Ekaterina Lobacheva

Full Bayesian inference

2021

Slides are partially based on lectures of Dmitry Vetrov, Dmitry Kropotov and Kirill Struminsky, deepbayes.ru/2018

Bayesian ML models

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\boxed{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \boxed{\int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta}$$

When the integrals are tractable?

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \qquad \longrightarrow \qquad p(\theta \mid x) \in \mathcal{A}\left(\alpha'\right)$$

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \qquad \longrightarrow \qquad p(\theta \mid x) \in \mathcal{A}\left(\alpha'\right)$$

Intuition:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta}$$

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \quad \longrightarrow \quad p(\theta \mid x) \in \mathcal{A}(\alpha')$$

Intuition:

$$p(\theta \mid x) = \frac{\boxed{p(x \mid \theta)p(\theta)}}{\int p(x \mid \theta)p(\theta)d\theta} \longleftarrow \text{conjugate}$$

- Denominator is tractable since any distribution in $\mathcal{A}$ is normalized
- All we need is to compute α'

Full Bayesian inference

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\boxed{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \boxed{\int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta}$$

Integrals are tractable if prior and likelihood are conjugate

Full Bayesian inference

- Easy to use - analytical formulas for training and testing stages
- Strong assumptions on the model - conjugacy of prior and likelihood
    - → Choose conjugate prior
    - → Only simple models (not flexible enough for most of the cases)

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \ldots, x_n)\,, \quad x \in \{0, 1\}$

Head (H)

Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta) p(\theta)$$

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \dots, x_n)\,, \quad x \in \{0, 1\}$

Head (H) Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta)p(\theta)$$

Likelihood: $Bern(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \ldots, x_n)\,, \quad x \in \{0, 1\}$

Head (H)

Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta) p(\theta)$$

Likelihood: $Bern(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$

Prior: ???

Example: coin tossing

How to choose a prior?

- Correct domain: $\theta \in [0, 1]$
- Include prior knowledge: a coin is most likely fair
- Inference complexity: use conjugate prior

Example: coin tossing

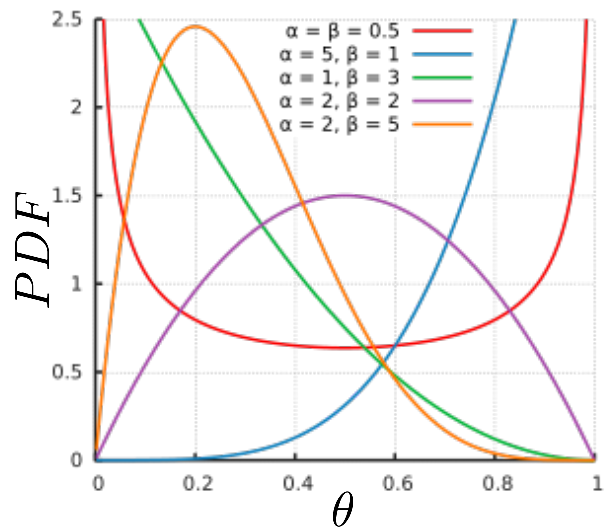
How to choose a prior?

- Correct domain: $\theta \in [0, 1]$
- Include prior knowledge: a coin is most likely fair
- Inference complexity: use conjugate prior

Beta distribution matches all requirements:

$$Beta(\theta \mid a, b) = \frac{1}{\mathrm{B}(a, b)} \theta^{a-1} (1-\theta)^{b-1}$$

Beta distribution



Example: coin tossing

How to choose a prior?

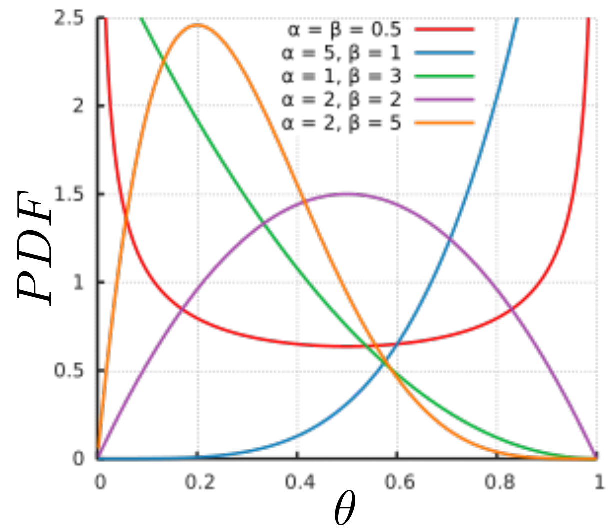
- Correct domain: $\theta \in [0, 1]$
- Include prior knowledge: a coin is most likely fair
- Inference complexity: use conjugate prior

Beta distribution matches all requirements:

$$Beta(\theta \mid a, b) = \frac{1}{\mathrm{B}(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

* May be also used for the case of most likely unfair coin

Beta distribution



Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = C\theta^C (1-\theta)^C$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = C\theta^C(1-\theta)^C$$

$$p(\theta \mid x) = \frac{1}{C}p(x \mid \theta)p(\theta) = \frac{1}{C}\theta^x(1-\theta)^{1-x}\frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1} =$$

$$= C\theta^C(1-\theta)^C$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = \boxed{C\theta^C (1-\theta)^C} \text{ conjugacy}$$

$$p(\theta \mid x) = \frac{1}{C} p(x \mid \theta) p(\theta) = \frac{1}{C} \theta^x (1-\theta)^{1-x} \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} =$$

$$= \boxed{C\theta^C (1-\theta)^C} \text{ conjugacy}$$

Here different constants are denoted with the same letter C for demonstration reasons.

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) =$$

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z}\left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right] p(\theta) =$$

$$= \frac{1}{Z}\left[\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\right] \frac{1}{\mathrm{B}(a, b)} \theta^{a-1}(1-\theta)^{b-1} =$$

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$
\begin{aligned}
p(\theta \mid X) &= \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) = \\
&= \frac{1}{Z} \left[\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right] \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} = \\
&= \frac{1}{Z'} \theta^{a + \sum_{i=1}^{n} x_i - 1} (1-\theta)^{b + n - \sum_{i=1}^{n} x_i - 1}
\end{aligned}
$$

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z}\left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right] p(\theta) =$$

$$= \frac{1}{Z}\left[\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\right] \frac{1}{\mathrm{B}(a, b)} \theta^{a-1}(1-\theta)^{b-1} =$$

$$= \frac{1}{Z'} \theta^{a+\sum_{i=1}^{n} x_i - 1}(1-\theta)^{b+n-\sum_{i=1}^{n} x_i - 1} = Beta\left(\theta \mid a', b'\right)$$

New parameters: $\quad a' = a + \sum_{i=1}^{n} x_i \qquad b' = b + n - \sum_{i=1}^{n} x_i$

Full Bayesian inference

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\boxed{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \boxed{\int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta}$$

Integrals are tractable if prior and likelihood are conjugate

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max \boxed{p\left(\theta \mid X_{tr}, Y_{tr}\right)} = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

We do not need to calculate the normalisation constant

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) \boxed{p\left(\theta \mid X_{tr}, Y_{tr}\right)} d\theta \approx p\left(y \mid x, \theta_{MP}\right)$$

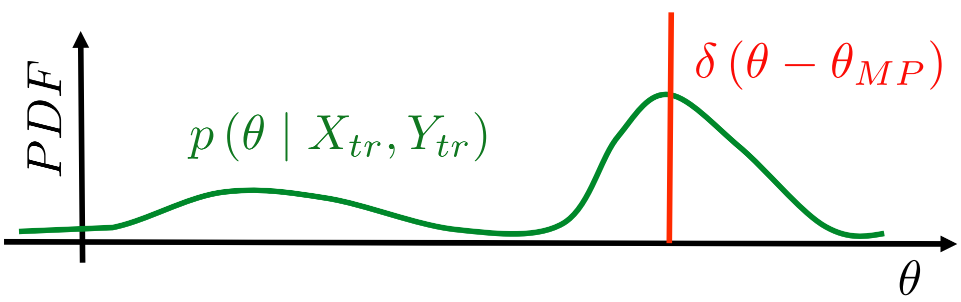
What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$



What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$

$\ast$ Not the same as θ_{ML} — here we use prior

Inference methods: summary

Probabilistic model: $p(x, \theta)$

We want to compute: $p(\theta \mid x)$

Approximation		Inference
Exact	$p(\theta \mid x)$	Full Bayesian inference
More advanced techniques		
Delta function	$p(\theta \mid x) \approx \delta(\theta - \theta_{MP})$	MP inference
No prior	θ_{ML}	MLE