MS&E 226: Fundamentals of Data Science

Lecture 12: Bayesian inference

Ramesh Johari

Priors

Frequentist vs. Bayesian inference

- Frequentists treat the *parameters* as fixed (deterministic).
  - Considers the training data to be a random draw from the population model.
  - Uncertainty in estimates is quantified through the *sampling distribution*: what is seen if the estimation procedure is repeated over and over again, over many sets of training data ("parallel universes").

Frequentist vs. Bayesian inference

- Frequentists treat the *parameters* as fixed (deterministic).
  - Considers the training data to be a random draw from the population model.
  - Uncertainty in estimates is quantified through the *sampling distribution*: what is seen if the estimation procedure is repeated over and over again, over many sets of training data ("parallel universes").
- Bayesians treat the *parameters* as random.
  - Key element is a *prior* distribution on the parameters.
  - Using Bayes' theorem, combine prior with data to obtain a *posterior* distribution on the parameters.
  - Uncertainty in estimates is quantified through the posterior distribution.

Bayes' rule: Discrete case

Suppose that we generate discrete data $Y_1, \ldots, Y_n$, given a parameter θ that can take one of finitely many values.

Recall that the distribution of the data given θ is the *likelihood* $\mathbb{P}(\mathbf{Y} = \mathbf{y} | \theta = t)$.

The Bayesian adds to this a *prior distribution* $\mathbb{P}(\theta = t)$, expressing the belief that θ takes on a given value. Then Bayes' rule says:

POSTERIOR

$$\mathbb{P}(\theta = t | \mathbf{Y} = \mathbf{y}) = \frac{\mathbb{P}(\theta = t, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})}$$

PRIOR

$$= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} | \theta = t)\mathbb{P}(\theta = t)}{\mathbb{P}(\mathbf{Y} = \mathbf{y})}.$$

LIKELIHOOD

$$= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} | \theta = t)\mathbb{P}(\theta = t)}{\sum_\tau \mathbb{P}(\mathbf{Y} = \mathbf{y} | \theta = \tau)\mathbb{P}(\theta = \tau)}.$$

POSTERIOR $\propto$ LIKELIHOOD $\times$ PRIOR

"PROPORTIONAL TO"

Bayes' rule: Continuous case

If data and/or parameters are continuous, we use densities instead of distributions. E.g., if both data and parameters are continuous, Bayes' rule says:

$$f(\boldsymbol{\theta}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{Y})},$$

where

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\tilde{\boldsymbol{\theta}})f(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}}.$$

Bayes' rule: In words

The *posterior* is the distribution of the parameter, given the data.

Bayes' rule says:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Here "$\propto$" means "proportional to"; the missing constant is $1/f(\mathbf{Y})$, the [inverse of] unconditional probability of the data.

Note that this constant *does not depend on the parameter* $\boldsymbol{\theta}$.

Example: Biased coin flipping

We flip a biased coin 5 times, and get H, H, T, H, T. What is your estimate of the bias q?

A Bayesian starts with a *prior* for q: $f(q)$ (pdf).

Example: Biased coin flipping

We flip a biased coin 5 times, and get H, H, T, H, T. What is your estimate of the bias q?

A Bayesian starts with a *prior* for q: $f(q)$ (pdf).

This is combined with the *likelihood* of the data, given q. In our example:

$$\text{likelihood given } q = \mathbb{P}(\mathbf{Y}|q) = q^3(1-q)^2.$$

Example: Biased coin flipping

We flip a biased coin 5 times, and get H, H, T, H, T. What is your estimate of the bias q?

A Bayesian starts with a *prior* for q: $f(q)$ (pdf).

This is combined with the *likelihood* of the data, given q. In our example:

$$\text{likelihood given } q = \mathbb{P}(\mathbf{Y}|q) = q^3(1-q)^2.$$

The posterior is the distribution of q, given the data we saw; we get this using Bayes' rule:

$$f(q|\mathbf{Y}) = \frac{\mathbb{P}(\mathbf{Y}|q)f(q)}{\int_0^1 \mathbb{P}(\mathbf{Y}|q')f(q')dq'}.$$

Example: Biased coin flipping

As an example, suppose that $f(q)$ was the uniform distribution on $[0, 1]$.

Then the posterior after n flips with k H's and $n - k$ T's is:

$$f(q|\mathbf{Y}) = \frac{1}{B(k+1, n-k+1)} q^k (1-q)^{n-k},$$

the Beta$(k + 1, n - k + 1)$ distribution.

In fact: if the *prior is a* Beta(a, b) *distribution, the posterior is a* Beta$(a + k, b + n - k)$ *distribution.*[1] (The uniform distribution is a Beta$(1, 1)$ distribution.

[1] We say the beta distribution (the prior on the parameter) is *conjugate* to the binomial distribution (the likelihood).

Bayesian inference

The goals of inference

Recall the two main goals of inference:

- What is a good guess of the population model (the true parameters)?
- How do I quantify my uncertainty in the guess?

Bayesian inference answers both questions directly through the posterior.

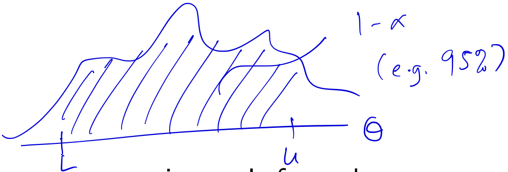
Using the posterior

The posterior can be used in many ways to estimate the parameters. For example, you might compute:

- The mean
- The median
- The mode
- etc.

Depending on context, these are all potentially useful ways to estimate model parameters from a posterior.

Using the posterior



In the same way, it is possible to construct intervals from the posterior. These are called *credible* intervals (in contrast to "confidence" intervals in frequentist inference).

Given a posterior distribution on a parameter θ, a $1 - \alpha$ credible interval $[L, U]$ is an interval such that:

$$\mathbb{P}(L \leq \theta \leq U | \mathbf{Y}) \geq 1 - \alpha.$$

Note that here, in contrast to frequentist confidence intervals, the endpoints L and U are *fixed* and the parameter θ is *random*!

Using the posterior

More generally, because the posterior is a full distribution on the parameters, it is possible to make all sorts of probabilistic statements about their values, for example:

- "I am 95% sure that the true parameter is bigger than 0.5."
- "There is a 50% chance that θ_1 is larger than θ_2.
- etc.

In Bayesian inference, you should not limit yourself to just point estimates and intervals; visualization of the posterior distribution is often quite valuable and yields significant insight.

Example: Biased coin flipping

Recall that with a Beta(a, b) prior on q, the posterior (with k heads and $n-k$ tails) is Beta$(a+k, b+n-k)$.

The mode of this distribution is $\hat{q} = \frac{a+k-1}{a+b+n-2}$.

Recall that the MLE is k/n, and the mode of the Beta(a, b) prior is $(a-1)/(a+b-2)$. So in this case we can also write:

$$\text{posterior mode} = c_n \text{MLE} + (1 - c_n)(\text{prior mode}),$$

where:

$$c_n = \frac{n}{a+b+n-2}.$$

Bayesian estimation and the MLE

The preceding example suggests a close connection between Bayesian estimation and the MLE. This is easier to see by recalling that:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

So if the prior is *flat* (i.e., uniform), then the parameter estimate that maximizes the posterior (the mode, also called the *maximum a posteriori* estimate or *MAP*) is the same as the maximum likelihood estimate.

Bayesian estimation and the MLE

The preceding example suggests a close connection between Bayesian estimation and the MLE. This is easier to see by recalling that:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

So if the prior is *flat* (i.e., uniform), then the parameter estimate that maximizes the posterior (the mode, also called the *maximum a posteriori* estimate or *MAP*) is the same as the maximum likelihood estimate.

In general:

- Uniform priors may not make sense (because, e.g., the parameter space is unbounded).
- A non-uniform prior will make the MAP estimate different from the MLE.

Example: Normal data

Suppose that $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}(\mu, 1)$. Suppose a prior on μ is that $\mu \sim \mathcal{N}(a, b^2)$. Then it can be shown that the posterior for μ is $\mathcal{N}(\hat{a}, \hat{b}^2)$, where:

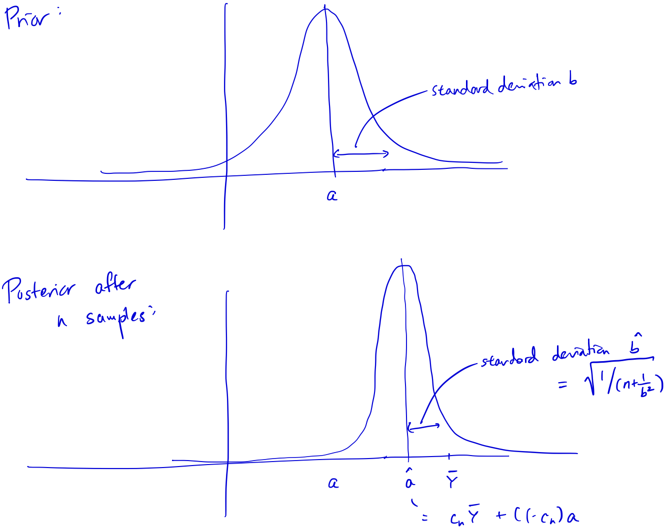
$$\begin{aligned} \hat{a} &= c_n \overline{Y} + (1 - c_n) a; \\ \hat{b}^2 &= \frac{1}{n + 1/b^2}; \\ c_n &= \frac{n}{n + 1/b^2}. \end{aligned}$$

So the MAP estimate is $\hat{a}$; and a 95% credible interval for μ is $[\hat{a} - 1.96\hat{b}, \hat{a} + 1.96\hat{b}]$.

Note that for large n, $\hat{a} \approx \overline{Y}$, and $\hat{b} \approx 1/\sqrt{n} = \text{SE}$, the frequentist standard error of the MLE.

Example: Normal data

A picture:



Maximum likelihood and Bayesian inference

The preceding observations are more general.

If the prior is "reasonable" then the posterior is *asymptotically normal*, with mean that is the MLE $\hat{\theta}_{\text{MLE}}$, and variance that is $\hat{\text{SE}}^2$, where $\hat{\text{SE}}$ is the standard error of the MLE.

So for example:

- In large samples, the posterior mean (or mode) and the MLE are approximately the same.
- In large samples, the $1 - \alpha$ normal (frequentist) confidence interval is the same as the the $1 - \alpha$ (Bayesian) credible interval.

Computation

Even simple Bayesian inference problems can rapidly become computationally intractable: computing the posterior is often not straightforward.

The last few decades have seen a revolution in computational methods for Bayesian statistics, headlined by *Markov chain Monte Carlo* (MCMC) techniques for estimating the posterior.

Though not perfect, these advances mean that you don't need to consider computational advantages when choosing one approach over another.

When do Bayesian methods work well?

Bayesian methods work well, quite simply, when *prior information matters*.

An example with biased coin flipping: if a perfectly normal coin is flipped 10 times, with 8 heads, what is your guess of the bias?

- A frequentist would say 0.8.
- A Bayesian would likely use a prior that is very strongly peaked around 0.5, so the new evidence from just ten flips would not change her belief very much.

When do Bayesian methods work poorly?

Analogously, Bayesian methods work poorly when the prior is poorly chosen.

For example, suppose you try out a new promotion on a collection of potential customers.

Previous promotions may have failed spectacularly, leading you to be pessimistic as a Bayesian.

However, as a consequence, you will be more unlikely to detect an objectively successful experiment.

Combining methods

A good frequentist estimation procedure:

- Uses only the data for inferences
- Provides guarantees on how the procedure will perform, if repeatedly used

A good Bayesian estimation procedure:

- Leverages available prior information effectively
- Combines prior information and the data into a single distribution (the posterior)
- Ensures the choice of estimate is “optimal” given the posterior (e.g., maximum *a posteriori* estimation)

Combining methods

It is often valuable to:

- Ask that Bayesian methods have good frequentist properties
- Ask that estimates computed by frequentist methods "make sense" given prior understanding

Having both approaches in your toolkit is useful for this reason.

Choosing the prior

Where did the prior come from?

There are two schools of thought on the prior:

- *Subjective* Bayesian
  - The prior is a summary of our subjective beliefs about the data.
  - E.g., in the coin flipping example: the prior for q should be strongly peaked around $1/2$.

Where did the prior come from?

There are two schools of thought on the prior:

- *Subjective* Bayesian
  - The prior is a summary of our subjective beliefs about the data.
  - E.g., in the coin flipping example: the prior for q should be strongly peaked around $1/2$.
- *Objective* Bayesian
  - The prior should be chosen in a way that is "uninformed".
  - E.g., in the coin flipping example: the prior should be uniform on $[0, 1]$.

Where did the prior come from?

There are two schools of thought on the prior:

- *Subjective* Bayesian
  - The prior is a summary of our subjective beliefs about the data.
  - E.g., in the coin flipping example: the prior for q should be strongly peaked around $1/2$.
- *Objective* Bayesian
  - The prior should be chosen in a way that is "uninformed".
  - E.g., in the coin flipping example: the prior should be uniform on $[0, 1]$.

Objective Bayesian inference was a response to the basic criticism that subjectivity should not enter into scientific conclusions. (Worth considering whether this is appropriate in a business context...)

Objectivism: Flat priors

A *flat* prior is a uniform prior on the parameter: $f(\theta)$ is constant.

As noted this doesn't make sense when θ can be unbounded...or does it?

Example: Normal data

Suppose again that $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}(\mu, 1)$, but that now we use a "prior" $f(\mu) \equiv 1$.

Of course this prior is not a probability distribution, but it turns out that we can still formally carry out the calculation of a posterior, as follows:

- The product of the *likelihood* and the *prior* is just the likelihood.
- If we can *normalize* the likelihood to be a probability distribution over μ, then this will be a well-defined posterior.

Note that in this case the posterior is just a scaled version of likelihood, so the MAP estimate (posterior mode) is *exactly the same* as the MLE!

Improper priors

This example is one where the flat prior is *improper*: it is not a probability distribution on its own, but yields a well-defined posterior.

Flat priors are sometimes held up as evidence of why Bayesian estimates are at least as informative as frequentist estimates, since at worst by using a flat prior we can recover maximum likelihood estimation.

Another way to interpret this is as a form of *conservatism*: The most conservative thing to do is to assume you have no knowledge, except what is in the data; this is what the flat prior is meant to encode.

Jeffreys' priors [$*$]

But flat priors can also lead to some unusual behavior. For example, suppose we place a flat prior on μ. What is our prior on, e.g., μ^3? It is not flat:

$$f(\mu^3) = \frac{f(\mu)}{3\mu^2} = \frac{1}{3\mu^2}.$$

This is strange: if we have no information about μ, we should have no information about μ^3.

Jeffreys' priors [$*$]

But flat priors can also lead to some unusual behavior. For example, suppose we place a flat prior on μ. What is our prior on, e.g., μ^3? It is not flat:

$$f(\mu^3) = \frac{f(\mu)}{3\mu^2} = \frac{1}{3\mu^2}.$$

This is strange: if we have no information about μ, we should have no information about μ^3.

The problem is that flat priors are not *invariant* to transformations of the parameter.

Jeffreys showed that a reasonble uninformative prior that is also transformation invariant is obtained by setting $f(\theta)$ to the inverse of the Fisher information at θ; see [AoS], Section 11.6.

Applications

Bayesian linear regression

Assume a linear normal model $Y_i = \vec{X}\boldsymbol{\beta} + \varepsilon_i$, where the ε_i are $\mathcal{N}(0, \sigma^2)$.

In Bayesian linear regression, we also put a prior distribution on $\boldsymbol{\beta}$.

Here we look at two examples of this approach.

Bayesian linear regression: Normal prior

Suppose that the prior distribution on the coefficients $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda\sigma^2}\mathbf{I}\right).$$

In other words: the higher λ is, the higher the prior "belief" that $\boldsymbol{\beta}$ is close to zero.[2]

For this prior, the MAP estimator is the same as the *ridge regression* solution.

[2]Note that this is a "partial" Bayesian solution, since σ^2 is assumed known. In practice an estimate of σ^2 is used.

Bayesian linear regression: Laplace prior

Now suppose instead that the prior is that all the β_j are independent, each with density:

$$f(\beta) = \left(\frac{\lambda}{2\sigma}\right) \exp\left(-\frac{\lambda|\beta|}{\sigma}\right).$$

This is called the *Laplace* distribution; it is symmetric around zero, and more strongly peaked as λ grows.[3]

With this prior, the MAP estimator is the same as the *lasso* solution.

[3] Note that this is again a solution that assumes σ^2 is known. In practice an estimate of σ^2 is used.

Bayesian model selection

The BIC (Bayesian information criterion) is obtained from a Bayesian view of model selection.

Basic idea:

- Imagine there is a prior over possible models.

Bayesian model selection

The BIC (Bayesian information criterion) is obtained from a Bayesian view of model selection.

Basic idea:

- Imagine there is a prior over possible models.
- We would want to choose the model that has the highest posterior probability.

Bayesian model selection

The BIC (Bayesian information criterion) is obtained from a Bayesian view of model selection.

Basic idea:

- Imagine there is a prior over possible models.
- We would want to choose the model that has the highest posterior probability.
- In large samples, the effect of this prior becomes small relative to the effect of the data, so asymptotically BIC estimates the posterior probability of a model by (essentially) ignoring the effect of the prior.

Bayesian model selection

The BIC (Bayesian information criterion) is obtained from a Bayesian view of model selection.

Basic idea:

- Imagine there is a prior over possible models.
- We would want to choose the model that has the highest posterior probability.
- In large samples, the effect of this prior becomes small relative to the effect of the data, so asymptotically BIC estimates the posterior probability of a model by (essentially) ignoring the effect of the prior.
- Choosing the model that maximizes BIC is like choosing the model that has the highest posterior probability in this sense.