Dr. Peter Arndt

Master for AI and Data Science
Universität Düsseldorf

Math for AI
Winter term 2023/24

Exercise sheet 10

Exercise 37 (10 points)

A coin shows heads with probability p and tails with probability $(1-p)$. The random variable X counts how many times the coin is thrown until heads is shown for the first time. Thus X takes values in the natural numbers with distribution $P(X=k) = p(1-p)^{k-1}$ (a so-called geometric distribution).

In a series of n experiments, the numbers of throws until the first appearance of tails are: $k_1 + 1, \dots, k_n + 1$ (i.e. $k_1, \dots, k_n$ are the numbers of tails before the first head).

Let $a, b > 0$. Compute the a posteriori distribution for p on $(0,1)$, where the prior distribution on $(0,1)$ has the Beta(a,b)-distribution, i.e. density function

$$h(p) := \begin{cases} \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1} & \text{if } 0 < p < 1 \\ 0 & \text{otherwise} \end{cases}$$

Here $B(a,b) := \int_0^1 u^{a-1}(1-u)^{b-1} du$ is the normalizing constant ensuring that the above is really a density function on $(0,1)$. Show that the a posteriori distribution is again a Beta(a',b')-distribution. What are the new parameters a', b'?

Exercise 40 (6 points)

In a game a coin is thrown repeatedly until it shows heads for the first time. Let the random variable X count the total number of coin throws in the game. If the probability of showing heads in a single throw is p, then X has the geometric distribution:

$$P(X = n) = p(1-p)^{n-1}$$

You think that the heads side shows less than the tails side. More precisely, your beliefs about the bias of the coin are expressed by the density function $p \mapsto 3(p-1)^2$.

Now you play the game three times and get 2, 5 and 1 coin throws. Compute the maximum a posteriori estimate for p using this data.

Exercise 41 (14 points)

The Beta distribution with parameters α, β is the distribution on the unit interval $(0, 1)$ with density function $B(\alpha, \beta)$, given by

$$B(\alpha, \beta)(x) = C \cdot x^{\alpha-1}(1-x)^{\beta-1}.$$

The binomial distribution for n trials with parameter p is a discrete distribution on $\{0, \ldots, n\}$, giving probabilities for the number of successes in n experiments with success probability p. Its probability mass function is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The family of Beta distributions (parametrized by α, β) forms a conjugate family for the family of binomial distibutions (parametrized by p).

Suppose you plant 3 batches of 5 seeds each, and count how many of them sprout. From the first batch all 5 seeds sprout. From the second batch 3 seeds sprout. From the third batch 4 seeds sprout.

(a) (7 points) If your prior distribution was $B(\alpha, \beta)$ for some fixed numbers α, β, which member of the Beta family is your posterior distribution?

(b) (7 points) Compute the maximum a posteriori estimate for p as a function of α, β.

Exercise 42 (10 points)

a) (5 points) This exercise is about horseshoe crabs.

When mating, female horseshoe crabs walk to the shore accompanied by one male riding on their spine and often surrounded by a number of further males, called their satellites. The aim of this exercise is to predict the number of satellites based on the weight and colour of a female. Since we are trying to predict an integer, an appropriate model could be Poisson regression – see Example 5.10.5 of the notes.

In the usual folder you find the file `Crabs.dat` and a Jupyter notebook where it is loaded into a data frame.

The first column just provides a numbering of the observed female crabs – you can disregad it. The column `'y'` contains the number of satellites. Horseshoe crabs come in four colours, which are named `1,2,3,4` in the column `'colour'` – this is an example of categorical data, as opposed to numerical data. Just like in linear regression, if the non-categorical data falls into n categories, you can use such data also in generalized linear models by introducing $n-1$ new variables.

The provided notebook tells you how to compute a model for Poisson regression – do that.

b) (5 points) This exercise is about guessing from where an Australian possum is. Here you should do logistic regression.

Deadline: Friday 22nd of December, 10:00.
Upload your solution to this link.