Machine Learning

Lecture 4: Probabilistic Inference

Prof. Dr. Aleksandar Bojchevski

23.04.24

Outline

Maximum likelihood estimation

Bayesian inference

Maximum a posteriori estimation

Fully Bayesian

Posterior predictive distribution

Coin flips

We flip the same coin 10 times:

H T H H T H H H T H

Probability that the next coin flip is T?

Coin flips

We flip the same coin 10 times:

H T H H T H H H T H

Probability that the next coin flip is T?

$\sim 0 \qquad \sim 0.3 \qquad \sim 0.38 \qquad \sim 0.5 \qquad \sim 0.76 \qquad \sim 1$

Coin flips

30% seems reasonable, but why?

Coin flips

30% seems reasonable, but why?

Every flip is random. So every sequence of flips is random – it has some probability to be observed.

Coin flips

30% seems reasonable, but why?

Every flip is random. So every sequence of flips is random – it has some probability to be observed.

For the i-th coin flip we write $p_i(F_i = \text{T}) = \theta_i$.

Coin flips

30% seems reasonable, but why?

Every flip is random. So every sequence of flips is random – it has some probability to be observed.

For the i-th coin flip we write $p_i(F_i = \text{T}) = \theta_i$.

To denote that the probability distribution depends on θ_i, we write

$$p_i(F_i = \text{T} \mid \theta_i) = \mathrm{Ber}(F_i = \text{T} \mid \theta_i) = \theta_i$$

i.e. $F_i \sim \mathrm{Ber}(\theta_i)$.

Note the i in the index! We are trying to reason about θ_{11}.

Coin flips

All the randomness of a sequence of flips is governed (*modeled*) by the parameters $\theta_1, \dots, \theta_{10}$:

$$p(\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{H}\,\text{T}\,\text{H} \mid \theta_1, \theta_2, \dots, \theta_{10})$$

What do we know about $\theta_1, \dots, \theta_{10}$? Can we infer something about θ_{11}?
At first sight, there is no connection.

Find θ_i's such that the $p(\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{H}\,\text{T}\,\text{H} \mid \theta_1, \theta_2, \dots, \theta_{10})$ is as high as possible. This is a very important principle:

Maximum likelihood: maximize the likelihood of our observations.

Coin flips

We need to model $p(\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{H}\,\text{T}\,\text{H} \mid \theta_1, \theta_2, \ldots, \theta_{10})$.

First assumption: The coin flips do not affect each other – **independence**.

$$
\begin{aligned}
& p(\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{H}\,\text{T}\,\text{H}) \mid \theta_1, \theta_2, \ldots, \theta_{10}) \\
= & \, p_1(F_1 = \text{H} \mid \theta_1) \cdot p_2(F_2 = \text{H} \mid \theta_2) \cdot \cdots \cdot p_{10}(F_{10} = \text{H} \mid \theta_{10}) \\
= & \prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i)
\end{aligned}
$$

Notice the i in p_i, θ_i! This indicates that the coin flip at time 1 is different from the one at time 2, time 3, ... But the coin does not change.

Coin flips

Second assumption: The flips are qualitatively the same – **identical distribution**.

$$\prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i) = \prod_{i=1}^{10} p(F_i = f_i \mid \theta)$$

In total: The 10 flips are **independent and identically distributed (i.i.d.).**

Remember θ_{11}? With the i.i.d. assumption we can link it to $\theta_1, \ldots, \theta_{10}$.

Now we can write down the probability of our sequence with respect to θ:

$$\begin{aligned}\prod_{i=1}^{10} p(F_i = f_i \mid \theta) &= (1-\theta)\theta(1-\theta)(1-\theta)\theta(1-\theta)(1-\theta)(1-\theta)\theta(1-\theta) \\ &= \theta^3(1-\theta)^7\end{aligned}$$

Coin flips

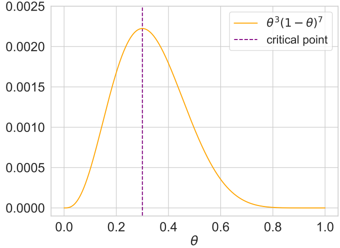
Under our model assumptions (i.i.d.): $p(\underbrace{\text{H T H H T H H H T H}}_{\text{observed data, } \mathcal{D}} \mid \theta) = \theta^3(1-\theta)^7$

Coin flips

Under our model assumptions (i.i.d.): $p(\underbrace{\text{H T H H T H H H T H}}_{\text{observed data, } \mathcal{D}} \mid \theta) = \theta^3(1-\theta)^7$

This can be interpreted as a function $f(\theta) := p(\mathcal{D} \mid \theta)$. We want to find the maxima of this function (maximum likelihood).

Our goal: $\theta_{\text{MLE}} = \arg\max_{\theta \in [0,1]} f(\theta)$.



Very important: the likelihood function is not a probability distribution over θ since $\int p(\mathcal{D} \mid \theta) d\theta \neq 1$ in general.

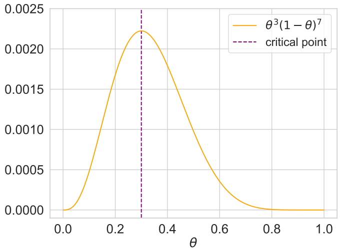
How do we maximize the likelihood function?

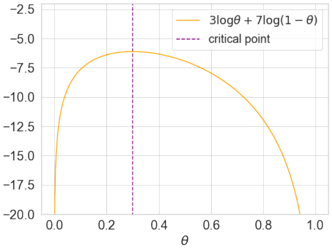
Take the derivative $\frac{\mathrm{d}f}{\mathrm{d}\theta}$, set it to 0, and solve for θ. Check these *critical points* by checking the second derivative.

This is possible, but even for our simple $f(\theta)$ the math is rather ugly.

Can we simplify the problem? Monotonic functions preserve critical points!

$$\arg\max_{\theta\in[0,1]} f(\theta) = \arg\max_{\theta\in[0,1]} \log f(\theta)$$





Maximum Likelihood Estimation (MLE)

Can we generalize this to *any* coin sequence?

Maximum Likelihood Estimation (MLE)

Can we generalize this to *any* coin sequence? We get (derivation in the exercises):

$$\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$$

where $|T|, |H|$ denote number of T, H, respectively.

Remember we wanted to find the probability the next coin flip is T

$$F_{11} \sim \text{Ber}(\theta_{\text{MLE}})$$

$$p(F_{11} = \text{T} \mid \theta_{\text{MLE}}) = \text{Ber}(F_{11} = \text{T} \mid \theta_{\text{MLE}}) = \theta_{\text{MLE}} = \frac{|T|}{|T| + |H|}$$

This justifies 30% as a reasonable answer to our initial question.

Problem solved?!

MLE for a different sequence

Just for fun, a totally different sequence (*same coin!*):

MLE for a different sequence

Just for fun, a totally different sequence (*same coin!*):

H H

$\theta_{\text{MLE}} = 0$. But even a fair coin ($\theta = 0.5$) has 25% chance of showing this result!

The MLE solution seems counter-intuitive. Why?

MLE for a different sequence

Just for fun, a totally different sequence (*same coin!*):

$\theta_{\mathrm{MLE}} = 0$. But even a fair coin ($\theta = 0.5$) has 25% chance of showing this result!

The MLE solution seems counter-intuitive. Why?

We have **prior beliefs**: *"Coins usually don't land heads all the time".*

How can we

- represent such beliefs mathematically?
- incorporate them into our model?

Outline

Maximum likelihood estimation

Bayesian inference

Maximum a posteriori estimation

Fully Bayesian

Posterior predictive distribution

Priors

How can we represent our beliefs about θ mathematically?

Bayesian interpretation: the **prior distribution** $p(\theta)$ reflects our subjective beliefs about θ, **before** we observe any data.

Priors

How can we represent our beliefs about θ mathematically?

Bayesian interpretation: the **prior distribution** $p(\theta)$ reflects our subjective beliefs about θ, **before** we observe any data.

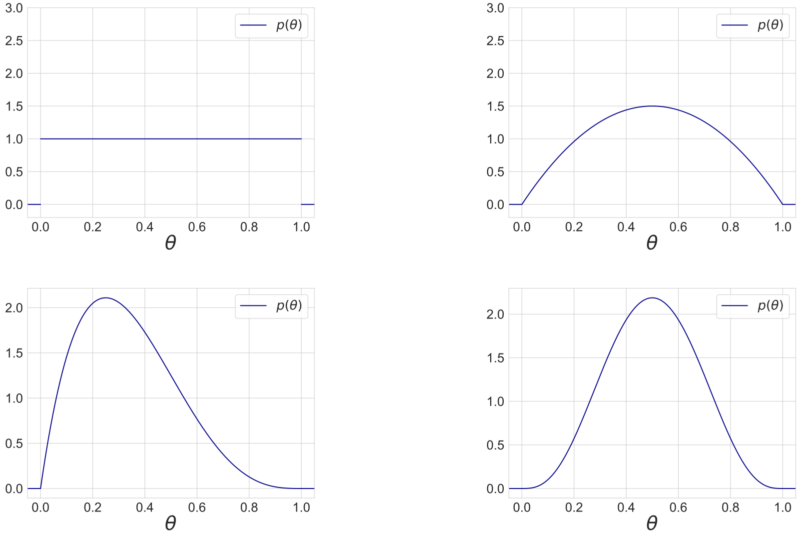
How do we choose $p(\theta)$? The only constraints are:

1. It **must not** depend on the data
2. $p(\theta) \geq 0$ for all θ
3. $\int p(\theta)\mathrm{d}\theta = 1$

Properties 2 and 3 have to hold on the support (i.e., feasible values) of θ. In our setting, only values $\theta \in [0, 1]$ make sense.

This leaves room for (possibly subjective) model choices!

Some possible choices for the prior on θ



Bayes formula

Tells us how to update our beliefs about θ after observing the data $\mathcal{D}$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) \cdot p(\theta)}{p(\mathcal{D})}$$

Here, $p(\theta \mid \mathcal{D})$ is the **posterior** distribution.

It encodes our beliefs in the value of θ **after** observing data.

The posterior depends on the following terms:

- $p(\mathcal{D} \mid \theta)$ is the **likelihood**.
- $p(\theta)$ is the **prior** that encodes our beliefs before observing the data.
- $p(\mathcal{D})$ is the **evidence** that acts as a normalizing constant that ensures that the posterior distribution integrates to 1.

Bayes formula

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) \cdot p(\theta)}{p(\mathcal{D})}$$

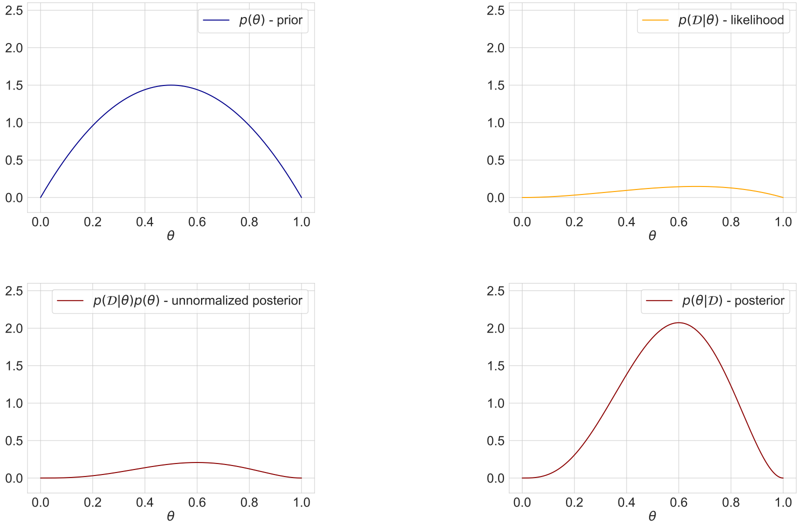
posterior $\propto$ likelihood $\cdot$ prior

We usually specify our model via the likelihood $p(\mathcal{D} \mid \theta)$ and the prior $p(\theta)$.

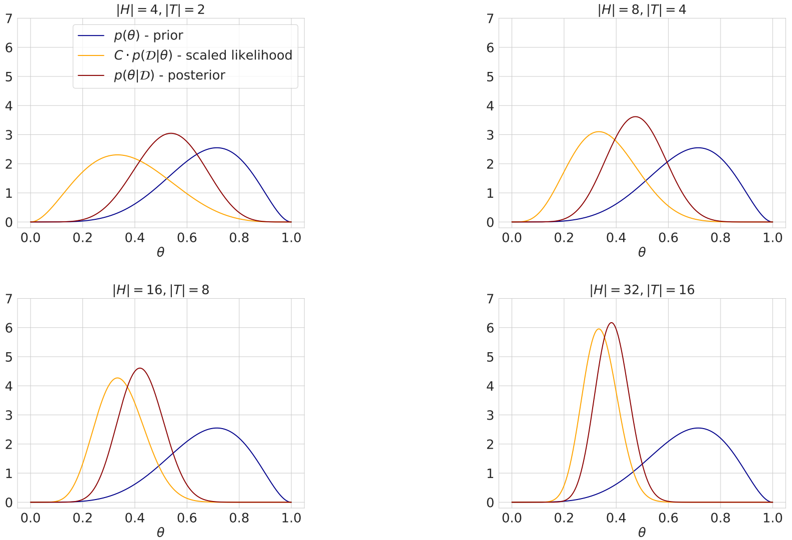
We can obtain the evidence using the sum rule of probability

$$p(\mathcal{D}) = \int p(\mathcal{D}, \theta)\mathrm{d}\theta = \int p(\mathcal{D} \mid \theta)p(\theta)\mathrm{d}\theta$$

The Bayes formula tells us how to update our beliefs given data



Observing more data increases our confidence



Priors

What happens if $p(\theta) = 0$ for some particular θ?

Priors

What happens if $p(\theta) = 0$ for some particular θ?

Recall:

$$\begin{array}{ccccc} \text{posterior} & \propto & \text{likelihood} & \cdot & \text{prior} \\ p(\theta \mid \mathcal{D}) & \propto & p(\mathcal{D} \mid \theta) & \cdot & p(\theta) \end{array}$$

Priors

What happens if $p(\theta) = 0$ for some particular θ?

Recall:

$$\begin{array}{ccccc} \text{posterior} & \propto & \text{likelihood} & \cdot & \text{prior} \\ p(\theta \mid \mathcal{D}) & \propto & p(\mathcal{D} \mid \theta) & \cdot & p(\theta) \end{array}$$

Posterior will always be zero for that particular θ regardless of the likelihood/data.

Outline

Maximum likelihood estimation

Bayesian inference

Maximum a posteriori estimation

Fully Bayesian

Posterior predictive distribution

Back to the coin flips

How do we estimate θ from the data?

In MLE, we were asking the wrong question:

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta} \; p(\mathcal{D} \mid \theta)$$

MLE ignores our prior beliefs and performs poorly if little data is available.

Back to the coin flips

How do we estimate θ from the data?

In MLE, we were asking the wrong question:

$$\theta_{\text{MLE}} = \arg\max_{\theta} \; p(\mathcal{D} \mid \theta)$$

MLE ignores our prior beliefs and performs poorly if little data is available.

Actually, we should care about the posterior distribution $p(\theta \mid \mathcal{D})$.

What if we instead maximize the posterior probability?

$$\theta_{\text{MAP}} = \arg\max_{\theta} \; p(\theta \mid \mathcal{D})$$

This approach is called **maximum a posteriori (MAP)** estimation.

Maximum a posterior estimation

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta}\ p(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta}\ \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}\end{aligned}$$

We can ignore $\frac{1}{p(\mathcal{D})}$ since it's a (positive) constant independent of θ

$$\theta_{\text{MAP}} = \arg\max_{\theta}\ p(\mathcal{D} \mid \theta)p(\theta)$$

We already know the likelihood $p(\mathcal{D} \mid \theta)$, how do we choose the prior $p(\theta)$?

Choosing the prior

Often, we choose the prior to make subsequent calculations easier.

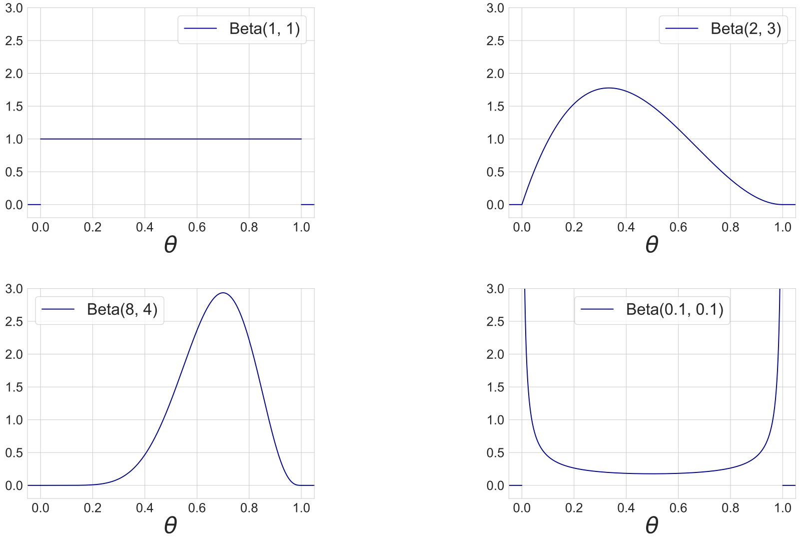
We choose Beta distribution for reasons that will become clear later.

$$\mathrm{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0,1]$$

where

- $a > 0, b > 0$ are the distribution parameters,
- $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ is the gamma function.

The PDF of the Beta for different choices of a and b



Putting everything together

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) \cdot p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \theta) \cdot p(\theta)$$

because $p(\mathcal{D})$ is constant w.r.t. θ.

Putting everything together

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) \cdot p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \theta) \cdot p(\theta)$$

because $p(\mathcal{D})$ is constant w.r.t. θ.

We know

$$p(\mathcal{D} \mid \theta) = \theta^{|T|}(1-\theta)^{|H|},$$

$$p(\theta) \equiv p(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

So we get:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto \theta^{|T|}(1-\theta)^{|H|} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}. \end{aligned}$$

MAP

We are looking for

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \; \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}\end{aligned}$$

As before, the problem becomes much easier if we consider the logarithm

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} \; \log p(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \, (|T|+a-1)\log\theta + (|H|+b-1)\log(1-\theta)\end{aligned}$$

With some algebra we obtain

$$\theta_{\text{MAP}} = \frac{|T|+a-1}{|H|+|T|+a+b-2}$$

Outline

Maximum likelihood estimation

Bayesian inference

Maximum a posteriori estimation

Fully Bayesian

Posterior predictive distribution

Estimating the posterior distribution

What we have so far

$$\theta_{\text{MAP}} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$$

The most probable value of θ under the posterior distribution.

Is this the best we can do?

- How certain are we in our estimate?
- What is the probability that θ lies in some interval?

For this, we need to consider the **entire posterior** distribution $p(\theta \mid \mathcal{D})$, not just its mode θ_{MAP}.

Unnormalized posterior

We know the posterior up to a normalizing constant (slide 24)

$$p(\theta \mid \mathcal{D}) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}.$$

Finding the true posterior $p(\theta \mid \mathcal{D})$ boils down to finding the normalization constant, such that the distribution integrates to 1.

Unnormalized posterior

We know the posterior up to a normalizing constant (slide 24)

$$p(\theta \mid \mathcal{D}) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}.$$

Finding the true posterior $p(\theta \mid \mathcal{D})$ boils down to finding the normalization constant, such that the distribution integrates to 1.

Option 1: Brute-force calculation $\int_0^1 \theta^{|T|+a-1}(1-\theta)^{|H|+b-1} \mathrm{d}\theta$.
This is tedious, difficult and boring. Any alternatives?

Option 2: Pattern matching. The unnormalized posterior $\theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$ looks similar to the PDF of the Beta distribution

$$\mathrm{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Can we use this fact?

Normalized posterior

The unnormalized posterior

$$p(\theta \mid \mathcal{D}) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$$

The Beta distribution

$$\mathrm{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Normalized posterior

The unnormalized posterior

$$p(\theta \mid \mathcal{D}) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$$

The Beta distribution

$$\mathrm{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Thus, we can conclude that the appropriate normalizing constant is

$$\frac{\Gamma(|T|+a+|H|+b)}{\Gamma(|T|+a)\Gamma(|H|+b)}$$

and the posterior is a Beta distribution

$$p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a+|T|, b+|H|)$$

Remember this when solving integrals involving known (up to a constant) pdfs.

Conjugate priors

We started with the following prior distribution

$$p(\theta) = \mathrm{Beta}(\theta \mid a, b)$$

And obtained the following posterior

$$p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a + |T|, b + |H|)$$

Was this just a lucky coincidence?

Conjugate priors

We started with the following prior distribution

$$p(\theta) = \text{Beta}(\theta \mid a, b)$$

And obtained the following posterior

$$p(\theta \mid \mathcal{D}) = \text{Beta}(\theta \mid a + |T|, b + |H|)$$

Was this just a lucky coincidence?

No, this is an instance of a more general principle. Beta distribution is a **conjugate prior** for the Bernoulli likelihood.

If a prior is conjugate for the given likelihood, then the posterior will be of the same family as the prior.

Conjugate priors

We started with the following prior distribution

$$p(\theta) = \mathrm{Beta}(\theta \mid a, b)$$

And obtained the following posterior

$$p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a + |T|, b + |H|)$$

Was this just a lucky coincidence?

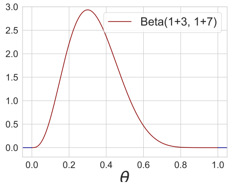
No, this is an instance of a more general principle. Beta distribution is a **conjugate prior** for the Bernoulli likelihood.

If a prior is conjugate for the given likelihood, then the posterior will be of the same family as the prior.

In our case, we can interpret the parameters a, b of the prior as the number of tails and heads that we saw in the past.

Advantages of the fully Bayesian approach

We have an entire distribution!
Not just a point estimate.



We can answer questions such as:

- What is the expected value of θ under $p(\theta \mid \mathcal{D})$?
- What is the variance of $p(\theta \mid \mathcal{D})$?
- Find a *credible interval* $[\theta_1, \theta_2]$, such that $\Pr(\theta \in [\theta_1, \theta_2] \mid \mathcal{D}) = 95\%$[1].

[1] Not to be confused with frequentist confidence intervals (see Lecture 5).

Three approaches for parameter estimation

Maximum likelihood estimation (MLE)

- Goal: Optimization problem $\max_\theta \log p(\mathcal{D} \mid \theta)$
- Result: Point estimate θ_{MLE}
- Coin example: $\theta_{\mathrm{MLE}} = \frac{|T|}{|T|+|H|}$

Maximum a posteriori (MAP) estimation

- Goal: Optimization problem $\max_\theta \log p(\theta \mid \mathcal{D})$
- Result: Point estimate θ_{MAP}
- Coin example: $\theta_{\mathrm{MAP}} = \frac{|T|+a-1}{|T|+|H|+a+b-2}$

Estimating the posterior distribution

- Goal: Find the normalizing constant $p(\mathcal{D})$
- Result: Full distribution $p(\theta \mid \mathcal{D})$
- Coin example: $p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a + |T|, b + |H|)$

The three approaches are closely connected

The posterior distribution is

$$p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a + |T|, b + |H|).$$

Recall that the mode of $\mathrm{Beta}(\alpha, \beta)$ is $\frac{\alpha-1}{\alpha+\beta-2}$, for $\alpha, \beta > 1$.

We see that the MAP solution is the mode of the posterior distribution

$$\theta_{\mathrm{MAP}} = \frac{|T| + a - 1}{|H| + |T| + a + b - 2}$$

If we choose a uniform prior (i.e. $a = b = 1$) we obtain the MLE solution

$$\theta_{\mathrm{MLE}} = \frac{|T| + 1 - 1}{|H| + |T| + 1 + 1 - 2} = \frac{|T|}{|H| + |T|}$$

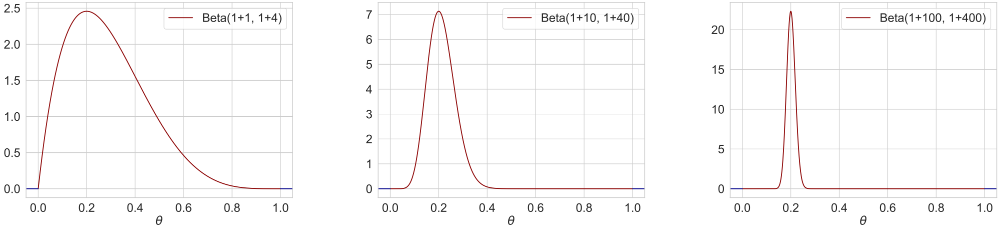
All these nice formulas are a consequence of choosing a conjugate prior. Had we chosen a non-conjugate prior, $p(\theta \mid \mathcal{D})$ and θ_{MAP} might not have a closed form.

Posterior for different number of observations

We had $p(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid a + |T|, b + |H|)$.

Visualize the posterior (for the prior $a = b = 1$):

$|T| = 1, |H| = 4$ $\quad$ $|T| = 10, |H| = 40$ $\quad$ $|T| = 100, |H| = 400$



With more data the posterior becomes more peaky – we are more certain about our estimate of θ.

Alternative view: a frequentist perspective

For MLE we had $\theta_{\mathrm{MLE}} = \frac{|T|}{|T|+|H|}$

Clearly, we get the same result for $|T| = 1, |H| = 4$ and $|T| = 10, |H| = 40$. Which one is *better*? Why?

Alternative view: a frequentist perspective

For MLE we had $\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$

Clearly, we get the same result for $|T| = 1, |H| = 4$ and $|T| = 10, |H| = 40$. Which one is *better*? Why?

How many flips? Hoeffding's Inequality for a *sampling complexity bound*:

$$p(|\theta_{\text{MLE}} - \theta| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2} \leq \delta,$$

where $N = |T| + |H|$.

For example, I want to know θ, within $\epsilon = 0.1$ error, with probability at least $1 - \delta = 0.99$. We have:

$$N \geq \frac{\ln(2/\delta)}{2\epsilon^2} \quad \rightarrow \quad N \approx 265$$

Outline

Maximum likelihood estimation

Bayesian inference

Maximum a posteriori estimation

Fully Bayesian

Posterior predictive distribution

To predict the next coin flip

For MLE:

1. Estimate $\theta_{\text{MLE}} = \frac{|T|}{|H|+|T|}$ from the data.
2. The probability that the next flip lands tails is

$$p(F_{11} = \text{T} \mid \theta_{\text{MLE}}) = \text{Ber}(F_{11} = \text{T} \mid \theta_{\text{MLE}}) = \theta_{\text{MLE}}$$

Similarly, for MAP:

1. Estimate $\theta_{\text{MAP}} = \frac{|T|+a-1}{|H|+|T|+a+b-2}$ from the data.
2. The probability that the next flip lands tails is

$$p(F_{11} = \text{T} \mid \theta_{\text{MAP}}) = \text{Ber}(F_{11} = \text{T} \mid \theta_{\text{MAP}}) = \theta_{\text{MAP}}$$

What if we have the entire posterior?

We have estimated the posterior distribution $p(\theta \mid \mathcal{D}, a, b)$ of the parameter θ.

Now, we want to compute the probability that the next coin flip is T, given observations $\mathcal{D}$ and prior belief a, b:

$$p(F = \text{T} \mid \mathcal{D}, a, b)$$

This distribution is called the **posterior predictive** distribution.

This is **different** from the posterior over the parameters $p(\theta \mid \mathcal{D}, a, b)$!

Posterior predictive distribution

For simplicity, denote the outcome of the next flip as $f \in \{0, 1\}$.

$$p(F = f \mid \mathcal{D}, a, b) = p(f \mid \mathcal{D}, a, b)$$

We already know the posterior over the parameters $p(\theta \mid \mathcal{D}, a, b)$.

Posterior predictive distribution

For simplicity, denote the outcome of the next flip as $f \in \{0, 1\}$.

$$p(F = f \mid \mathcal{D}, a, b) = p(f \mid \mathcal{D}, a, b)$$

We already know the posterior over the parameters $p(\theta \mid \mathcal{D}, a, b)$.

Using the sum rule of probability

$$\begin{aligned} p(f \mid \mathcal{D}, a, b) &= \int_0^1 p(f, \theta \mid \mathcal{D}, a, b) \mathrm{d}\theta \\ &= \int_0^1 p(f \mid \theta, \mathcal{D}, a, b) p(\theta \mid \mathcal{D}, a, b) \mathrm{d}\theta \\ &= \int_0^1 p(f \mid \theta) p(\theta \mid \mathcal{D}, a, b) \mathrm{d}\theta \end{aligned}$$

The last equality follows from the conditional independence assumption: "If we know θ, the next flip f is independent of the previous flips $\mathcal{D}$."

Fully Bayesian analysis

Recall that $p(f \mid \theta) = \mathrm{Ber}(f \mid \theta) = \theta^f (1-\theta)^{1-f}$

and $p(\theta \mid \mathcal{D}, a, b) = \frac{\Gamma(|T|+a+|H|+b)}{\Gamma(|T|+a)\Gamma(|H|+b)} \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$.

Substituting these expressions and doing some (boring) algebra we get

$$\begin{aligned} p(f \mid \mathcal{D}, a, b) &= \int_0^1 p(f \mid \theta) p(\theta \mid \mathcal{D}, a, b) \mathrm{d}\theta \\ &= \frac{(|T|+a)^f (|H|+b)^{(1-f)}}{|T|+a+|H|+b} \end{aligned}$$

Fully Bayesian analysis

Recall that $p(f \mid \theta) = \mathrm{Ber}(f \mid \theta) = \theta^f (1-\theta)^{1-f}$

and $p(\theta \mid \mathcal{D}, a, b) = \frac{\Gamma(|T|+a+|H|+b)}{\Gamma(|T|+a)\Gamma(|H|+b)} \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$.

Substituting these expressions and doing some (boring) algebra we get

$$
\begin{aligned}
p(f \mid \mathcal{D}, a, b) &= \int_0^1 p(f \mid \theta) p(\theta \mid \mathcal{D}, a, b) \mathrm{d}\theta \\
&= \frac{(|T|+a)^f (|H|+b)^{(1-f)}}{|T|+a+|H|+b} \\
&= \mathrm{Ber}\left(f \,\middle|\, \frac{|T|+a}{|T|+a+|H|+b}\right)
\end{aligned}
$$

Note that the posterior predictive distribution doesn't contain θ — we have marginalized it out!

Prediction using different approaches

- MLE: $p(F = \text{T} \mid \theta_{\text{MLE}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|}{|T|+|H|}\right)$
- MAP: $p(F = \text{T} \mid \theta_{\text{MAP}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a-1}{|T|+a+|H|+b-2}\right)$
- Fully Bayesian: $p(F = \text{T} \mid \mathcal{D}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a}{|T|+a+|H|+b}\right)$

Prediction using different approaches

- MLE: $p(F = \text{T} \mid \theta_{\text{MLE}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|}{|T|+|H|}\right)$
- MAP: $p(F = \text{T} \mid \theta_{\text{MAP}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a-1}{|T|+a+|H|+b-2}\right)$
- Fully Bayesian: $p(F = \text{T} \mid \mathcal{D}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a}{|T|+a+|H|+b}\right)$

Given the prior $a = b = 5$ and the counts $|T| = 4, |H| = 8$

$$p_{\text{MLE}} = \frac{4}{12} \approx 0.33 \qquad p_{\text{MAP}} = \frac{8}{20} = 0.40 \qquad p_{\text{FB}} = \frac{9}{22} \approx 0.41$$

Prediction using different approaches

- MLE: $p(F = \text{T} \mid \theta_{\text{MLE}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|}{|T|+|H|}\right)$
- MAP: $p(F = \text{T} \mid \theta_{\text{MAP}}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a-1}{|T|+a+|H|+b-2}\right)$
- Fully Bayesian: $p(F = \text{T} \mid \mathcal{D}) = \text{Ber}\left(F = \text{T} \mid \frac{|T|+a}{|T|+a+|H|+b}\right)$

Given the prior $a = b = 5$ and the counts $|T| = 4, |H| = 8$

$$p_{\text{MLE}} = \frac{4}{12} \approx 0.33 \qquad p_{\text{MAP}} = \frac{8}{20} = 0.40 \qquad p_{\text{FB}} = \frac{9}{22} \approx 0.41$$

How about if we have $|T| = 304, |H| = 306$?

$$p_{\text{MLE}} = \frac{304}{610} \approx 0.50 \qquad p_{\text{MAP}} = \frac{308}{618} \approx 0.50 \qquad p_{\text{FB}} = \frac{309}{620} \approx 0.50$$

As we observe lots of data, the differences **in predictions** become less noticeable.

Summary

Three approaches to parameter estimation:

- Maximum likelihood: ignores prior information.
- Maximum a posteriori: finds the mode of the posterior.
- Fully Bayesian analysis: uses the entire posterior.

Posterior $\propto$ Likelihood $\cdot$ Prior.

The i.i.d. assumption.

Monotonic transforms for optimization.

Solving integrals by reverse-engineering densities (conjugate prior).

Reading material

Main reading

- ”Machine Learning: A Probabilistic Perspective” by Murphy
  [ch. 3.1 - 3.3]

Extra reading

- ”Probabilistic Machine Learning: An Introduction” by Murphy
  [ch. 4.2, 4.6]

Slides based on an older version by S. Gunnemann, themeselves based on a version by M. Sölch.