Assignment 2 – Fish Classifier

Elif Erden
Student ID: 041503038
Date: 27.02.2019

COMP 462
Introduction to Machine Learning

1. Algorithm Design and Explanation

In this assignment, a fish classification algorithm was developed and tested for the fish data. The dataset consists of 300 samples, from Salmon (Class 1) and Sea Bass (Class 2) fish classes. Each fish class is represented by a single measurement which is the length of a fish.

In order to create the classifier, firstly the training data was analyzed and plotted to find a decision rule for classifying the fishes. For that reason, the data was separated into two groups as fish types. Histograms for both salmon and seabass fish lengths were plotted as shown in Figure 1 and Figure 2. Then, histogram of salmon and seabass fish lengths in one plot was plotted (Figure 3). As it can be seen in histograms, both salmon and seabass fish lengths can be considered to be distributed normally.

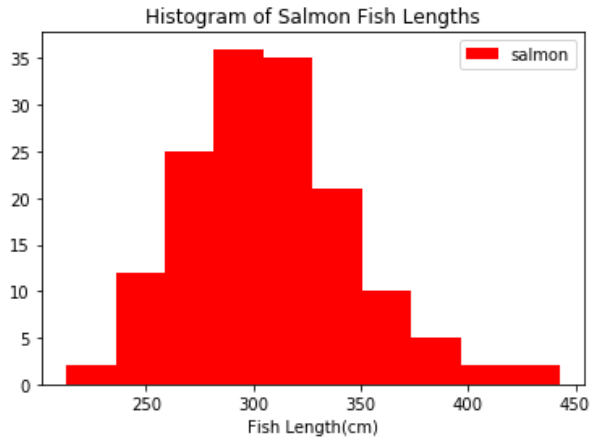

Figure 1. Histogram of salmon fish lengths.

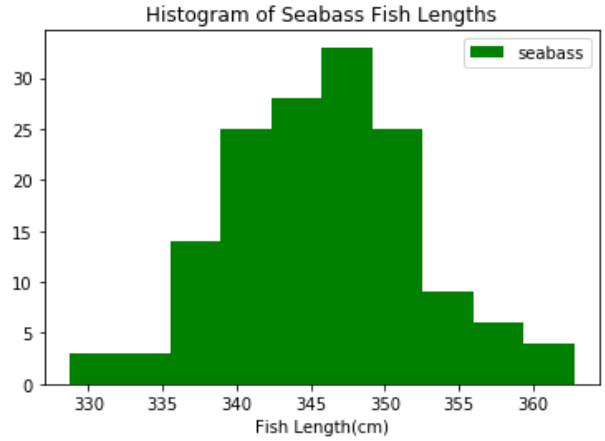

Figure 2. Histogram of seabass fish lengths.

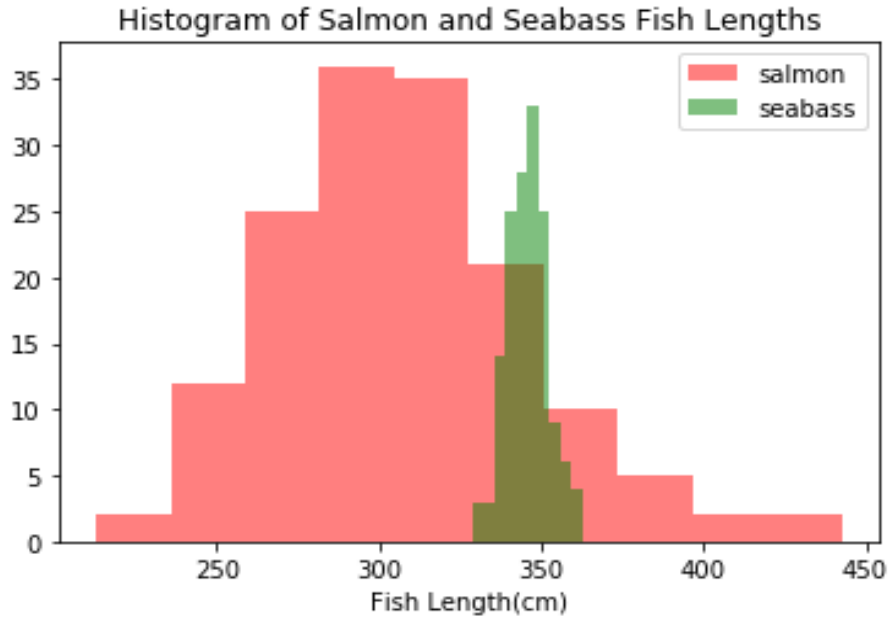

Figure 3. Histogram of salmon and seabass fish lengths.

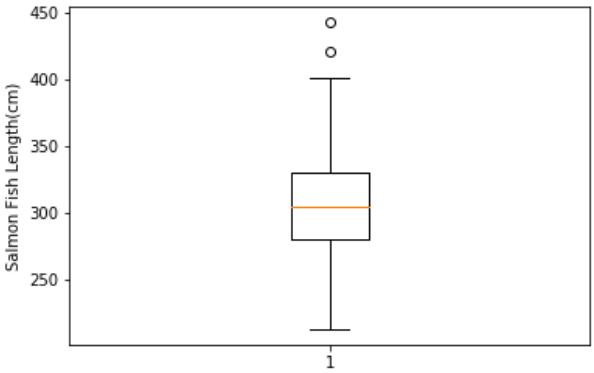

Figure 4. Boxplot of salmon fish lengths.

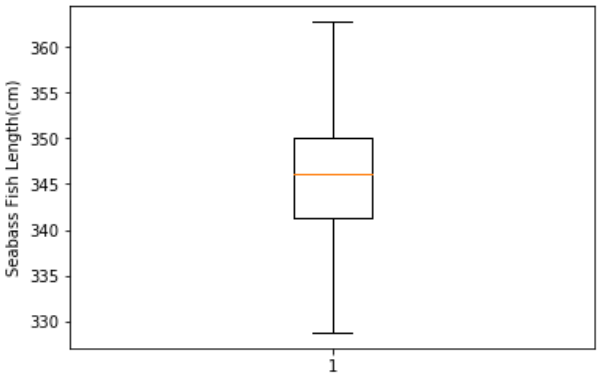

Figure 5. Boxplot of seabass fish lengths.

Boxplots for salmon and seabass fish lengths were plotted separately to support the normality decision. Since there are few exceptionally large or small values and boxplots for both fish lengths look symmetric, it can be said that two fish lengths are distributed normally. After plotting the data and as a result of normality decision, mean and standard deviation for both class were calculated as shown in Table 1.

Table 1. Descriptive statistics for fish lengths.

	Salmon	Seabass
Mean	307.45	345.79
Standard deviation	38.91	6.32

With descriptive statistics in Table 1, the classifier was implemented as a function. In predict function there is one input parameter which is a dataset to be classified. For every length in the dataset, firstly probabilities of being salmon and seabass for the related length are calculated according to normal distribution. If the probability of being salmon is greater than the probability of being seabass, the classifier labels this fish as salmon, otherwise it labels this fish as seabass.

2. Results

The predict classification algorithm was tested on the training data and the correct classification accuracy was calculated as 90.67% with the total error count of 28 among 300. For understanding the details of the result, confusion matrix is provided in Table 2. As shown in the confusion matrix, among 150 test samples from Class 1, 127 of them were classified as Class 1, and 23 of them were classified as Class 2. Similarly, for the Class 2, out of 150 samples, 5 of them are classified as Class 1 and 145 of them are correctly classified as Class 2.

Table 2. Confusion matrix for the fish classification problem.

	Predicted Class = 1	Predicted Class = 2
True Class = 1	127	23
True Class = 2	5	145