Ex2 Q1 solution

29 October 2017 09:21

Solution notes for Example Sheet 2 question 1

In the first four questions you will investigate racial bias in police stop-and-search behaviour. You will make inferences, and quantify your uncertainty about those inferences. The dataset is `https://teachingfiles.blob.core.windows.net/founds/stop-and-search.csv`, and we will restrict attention to records with police_force='cambridgeshire'. We will work with the model

$$\mathbb{P}(Y_i = \mathsf{find}) = \theta_{e_i}$$

where $Y_i \in \{\mathsf{find}, \mathsf{nothing}\}$ is the outcome listed for row i, e_i is the ethnicity, and

$$\theta = (\theta_{\mathsf{Asian}}, \theta_{\mathsf{Black}}, \theta_{\mathsf{Mixed}}, \theta_{\mathsf{Other}}, \theta_{\mathsf{White}})$$

is an unknown parameter.

Question 1 (Bayesian confidence interval).

(a) Let θ consist of 5 independent random variables drawn from the Beta(δ,δ) distribution, where $\delta = 0.5$. Calculate the posterior distribution of θ. Implement a function posterior_sample(size) that generates size independent samples of θ drawn from the posterior distribution. Each sample should be a vector of length 5.

Question 1 (Bayesian confidence interval).

(a) Let θ consist of 5 independent random variables drawn from the Beta(δ,δ) distribution, where $\delta = 0.5$. Calculate the posterior distribution of θ. Implement a function posterior_sample(size) that generates size independent samples of θ drawn from the posterior distribution. Each sample should be a vector of length 5.

Bayesian calculations: lectures 7+8, notes §3.2

3.2.1. BAYESIANISM

Data science is the process by which we change our beliefs about the world, in the light of data. There's no such thing as objective truth, there's only subjective degree of belief. One should represent belief by using a probability distribution, and one should update it using Bayes' rule.

1. Write down a distribution for prior belief

2. Use Bayes' rule to calculate the distribution for posterior belief.

What is the prior belief for this question?

The prior distribution:

"θ consists of 5 independent random variables drawn from the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution".

We'll need to apply Bayes' rule,

For two discrete random variables X and Y

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(Y = y|X = x)\,\mathbb{P}(X = x)}{\mathbb{P}(Y = y)}$$

For continuous X and discrete Y

$$\Pr(X = x|Y = y) = \frac{\mathbb{P}(Y = y|X = x)\,\Pr(X = x)}{\mathbb{P}(Y = y)}$$

[slides from lecture 8]

Applying Bayes' rule in this case, $\underbrace{\Pr(\Theta = \theta \mid \text{data})}_{\text{posterior density}} = \mathbb{P}(\text{data} \mid \Theta = \theta)\, \underbrace{\Pr(\theta)}_{\text{prior density}} \,/\, \text{const.}$

So, we first need to figure out the prior density. We're told "θ consists of 5 independent random variables" — what does that mean?

1.6. Independence and joint distributions

The concept of independent random variables is fundamental in modeling. Informally it means "knowing the value of one of them gives no information about the other." We've used the word several times so far, but we haven't defined it.

Definition. Two random variables X and Y are *independent* if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B) \quad \text{for all } A \text{ and } B.$$

For discrete random variables it's sufficient to check

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y) \quad \text{for all } x \text{ and } y,$$

and for continuous random variables with joint density function $f_{X,Y}(x, y)$, it's sufficient to check

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{for all } x \text{ and } y.$$

Notes Section 1.6

Also the topic of lecture 5

Aside on independence

How do we calculate $\mathbb{E}(PQ)$?

If P and Q are independent, $\mathbb{E}(PQ) = (\mathbb{E}P)(\mathbb{E}Q)$.

Loosely speaking, X and Y are independent if knowing the value of one of them is uninformative about the other.

Precisely speaking, X and Y are independent if $\Pr(X=x|Y=y)=\Pr(X=x)$ for all values of x and y. (and for continuous random variables, we need to write this slightly differently).

Independence is hugely important in modelling.
You can read the exact definition of independence in Section 1.6 of the handout, and you should work through the examples there.

In practical data modelling, we often just assert that values are independent. In practical data science, we test for independence. See Example Sheet 1 question 2 for an example.

Since we are told that θ consists of 5 independent random variables, the density is

$$f(\theta_{Asian}, \theta_{Black}, \cdots) = f_{Asian}(\theta_{Asian}) \times f_{Black}(\theta_{Black}) \times \cdots$$

or, in more useful notation,

$$\Pr(\Theta_{Asian} = \theta_{Asian}, \Theta_{Black} = \theta_{Black}, \cdots) = \Pr(\Theta_{Asian} = \theta_{Asian}) \times \Pr(\Theta_{Black} = \theta_{Black}) \times \cdots$$

And what is the actual density for one of them?

We're told "drawn from the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution". Look it up:

https://en.wikipedia.org/wiki/Beta_distribution

Notation	Beta(α, β)
Parameters	α > 0 shape (real) β > 0 shape (real)
Support	$x \in [0, 1]$ or $x \in (0, 1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$ where $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Translation:

If $X \sim$ Beta(α, β) then it has density

$\Pr(X = x) = x^{\alpha-1} (1-x)^{\beta-1} / \text{const.}$

So, the prior distribution has density function

So, the prior distribution has density function

$$\Pr\left(\Theta_{Asian} = \theta_{Asian},\ \Theta_{Black} = \theta_{Black}, \dots\right) = \theta_{Asian}^{-1/2}(1-\theta_{Asian})^{-1/2}\ \theta_{Black}^{-\frac{1}{2}}(1-\theta_{Black})^{-1/2} \cdots.$$

Question 1 (Bayesian confidence interval).

(a) Let θ consist of 5 independent random variables drawn from the Beta(δ,δ) distribution, where $\delta = 0.5$. Calculate the posterior distribution of θ. Implement a function posterior_sample(size) that generates size independent samples of θ drawn from the posterior distribution. Each sample should be a vector of length 5.

Bayesianism says: start with a prior distribution, and update it with Bayes' rule to get the posterior distribution. Remember Bayes' rule:

For two discrete random variables X and Y

$$\mathbb{P}(X=x|Y=y) = \frac{\mathbb{P}(Y=y|X=x)\,\mathbb{P}(X=x)}{\mathbb{P}(Y=y)}$$

For continuous X and discrete Y

$$\Pr(X=x|Y=y) = \frac{\mathbb{P}(Y=y|X=x)\,\Pr(X=x)}{\mathbb{P}(Y=y)}$$

[slides from Lecture 8]

Applying Bayes' rule in our case:

$$\Pr(\Theta=\theta \mid \text{data}) = \underbrace{\mathbb{P}(\text{data} \mid \Theta=\theta)}_{\text{What is the data?}}\ \underbrace{\Pr(\Theta=\theta)}_{\text{prior density}} \Big/ \text{const.}$$

In the first four questions you will investigate racial bias in police stop-and-search behaviour. You will make inferences, and quantify your uncertainty about those inferences. The dataset is `https://teachingfiles.blob.core.windows.net/founds/stop-and-search.csv`, and we will restrict attention to records with police_force='cambridgeshire'. We will work with the model

$$\mathbb{P}(Y_i = \text{find}) = \theta_{e_i}$$

where $Y_i \in \{\text{find}, \text{nothing}\}$ is the outcome listed for row i, e_i is the ethnicity, and

$$\theta = (\theta_{\text{Asian}}, \theta_{\text{Black}}, \theta_{\text{Mixed}}, \theta_{\text{Other}}, \theta_{\text{White}})$$

The question tells us the data is a list of $Y_i \in \{\text{find, nothing}\}$

$$\mathbb{P}(Y_i = y) = \begin{cases} \theta_{e_i} & \text{if } y = \text{find} \\ 1-\theta_{e_i} & \text{if } y = \text{nothing} \end{cases}$$

It doesn't actually tell us that the Y_i are independent, given θ. Let's just assume they are.

Assuming the Y_i are independent, given θ, the posterior density is thus

$$\Pr\left(\Theta_{Asian} = \theta_{Asian},\ \Theta_{Black} = \theta_{Black}, \dots\right) = \mathbb{P}(\text{data} \mid \Theta=\theta)\ \Pr(\Theta=\theta) \times \text{const}$$

$$= \left(\prod_i \begin{cases} \theta_{e_i} & \text{if } y_i = \text{find} \\ 1-\theta_{e_i} & \text{if } y_i = \text{nothing} \end{cases}\right) \times \theta_{Asian}^{-1/2}(1-\theta_{Asian})^{-1/2} \times \theta_{Black}^{-1/2}(1-\theta_{Black})^{-1/2} \times \cdots$$

$$\left[\theta^{n_{Asian,find}} (1-\theta)^{n_{Asian,nothing}}\ \theta^{-\frac{1}{2}} (1-\theta)^{-1/2}\right]$$

$$= \left[\theta_{Asian}^{n_{Asian,find}} (1-\theta_{Asian})^{n_{Asian,nothing}} \; \theta_{Asian}^{-\frac{1}{2}} (1-\theta_{Asian})^{-1/2} \right]$$

$$\times \left[\theta_{Black}^{n_{Black,find}} (1-\theta_{Black})^{n_{Black,nothing}} \; \theta_{Black}^{-1/2} (1-\theta_{Black})^{-\frac{1}{2}} \right] \times \cdots$$

$$= \left[\theta_{Asian}^{n_{Asian,find}-\frac{1}{2}} (1-\theta_{Asian})^{n_{Asian,nothing}-\frac{1}{2}} \right] \times \left[\theta_{Black}^{n_{Black,find}-\frac{1}{2}} (1-\theta_{Black})^{n_{Black,nothing}-\frac{1}{2}} \right] \times \cdots$$

which is the product of five terms, each of them the density function of a Beta distribution. Thus, the posterior distribution $(\Theta \mid data)$ consists of 5 independent Beta$(n_{e,find} + \frac{1}{2}, n_{e,nothing} + \frac{1}{2})$ distributions.

Question 1 (Bayesian confidence interval).

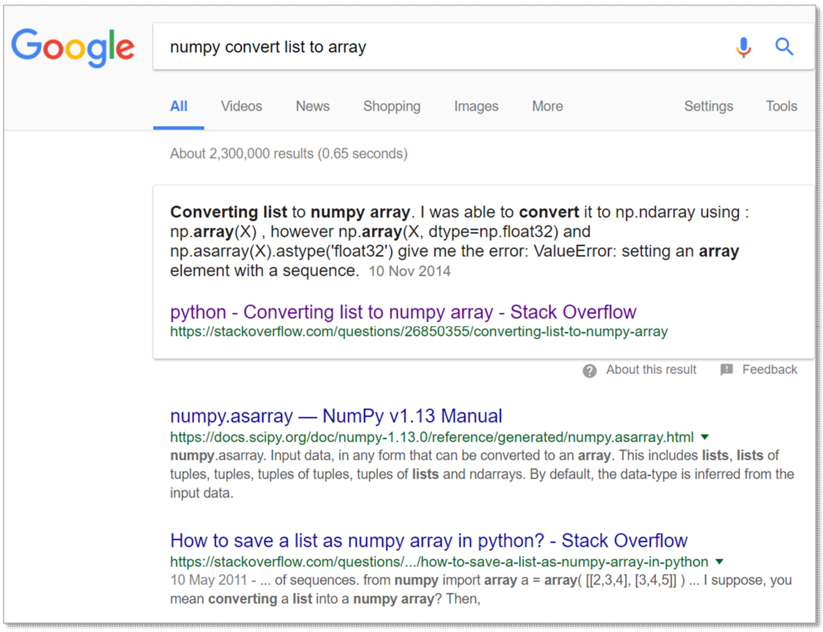
(a) Let θ consist of 5 independent random variables drawn from the Beta(δ, δ) distribution, where δ = 0.5. Calculate the posterior distribution of θ. Implement a function posterior_sample(size) that generates size independent samples of θ drawn from the posterior distribution. Each sample should be a vector of length 5.

```
def single_posterior_sample():
    return [np.random.beta(...., ....) for e in range(5)]
```

the coefficients depend on e

```
def posterior_sample(size):
    return [single_posterior_sample() for i in range(size)]
```

Or, I could return the samples as vectors rather than lists, and arrays rather than lists of lists. That way, it's easier to extract columns and apply maths to each element.


Google
numpy convert list to array
All
Videos
News
Shopping
Images
More
Settings
Tools
About 2,300,000 results (0.65 seconds)
Converting list to numpy array. I was able to convert it to np.ndarray using : np.array(X) , however np.array(X, dtype=np.float32) and np.asarray(X).astype('float32') give me the error: ValueError: setting an array element with a sequence. 10 Nov 2014
python - Converting list to numpy array - Stack Overflow
https://stackoverflow.com/questions/26850355/converting-list-to-numpy-array
About this result
Feedback
numpy.asarray — NumPy v1.13 Manual
https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.asarray.html
numpy.asarray. Input data, in any form that can be converted to an array. This includes lists, lists of tuples, tuples, tuples of tuples, tuples of lists and ndarrays. By default, the data-type is inferred from the input data.
How to save a list as numpy array in python? - Stack Overflow
https://stackoverflow.com/questions/.../how-to-save-a-list-as-numpy-array-in-python
10 May 2011 - ... of sequences. from numpy import array a = array([[2,3,4], [3,4,5]]) ... I suppose, you mean converting a list into a numpy array? Then,