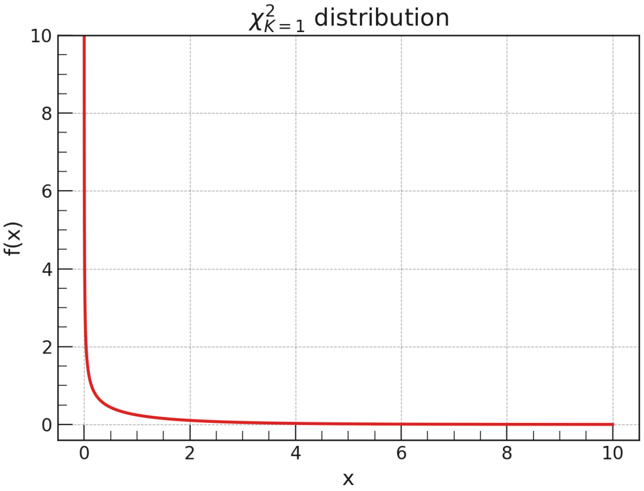
Advanced Methods in Applied Statistics
Problem set 3

Philip Kofoed-Djursner
(Tkv976)

17th of March 2023



1 Problem 2

If Wilks' theorem is satisfied the 2 times log-likelihood difference ($2\Delta LLH$) will follow a χ^2 distribution with degrees of freedom equal to the difference in parameter dimensionality. This can both be used as a test statistic and as a way of constructing confidence intervals. The problem is to find the value of $2\Delta LLH$ to construct the 77.9% confidence interval. The difference in parameter dimensionality is 4 as the maximum log-likelihood only exists for one set of g, u, q, and b. Therefore, following Wilks theorem, the $2\Delta LLH$ should follow a χ^2 distribution with 4 degrees of freedom. I found this by checking which value satisfied the equation:

$$CDF(\chi^2_{k=4}(x)) = 0.779 \tag{1}$$

where CDF is the cumulative distribution function and k is the degrees of freedom. I found that the value of $2\Delta LLH$ which defines the 77.9% confidence interval is

$$2\Delta LLH = 5.72$$

2 Exercise 3

2.1 Problem 3a

An estimation of the probability density can be done by doing a kernel density estimation (KDE). At each data point some density distribution is placed around it, the densities from each point is then added together and the kernel is normalized to 1. In this problem, a Gaussian distribution with a $\sigma = 25\,cm$ will be used as the kernel. The data over the white shark lengths and the kde is shown in Figure 1.

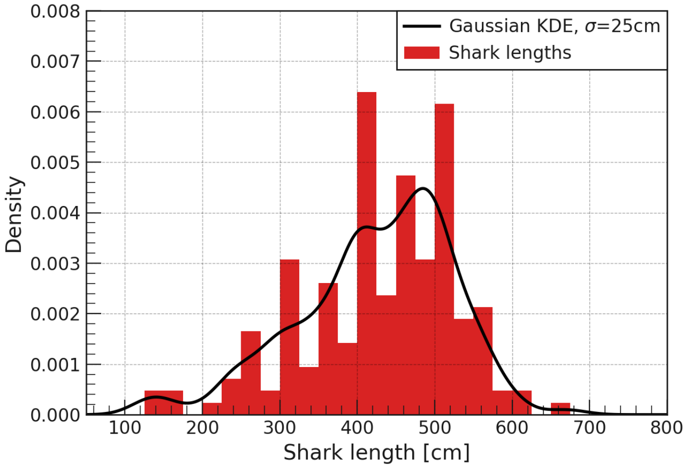


Figure 1: Histogram of white shark length and a gaussian KDE with $\sigma = 25\,cm$ overlayed.

From the KDE, it can be calculated the probability of finding a shark longer than 653 cm, just by integrating the KDE from 653 to infinity.

$$P_{KDE}(L > 653cm) = \int_{653}^{\infty} P_{KDE}(L)\mathrm{d}L = 0.0047 \tag{2}$$

2.2 Problem 3b

We were provided with two probability density functions that describe the probability of a male or female of a given length having a given weight. we are tasked with finding the probability, given these PDFs and the KDE form section 2.1, that a shark that weighs 763 kg is longer than 337 cm. This problem can be solved by utilizing Bayes' rule.

$$P(L > 337|W = 763) = \frac{P(W = 763|L > 337)P(L > 337)}{P(W = 763)} = \tag{3}$$

$$\frac{(P(W = 763|L > 337, M)P(M) + P(W = 763|L > 337, F)P(F))\, P(L > 337)}{\int_{-\infty}^{\infty} \mathrm{d}L\, P(W = 763|L)P(L)} = \tag{4}$$

$$\frac{(P(W = 763|L > 337, M)P(M) + P(W = 763|L > 337, F)P(F))\, P(L > 337)}{\int_{-\infty}^{\infty} \mathrm{d}L\; (P(W = 763|L, M)P(M) + P(W = 763|L, F)P(F))\, P(L)} = \tag{5}$$

$$\frac{\int_{337}^{\infty} \mathrm{d}L\, (P(W = 763|L, M)P(M) + P(W = 763|L, F)P(F))\, P(L)}{\int_{-\infty}^{\infty} \mathrm{d}L\; (P(W = 763|L, M)P(M) + P(W = 763|L, F)P(F))\, P(L)} \tag{6}$$

Here P(W = 763—L,M/F) is the PDF for male/female sharks' weight, P(M) and P(F) is the probability of finding each sex, which is given to be 50:50. The Prior, P(L), is take to be the KDE form Section 2.1. The marginal likelihood is expanded in all possibilities to normalize the posterior probability. In Figure 2, the likelihood, prior, and posterior are plotted as functions of the length. This is where the integral in the

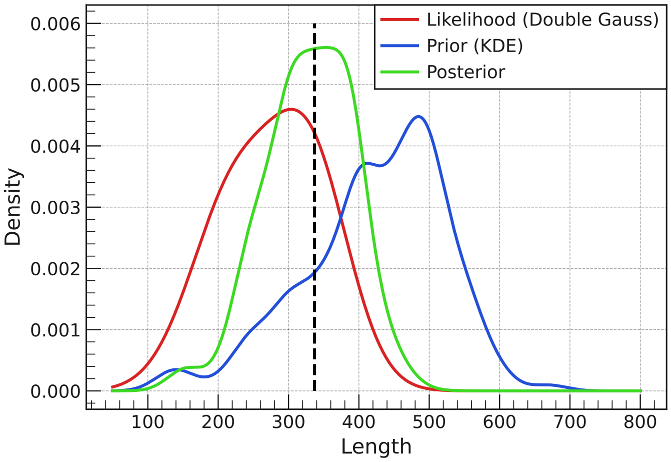


Figure 2: The likelihood, prior, and posterior are plotted as functions of the length. Black line is from where I integrate to find the solution to the problem

numerator is not calculated. It is very nice that one can see how the prior pulls the likelihood towards higher lengths. By evaluating the integral in Eq. 6 I get the probability of the shark being longer than 337 cm.

$$P(L > 337|W = 763) = 0.48 \tag{7}$$

2.3 Problem 3c

Now I am only interested in mature sharks, which are sharks of length longer than 201 cm. Therefore, the likelihood and prior must be truncated to go from L>201cm to infinity, and then renormalized to still have an area equal to 1. As the mature shark population still has a 50:50 split in the sexes the two PDFs for each sex must b renormalized separately otherwise, we cannot make sure this condition is fulfilled. The equations are the same as in Eq. 3-6, but with truncated PDFs and KDE. In Figure 3, is shown the truncated PDFs and KDE as functions of L. If integrating again from 337 to infinity I get a probability of a mature shark that weighs 763 kg being longer than 337cm at $P = 0.45$

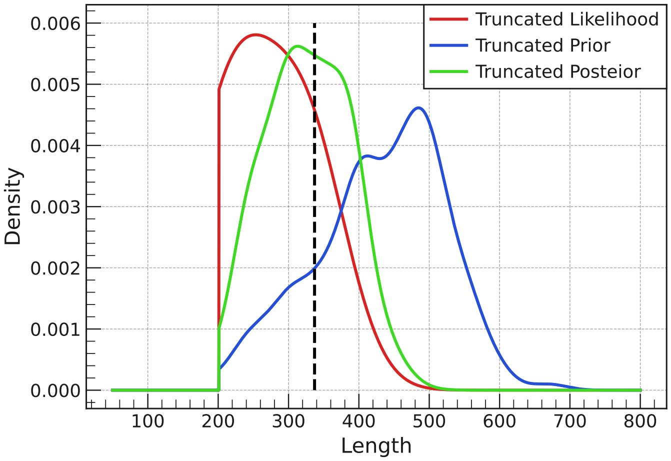


Figure 3: The truncated likelihood, prior, and posterior are plotted as functions of the length. Black line is from where I integrate to find the solution to the problem

3 Exercise 4

3.1 Problem 4a

I choose to use a gradient boosted decision tree (BDT) to predict given the data if the user would lead to revenue or not. In Figure 4, you see the scores assigned by the BDT. It is good at finding a lot of users which for sure will not generate revenue but it is very unsure and bad at finding good separation for most cases. Also, the data has a lot more entries where the user does not generate revenue which screws the result. If I place the separation value at -1.8 it is equally good at predicting no revenue as revenue and the TPR and TNR are 84.5%

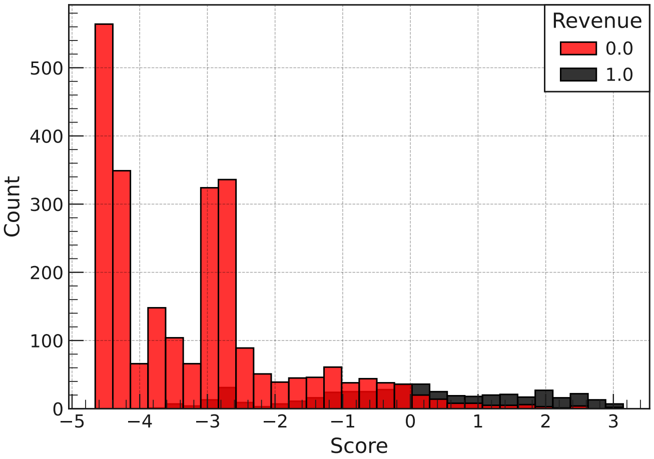


Figure 4: All the scores assigned by the BDT

3.2 Problem 4b

I ran a few tests of different parameters for the BDT and continuously followed both the success rate on the training and test set. I made sure not to increase the complexity of the model if it lead to a larger separation in success rate between the training and test dataset. Thereby, creating the simplest model which still got equally good results.

3.3 Problem 4c

Uploaded to Absalon