Introduction to Bayesian Inference

Christoph Teichmann Antoine Venant

November 15, 2017

The Bayesian way of life

1. Find a dataset "relevant" to your problem.

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.
3. Build a **joint prior probabilistic model** $p(X, Y)$ describing **both** the data O and the quantity X you want to estimate.

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.
3. Build a **joint prior probabilistic model** $p(X, Y)$ describing **both** the data O and the quantity X you want to estimate.

$\rightarrow$ Obviously, the data and quantity of interrest should **not** be independent and step 3 should reflect that fact!

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.
3. Build a **joint prior probabilistic model** $p(X, Y)$ describing **both** the data O and the quantity X you want to estimate.

→ Obviously, the data and quantity of interrest should **not** be independent and step 3 should reflect that fact!

4. Condition on all observed data o, and get a **posterior** distribution for X: $p(X \mid O = o)$.

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.
3. Build a **joint prior probabilistic model** $p(X, Y)$ describing **both** the data O and the quantity X you want to estimate.

→ Obviously, the data and quantity of interrest should **not** be independent and step 3 should reflect that fact!

4. Condition on all observed data o, and get a **posterior** distribution for X: $p(X \mid O = o)$.
5. (Ideally) Assert how fit your model it, critize it and make a more realistic one.

The Bayesian way of life

1. Find a dataset "relevant" to your problem.
2. Reduce your problem to estimating the distribution of some variable quantity of interest X.
3. Build a **joint prior probabilistic model** $p(X, Y)$ describing **both** the data O and the quantity X you want to estimate.

→ Obviously, the data and quantity of interrest should **not** be independent and step 3 should reflect that fact!

4. Condition on all observed data o, and get a **posterior** distribution for X: $p(X \mid O = o)$.
5. (Ideally) Assert how fit your model it, critize it and make a more realistic one.
6. Use the posterior distribution of Y to solve your problem.

A toy example: Bayesian dating service.

Problem

In order to build an automatic online dating service, we would like to assess a *matching score* between two registered users, as a basis for the system to determine whether it should propose these two users a date with each other. The score should be a real bumber ranging between 0 and 1.

Data

When registering to the dating website, each of the two user answered the same set of n multiple choice questions, and we dispose of these answers.

Our data look like:

Q	ans. User 0	ans. User 1
Q_1	$a)$	$b)$
Q_2	$c)$	$c)$
$\vdots$	$\vdots$	$\vdots$
Q_n	$a)$	$d)$

(*e.g.*: Q_1 : *You prefer a)a good movie, b)a nice book, c)a delicious meal.*)

Simplified data

- To achieve maximal simplicity, we will not use the particular content of questions.
- We will use only the information of whether the two users' answers for a given question are the same or not.
- We thus extract a simplified dataset as follows:

Q	ans. User 0	ans. User 1
Q_1	$a)$	$b)$
Q_2	$c)$	$c)$
$\vdots$	$\vdots$	$\vdots$
Q_n	$a)$	$d)$

$\Rightarrow$

Q	Match
Q_1	no
Q_2	yes
$\vdots$	$\vdots$
Q_n	no

First step: target and observations as random variables

- Assume an underlying probability space $\langle \Omega, p \rangle$.
- Our problem reduces to estimating the distribution of a *matching score* random variable $\Phi : \Omega \mapsto [0, 1]$ after observing the answers made by both users.
- Observed data is represented as a sequence $\langle o_1, \ldots o_n \rangle \in \{0, 1\}^n$
- $o_n = 1$ means that both users choosed the same answer to the n^{th} question of the questionaire, $o_n = 0$ means that they did not.
- Accordingly, assume that the observed data is in fact one outcome of a random vector $O : \Omega \mapsto \{0, 1\}^n$. Equivalently, we can see O as a vector of random variables $(O_1, \ldots, O_n)$ where $(O_i : \Omega \mapsto \{0, 1\})$.

Second step: build the joint Model

- So far we only said "there exists some joint distribution". Did not precise which one.
- Idea: assume that both users answers each question independently of other questions, and that probability of choosing matching answer equals the *matching score*.
- This translates into:

$$p(O = (o_1, \dots o_n) \mid \Phi = \phi) = \phi^{\alpha}(1 - \phi)^{n-\alpha}$$

  where $\alpha = \Sigma_{i=1}^{n} o_i$
- We also need to set a prior probability on Φ. Simply assume it uniform on $[0, 1]$: for $I \subseteq [0, 1]$

$$P(\Phi \in I) = \int_I 1dx$$

  (for instance $P(\Phi \leq 1/2) = \int_o^{1/2} 1dx = [x]_0^{1/2} = 1/2$)

Mixing continuous and discrete models

Not so obvious:

- $\forall \phi \in [0,1]\, P(\Phi = \phi) = \int_{\phi}^{\phi} 1 dx = 0.$
- So shouldn't $P(o \mid \Phi = \phi) = \frac{P(o, \Phi = \phi)}{P(\Phi = \phi)}$ be undefined if $P(\Phi = \phi) = 0$?
- So far, no **full** model: what is for instance $P(O = (o_1, \ldots o_n))$?

Short answer:

- True that 'classic' conditioning does not define $P(o \mid \Phi = \phi)$.
- But $P(o, \Phi = \phi) = P(o \mid \Phi = \phi) P(\phi) = 0$ still true under our definition. So we're not arming the axioms of probability theory.
- define $P(O = (o1, \ldots, o_n), \Phi \in I) = \int_I p(O = (o_1, \ldots, o_n) \mid \Phi = x) dx$. (Exercise: check that this is a probability distribution).

Third step: conditioning on observations $o_1, \ldots, o_n$.

$$
\begin{aligned}
p(\Phi \in I \mid O = (o_1, \ldots, o_n)) &= \frac{p(\Phi \in I, O = (o_1, \ldots, o_n))}{p(O = (o_1, \ldots, o_n))} \\
&= \frac{\int_I p(O = (o_1, \ldots, o_n) \mid \Phi = \phi) d\phi}{p(O = (o_1, \ldots, o_n))} \\
&= \int_I \frac{p(O = (o_1, \ldots, o_n) \mid \Phi = \phi)}{N} d\phi
\end{aligned}
$$

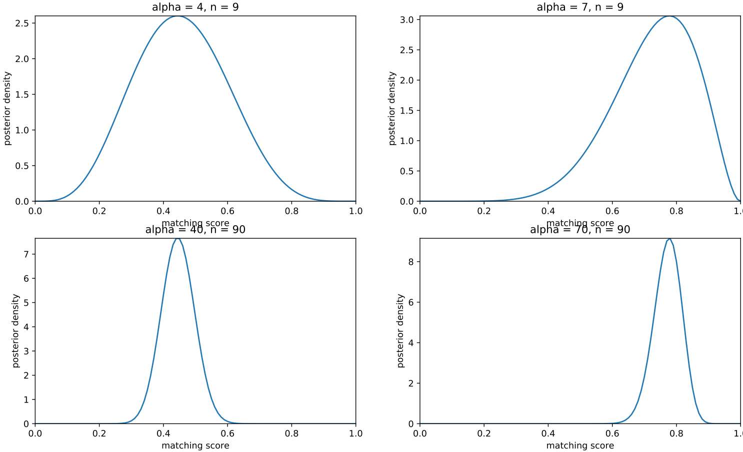
Where $N = p(O = (o_1, \ldots, o_n))$

- We see that the **posterior** distribution of the matching score Φ admits a density $f_{\text{post}} = \frac{p(O=(o_1,\ldots,o_n)|\Phi=\phi)}{N} = \frac{\phi^\alpha(1-\phi)^{n-\alpha}}{N}$

Predictions

- Need to determine normalizing factor
$N = p(O = (o_1, \ldots, o_n)) = \int_0^1 p(O = (o_1, \ldots, o_n) \mid \Phi = \phi) d\phi$
- Hard to do in general! We'll use different techniques to avoid this computation in the seminar.
- But easy, in the present case. Closed form solution: $N = \frac{1}{(n+1)C_n^{\alpha}}$ (Exercise: proove it!).
- Posterior density $f_{\text{post}}(\phi) = \frac{\phi^{\alpha}(1-\phi)^{n-\alpha}}{N}$ is an instance of the **Beta distribution** (obtained with pair of parameters $(\alpha + 1, n - \alpha + 1)$). We'll encounter this again in the future!
- (Posterior-) Expected value for ϕ: $\frac{\alpha-1}{n-1}$.

The more data...



Posterior distributions:

$\alpha = 4, n = 9$

I	$P(\Phi \in I)$
$[0, 0.1]$	0.0016
$[0.1, 0.2]$	0.0312
$[0.2, 0.3]$	0.1175
$[0.3, 0.4]$	0.2166
$[0.4, 0.5]$	0.2562
$[0.5, 0.6]$	0.2107
$[0.6, 0.7]$	0.1189
$[0.7, 0.8]$	0.0410
$[0.8, 0.9]$	0.0062
$[0.9, 1.0]$	0.0001

$\alpha = 40, n = 90$

I	$P(\Phi \in I)$
$[0, 0.1]$	0.0000
$[0.1, 0.2]$	0.0000
$[0.2, 0.3]$	0.0017
$[0.3, 0.4]$	0.1880
$[0.4, 0.5]$	0.6630
$[0.5, 0.6]$	0.1458
$[0.6, 0.7]$	0.0014
$[0.7, 0.8]$	0.0000
$[0.8, 0.9]$	0.0000
$[0.9, 1.0]$	0.0000