\ &= \mathbb{E} \left[\mu \mid \mathbf{y}_{\text{train}} \right] \end{aligned}$$

Beta-Bernoulli: Predictive Distribution

Alternatively, calculating the predictive probability directly:

$$
\begin{aligned}
p(y_{new} = 1 \mid \mathbf{y}_{\text{train}}) &= \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}_{\text{train}}) \; d\mu \\
&= \int_0^1 \mu \; p(\mu \mid \mathbf{y}_{\text{train}}) \; d\mu \\
&= \mathbb{E}\left[\mu \mid \mathbf{y}_{\text{train}}\right] \\
&= \frac{12 + 1}{40 + 2} = 0.31
\end{aligned}
$$

Beta-Bernoulli: Predictive Distribution

Alternatively, calculating the predictive probability directly:

$$
\begin{aligned}
p(y_{new} = 1 \mid \mathbf{y}_{\text{train}}) &= \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \\
&= \int_0^1 \mu \, p(\mu \mid \mathbf{y}_{\text{train}}) \, d\mu \\
&= \mathbb{E}\left[\mu \mid \mathbf{y}_{\text{train}}\right] \\
&= \frac{12+1}{40+2} = 0.31
\end{aligned}
$$

In this case, the predictive probability of interest is just the **posterior mean** (this will not always be true, however).

Example: Gamma-Poisson

- Suppose our taqueria has had a total of y_{train} customers in the last N Saturdays.

Example: Gamma-Poisson

- Suppose our taqueria has had a total of y_{train} customers in the last N Saturdays.
- How many customers should we expect next Saturday?

Example: Gamma-Poisson

- Suppose our taqueria has had a total of y_{train} customers in the last N Saturdays.
- How many customers should we expect next Saturday?
- A simple likelihood and prior (ignoring seasonal variation, etc.):

$$
\begin{aligned}
y_{\text{train}} \mid \lambda &\sim \text{Poisson}(N\lambda) \\
\lambda &\sim \text{Gamma}(a_0, b_0)
\end{aligned}
$$

Example: Gamma-Poisson

- Suppose our taqueria has had a total of y_{train} customers in the last N Saturdays.
- How many customers should we expect next Saturday?
- A simple likelihood and prior (ignoring seasonal variation, etc.):

$$
\begin{aligned}
y_{\text{train}} \mid \lambda &\sim \text{Poisson}(N\lambda) \\
\lambda &\sim \text{Gamma}(a_0, b_0)
\end{aligned}
$$

- If we knew λ we would expect λ customers per day on average

Gamma-Poisson: Predicting with a Point Estimate

If we predict using $p(y_{\text{new}} \mid \hat{\lambda})$ with

$$\hat{\lambda}_{\text{MLE}} = \frac{y_{\text{train}}}{N}$$

then

$$\mathbb{E}\left[y_{new} \mid \hat{\lambda}\right] = \hat{\lambda} = \frac{y_{\text{train}}}{N}$$

If instead we use

$$\hat{\lambda} = \mathbb{E}\left[\lambda \mid y_{\text{train}}\right] = \frac{a_{\text{post}}}{b_{\text{post}}}$$

where

$$a_{\text{post}} = a_0 + y_{\text{train}} \qquad b_{\text{post}} = b_0 + N$$

then

$$\mathbb{E}\left[y_{new} \mid \hat{\lambda}\right] = \hat{\lambda} = \frac{a_{\text{post}}}{b_{\text{post}}} = \frac{a_0 + y_{\text{train}}}{b_0 + N}$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$p(y_{\text{new}} \mid y_{\text{train}}) = \int_0^\infty p(y_{\text{new}} \mid \lambda) p(\lambda \mid y_{\text{train}}) \, d\lambda$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$
\begin{aligned}
p(y_{\text{new}} \mid y_{\text{train}}) &= \int_0^\infty p(y_{\text{new}} \mid \lambda) p(\lambda \mid y_{\text{train}}) \, d\lambda \\
&= \frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}})} \int_0^\infty \frac{e^{-\lambda} \lambda^{y_{\text{new}}}}{y_{\text{new}}!} \lambda^{a_{\text{post}}-1} e^{-b_{\text{post}} \lambda} \, d\lambda
\end{aligned}
$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$
\begin{aligned}
p(y_{\text{new}} \mid y_{\text{train}}) &= \frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}})} \int_0^\infty \frac{e^{-\lambda} \lambda^{y_{\text{new}}}}{y_{\text{new}}!} \lambda^{a_{\text{post}}-1} e^{-b_{\text{post}}\lambda} \, d\lambda \\
&= \frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \int_0^\infty \lambda^{a_{\text{post}}+y_{\text{new}}-1} e^{-(b_{\text{post}}+1)\lambda} \, d\lambda
\end{aligned}
$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$
\begin{aligned}
p(y_{\text{new}} \mid y_{\text{train}}) &= \frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \int_0^\infty \lambda^{a_{\text{post}}+y_{\text{new}}-1} e^{-(b_{\text{post}}+1)\lambda} \, d\lambda \\
&= \left(\frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \right) \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{(b_{\text{post}} + 1)^{a_{\text{post}}+y_{\text{new}}}} \right)
\end{aligned}
$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$
\begin{aligned}
p(y_{\text{new}} \mid y_{\text{train}}) &= \left(\frac{b_{\text{post}}^{a_{\text{post}}}}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \right) \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{(b_{\text{post}} + 1)^{a_{\text{post}} + y_{\text{new}}}} \right) \\
&= \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \right) \left(\frac{1}{b_{\text{post}} + 1} \right)^{y_{\text{new}}} \left(\frac{b_{\text{post}}}{b_{\text{post}} + 1} \right)^{a_{\text{post}}}
\end{aligned}
$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$
\begin{aligned}
p(y_{\text{new}} \mid y_{\text{train}}) &= \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}})y_{\text{new}}!}\right)\left(\frac{1}{b_{\text{post}} + 1}\right)^{y_{\text{new}}}\left(\frac{b_{\text{post}}}{b_{\text{post}} + 1}\right)^{a_{\text{post}}} \\
&= \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}})y_{\text{new}}!}\right)\mu^{y_{\text{new}}}(1-\mu)^{a_{\text{post}}}
\end{aligned}
$$

where $\mu := \frac{1}{b_{\text{post}}+1}$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$p(y_{\text{new}} \mid y_{\text{train}}) = \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \right) \mu^{y_{\text{new}}} (1 - \mu)^{a_{\text{post}}}$$

where $\mu := \frac{1}{b_{\text{post}}+1}$ Note that if a_0 (representing the number of "virtual" customers visiting in b_0 days as encoded in the prior) is an integer, then so is $a_{\text{post}} = a_0 + y_{\text{train}}$, and we can write

$$p(y_{\text{new}} \mid y_{\text{train}}) = \frac{(a_{\text{post}} + y_{\text{new}} - 1)!}{(a_{\text{post}} - 1)! y_{\text{new}}!} \mu^{y_{\text{new}}} (1 - \mu)^{a_{\text{post}}}$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$p(y_{\text{new}} \mid y_{\text{train}}) = \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}}) y_{\text{new}}!} \right) \mu^{y_{\text{new}}} (1 - \mu)^{a_{\text{post}}}$$

where $\mu := \frac{1}{b_{\text{post}}+1}$ Note that if a_0 (representing the number of "virtual" customers visiting in b_0 days as encoded in the prior) is an integer, then so is $a_{\text{post}} = a_0 + y_{\text{train}}$, and we can write

$$\begin{aligned} p(y_{\text{new}} \mid y_{\text{train}}) &= \frac{(a_{\text{post}} + y_{\text{new}} - 1)!}{(a_{\text{post}} - 1)! y_{\text{new}}!} \mu^{y_{\text{new}}} (1 - \mu)^{a_{\text{post}}} \\ &= \binom{a_{\text{post}} + y_{\text{post}} - 1}{y_{\text{new}}} \mu^{y_{\text{new}}} (1 - \mu)^{a_{\text{post}}} \end{aligned}$$

This is a **Negative Binomial** distribution with parameters a_{post} and μ.

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$p(y_{\text{new}} \mid y_{\text{train}}) = \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}})y_{\text{new}}!}\right)\mu^{y_{\text{new}}}(1-\mu)^{a_{\text{post}}}$$

where $\mu := \frac{1}{b_{\text{post}}+1}$ Note that if a_0 (representing the number of "virtual" customers visiting in b_0 days as encoded in the prior) is an integer, then so is $a_{\text{post}} = a_0 + y_{\text{train}}$, and we can write

$$p(y_{\text{new}} \mid y_{\text{train}}) = \binom{a_{\text{post}} + y_{\text{post}} - 1}{y_{\text{new}}}\mu^{y_{\text{new}}}(1-\mu)^{a_{\text{post}}}$$

This is a **Negative Binomial** distribution with parameters a_{post} and μ. Its mean and variance are

$$\mathbb{E}\left[y_{\text{new}} \mid y_{\text{train}}\right] = \left(\frac{\mu}{1-\mu}\right)a_{\text{post}} = \frac{a_{\text{post}}}{b_{\text{post}}}$$

$$\mathbb{V}\text{ar}\left[y_{\text{new}} \mid y_{\text{train}}\right] = \left(\frac{\mu}{1-\mu}\right)\frac{a_{\text{post}}}{1-\mu} = \frac{a_{\text{post}}}{b_{\text{post}}}\frac{1+b_{\text{post}}}{b_{\text{post}}}$$

Gamma-Poisson: Predictive Distribution

Alternatively, calculating the predictive distribution directly:

$$p(y_{\text{new}} \mid y_{\text{train}}) = \left(\frac{\Gamma(a_{\text{post}} + y_{\text{new}})}{\Gamma(a_{\text{post}})y_{\text{new}}!}\right)\mu^{y_{\text{new}}}(1-\mu)^{a_{\text{post}}}$$

where $\mu := \frac{1}{b_{\text{post}}+1}$ Note that if a_0 (representing the number of "virtual" customers visiting in b_0 days as encoded in the prior) is an integer, then so is $a_{\text{post}} = a_0 + y_{\text{train}}$, and we can write

$$p(y_{\text{new}} \mid y_{\text{train}}) = \binom{a_{\text{post}} + y_{\text{post}} - 1}{y_{\text{new}}}\mu^{y_{\text{new}}}(1-\mu)^{a_{\text{post}}}$$

This is a **Negative Binomial** distribution with parameters a_{post} and μ. Its mean and variance are

$$\mathbb{E}\left[y_{\text{new}} \mid y_{\text{train}}\right] = \left(\frac{\mu}{1-\mu}\right)a_{\text{post}} = \frac{a_{\text{post}}}{b_{\text{post}}}$$

$$\mathbb{V}\text{ar}\left[y_{\text{new}} \mid y_{\text{train}}\right] = \left(\frac{\mu}{1-\mu}\right)\frac{a_{\text{post}}}{1-\mu} = \frac{a_{\text{post}}}{b_{\text{post}}}(1 + b_{\text{post}}^{-1})$$

Gamma-Poisson: Predictive Distribution

We have shown that, with data-generating model and prior

$$
\begin{aligned}
y_{\text{train}} \mid \lambda &\sim \mathsf{Poisson}(N\lambda) \\
\lambda &\sim \mathsf{Gamma}(a_0, b_0)
\end{aligned}
$$

we get posterior, and posterior predictive distributions:

$$
\begin{aligned}
\lambda \mid y_{\text{train}} &\sim \mathsf{Gamma}(a_0 + y_{\text{train}}, b_0 + N) \\
y_{\text{new}} \mid y_{\text{train}} &\sim \mathsf{NegBinom}(a_0 + y_{\text{train}}, (b_0 + N + 1)^{-1})
\end{aligned}
$$

with predictive mean and variance:

$$
\begin{aligned}
\mathbb{E}\left[y_{\text{new}} \mid y_{\text{train}}\right] &= \frac{a_0 + y_{\text{train}}}{b_0 + N} = \left(\frac{b_0}{b_0 + N}\right)\mathbb{E}\left[\lambda\right] + \left(\frac{N}{b_0 + N}\right)\hat{\lambda}_{MLE} \\
\mathbb{V}\text{ar}\left[y_{\text{new}} \mid y_{\text{train}}\right] &= \left(\frac{a_0 + y_{\text{train}}}{b_0 + N}\right)\left(1 + (b_0 + N)^{-1}\right)
\end{aligned}
$$

Note that as $N \to \infty$, both the predictive mean and variance converge to their values when $\lambda = \hat{\lambda}_{MLE}$.

Outline

The Predictive Distribution

Model Selection and Bayesian Occam's Razor

Model Selection

- In many cases we have more than one **family** of data-generating distributions under consideration. E.g.

$$\text{Set of candidate model families } = \{\mathcal{M}_k\}_{k=1}^{K}$$

Model Selection

- In many cases we have more than one **family** of data-generating distributions under consideration. E.g.

$$\text{Set of candidate model families } = \{\mathcal{M}_k\}_{k=1}^{K}$$

- Each of these may depend on some parameter vector θ (it may not have the same size for all of them)

Model Selection

- In many cases we have more than one **family** of data-generating distributions under consideration. E.g.

$$\text{Set of candidate model families } = \{\mathcal{M}_k\}_{k=1}^{K}$$

- Each of these may depend on some parameter vector θ (it may not have the same size for all of them)
- We can construct a **hierarchical prior** to *simultaneously* infer $\mathcal{M}$ and θ:

$$p(\mathcal{M}, \theta) = p(\mathcal{M})p(\theta \mid \mathcal{M})$$

Model Selection

- In many cases we have more than one **family** of data-generating distributions under consideration. E.g.

$$\text{Set of candidate model families } = \{\mathcal{M}_k\}_{k=1}^K$$

- Each of these may depend on some parameter vector θ (it may not have the same size for all of them)
- We can construct a **hierarchical prior** to *simultaneously* infer $\mathcal{M}$ and θ:

$$p(\mathcal{M}, \theta) = p(\mathcal{M})p(\theta \mid \mathcal{M})$$

- To examine the **posterior plausibility of each model class** (**averaging** over possible θ), we are interested in

$$p(\mathcal{M} \mid \mathbf{y}) = k_{\mathbf{y}} p(\mathbf{y} \mid \mathcal{M})p(\mathcal{M})$$

Marginal Likelihood

To find $p(\mathcal{M} \mid \mathbf{y})$, we need $p(\mathcal{M})$ (which we specify as part of the prior), and $p(\mathbf{y} \mid \mathcal{M})$.

Marginal Likelihood

To find $p(\mathcal{M} \mid \mathbf{y})$, we need $p(\mathcal{M})$ (which we specify as part of the prior), and $p(\mathbf{y} \mid \mathcal{M})$.

The latter is called the **marginal likelihood**:

Marginal Likelihood

The **marginal likelihood** for a dataset $\mathbf{y}$ given a model class, $\mathcal{M}$ is

$$p(\mathbf{y}_{\text{train}} \mid \mathcal{M}) = \int p(\mathbf{y}_{\text{train}} \mid \theta, \mathcal{M}) p(\theta \mid \mathcal{M}) \, d\theta$$

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $\mathcal{M}_{\text{fair}}$, we set

$$p(\mu \mid \mathcal{M}_{\text{fair}}) = I(\mu = 0.5)$$

(a "degenerate" PMF on μ)

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $\mathcal{M}_{\text{fair}}$, we set

$$p(\mu \mid \mathcal{M}_{\text{fair}}) = I(\mu = 0.5)$$

  (a "degenerate" PMF on μ)
- For a biased, coin, we put a uniform prior on μ:

$$p(\mu \mid \mathcal{M}_{\text{biased}} = 1 \cdot I(0 \leq \mu \leq 1)$$

  (a PDF on $[0, 1]$)

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $\mathcal{M}_{\text{fair}}$, we set

$$p(\mu \mid \mathcal{M}_{\text{fair}}) = I(\mu = 0.5)$$

  (a "degenerate" PMF on μ)
- For a biased, coin, we put a uniform prior on μ:

$$p(\mu \mid \mathcal{M}_{\text{biased}} = 1 \cdot I(0 \le \mu \le 1)$$

  (a PDF on $[0, 1]$)
- After 40 flips, we see 25 heads.

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $\mathcal{M}_{\text{fair}}$, we set

$$p(\mu \mid \mathcal{M}_{\text{fair}}) = I(\mu = 0.5)$$

  (a "degenerate" PMF on μ)
- For a biased, coin, we put a uniform prior on μ:

$$p(\mu \mid \mathcal{M}_{\text{biased}} = 1 \cdot I(0 \le \mu \le 1)$$

  (a PDF on $[0, 1]$)
- After 40 flips, we see 25 heads.
- This gives conditional posteriors:

$$\mu \mid \mathbf{y}, \mathcal{M}_{\text{fair}} \sim I(\mu = 0.5)$$
$$\mu \mid \mathbf{y}, \mathcal{M}_{\text{biased}} \sim \text{Beta}(25 + 1, 15 + 1)$$

Fair Coin: Prior and Posterior

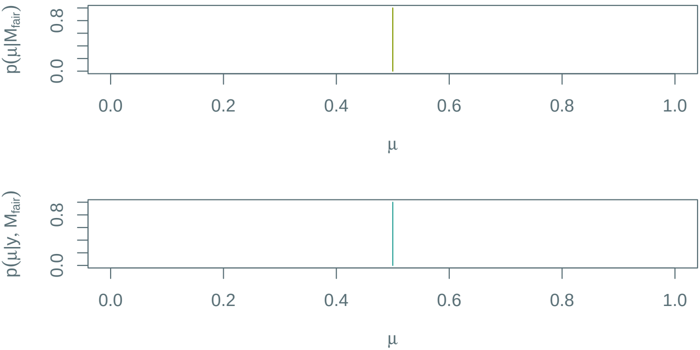


Figure: Top: Prior on μ, conditioned on the coin being fair. Bottom: Posterior on μ, conditioned on the coin being fair. Note that conditioning on the coin being fair makes the data irrelevant for inferring μ

Biased Coin: Prior and Posterior

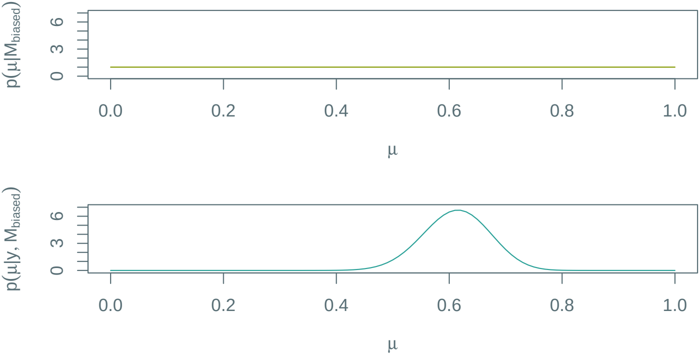


Figure: Top: Prior on μ, conditioned on the coin being biased. Bottom: Posterior on μ, conditioned on the coin being biased. When the coin can have any bias, the posterior concentrates mass near the observed proportion of heads

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$p(y \mid \mathcal{M}_{\text{biased}}) = \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \, d\mu$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$p(y \mid \mathcal{M}_{\text{biased}}) = \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \, d\mu$$
$$= \int_0^1 \binom{40}{25} \mu^y (1-\mu)^{40-y} \times 1 \, d\mu$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid \mathcal{M}_{\text{biased}}) &= \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \, d\mu \\
&= \int_0^1 \binom{40}{25} \mu^y (1-\mu)^{40-y} \times 1 \, d\mu \\
&= \binom{40}{25} \int_0^1 \mu^{26-1} (1-\mu)^{16-1} \, d\mu
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid \mathcal{M}_{\text{biased}}) &= \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \; d\mu \\
&= \binom{40}{25} \int_0^1 \mu^{26-1}(1-\mu)^{16-1} \; d\mu \\
&= \binom{40}{25} \frac{\Gamma(26)\Gamma(16)}{\Gamma(42)}
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid \mathcal{M}_{\text{biased}}) &= \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \; d\mu \\
&= \binom{40}{25} \frac{\Gamma(26)\Gamma(16)}{\Gamma(42)} \\
&= \frac{40!}{25!15!} \frac{25!15!}{41!}
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid \mathcal{M}_{\text{biased}}) &= \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \, d\mu \\
&= \frac{40!}{25!15!} \frac{25!15!}{41!} \\
&= 1/41
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid \mathcal{M}_{\text{biased}}) &= \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \; d\mu \\
&= 1/41 \\
&= \mathbf{0.0243}
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$p(y \mid \mathcal{M}_{\text{biased}}) = \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}}) \, d\mu$$
$$= \mathbf{0.0243}$$

If the coin is fair (i.e., $\mu = 0.5$ with probability 1), then the marginal likelihood is just

$$p(y \mid \mathcal{M}_{\text{fair}}) = \binom{40}{25} (1/2)^{25} (1/2)^{15} = \mathbf{0.0366}$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$p(y \mid \mathcal{M}_{\text{biased}}) = \int_0^1 p(y \mid \mu, \mathcal{M}_{\text{biased}}) p(\mu \mid \mathcal{M}_{\text{biased}})\, d\mu$$
$$= \mathbf{0.0243}$$

If the coin is fair (i.e., $\mu = 0.5$ with probability 1), then the marginal likelihood is just

$$p(y \mid \mathcal{M}_{\text{fair}}) = \binom{40}{25}(1/2)^{25}(1/2)^{15} = \mathbf{0.0366}$$

and so the "fair coin hypothesis" yields a higher **marginal likelihood** than the "Bayesian alternative" with a uniform prior.

Bayes Factor

- How does this data affect the plausibility that the coin is biased?

Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$
\begin{aligned}
\frac{p(\mathcal{M}_{\text{biased}} \mid y)}{p(\mathcal{M}_{\text{fair}} \mid y)} &= \frac{p(\mathcal{M}_{\text{biased}})p(y \mid \mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})p(y \mid \mathcal{M}_{\text{fair}})} \\
&= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times \frac{0.0243}{0.0366} \\
&= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times 0.663
\end{aligned}
$$

Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$\frac{p(\mathcal{M}_{\text{biased}} \mid y)}{p(\mathcal{M}_{\text{fair}} \mid y)} = \frac{p(\mathcal{M}_{\text{biased}})p(y \mid \mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})p(y \mid \mathcal{M}_{\text{fair}})}$$
$$= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times \frac{0.0243}{0.0366}$$
$$= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times 0.663$$

- Thus, relative to what we believed before seeing the data, our **subjective odds** that the coin is biased **should go down** after seeing 25 heads out of 40! (with the "uniform" notion of what "bias" looks like)

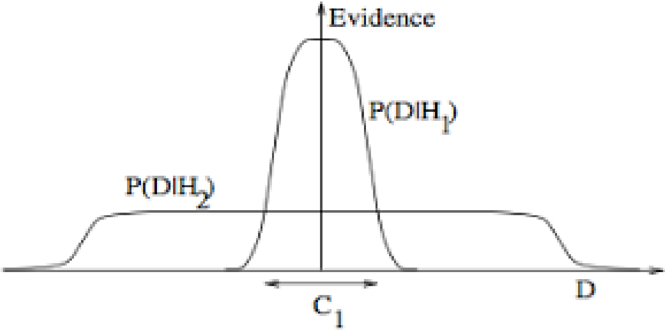
Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$
\begin{aligned}
\frac{p(\mathcal{M}_{\text{biased}} \mid y)}{p(\mathcal{M}_{\text{fair}} \mid y)} &= \frac{p(\mathcal{M}_{\text{biased}})p(y \mid \mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})p(y \mid \mathcal{M}_{\text{fair}})} \\
&= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times \frac{0.0243}{0.0366} \\
&= \frac{p(\mathcal{M}_{\text{biased}})}{p(\mathcal{M}_{\text{fair}})} \times 0.663
\end{aligned}
$$

- Thus, relative to what we believed before seeing the data, our **subjective odds** that the coin is biased **should go down** after seeing 25 heads out of 40! (with the "uniform" notion of what "bias" looks like)
- The ratio of marginal likelihoods, by which our "belief ratio" is scaled, is called the **Bayes Factor**

Conservation of Explanatory Power



Marginal likelihood “rewards” specific predictions

Conservation of Explanatory Power

Probabilistic Occam's Razor

Bayesian Occam's Razor

A "possible world" consists of a model $\mathcal{M}$, along with a (possibly trivial) parameter-setting, θ

$$p(\mathcal{M}|\mathbf{y}) = \int \frac{p(\mathcal{M},\theta)p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta \; d\theta$$

$p(\mathbf{y}|\mathcal{M},\theta)$ Rewards specific predictions by $(\mathcal{M},\theta)$

Bayesian Occam's Razor

A "possible world" consists of a model $\mathcal{M}$, along with a (possibly trivial) parameter-setting, θ

$$
\begin{aligned}
p(\mathcal{M}|\mathbf{y}) &= \int \frac{p(\mathcal{M},\theta)p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta \\
&= \int \frac{p(\mathcal{M})p(\theta|\mathcal{M})p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})}\, d\theta
\end{aligned}
$$

$p(\mathbf{y}\|\mathcal{M},\theta)$	Rewards specific predictions by $(\mathcal{M},\theta)$
$p(\theta\|\mathcal{M})$	Penalizes flexibility of the model class

Bayesian Occam's Razor

$$p(\mathcal{M}|\mathbf{y}) = \int \frac{p(\mathcal{M},\theta)p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta \; d\theta$$

$p(\mathbf{y}\|\mathcal{M},\theta)$	Rewards specific predictions by $(\mathcal{M},\theta)$
$p(\theta\|\mathcal{M})$	Penalizes flexibility of the model class

Other examples

- Polynomial regression: θ = coefficients $p(\theta \mid \mathcal{M})$ will be "spread thin" for "higher order" polynomials, since there are more possibilities
- Classification: θ = shapes of each class $p(\theta \mid \mathcal{M})$ will be "spread thin" for more complex shapes, since there are more possibilities

Bayesian Occam's Razor

$$
\begin{aligned}
p(\mathcal{M}|\mathbf{y}) &= \int \frac{p(\mathcal{M},\theta)p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta \\
&= \int \frac{p(\mathcal{M})p(\theta|\mathcal{M})p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})}\ d\theta
\end{aligned}
$$

$p(\mathbf{y}\|\mathcal{M},\theta)$	Rewards specific predictions by $(\mathcal{M},\theta)$
$p(\theta\|\mathcal{M})$	Penalizes flexibility of the model class

Other examples

- Polynomial regression: θ = coefficients $p(\theta \mid \mathcal{M})$ will be "spread thin" for "higher order" polynomials, since there are more possibilities
- Classification: θ = shapes of each class $p(\theta \mid \mathcal{M})$ will be "spread thin" for more complex shapes, since there are more possibilities