Naive Bayes Classifier

Team 30
Dekel Viner(s2612925)   Andreea Glavan (s3083691)

September 2018

Assignment 1

1. The code for this plot, as well as the rest of the exercise, is provided in the appendix. It is notable that, in the plot, the two categories mirror each other, adding up to a probability of 1 for any index.

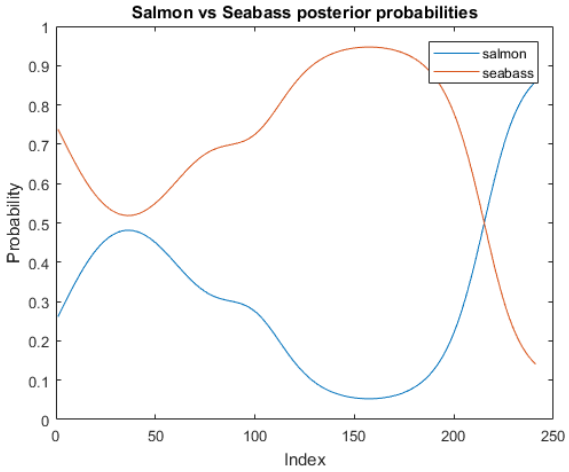

Figure 1: Posterior probabilities for the 2 fish types based on the Bayes formula

2. For a length value of 8, the probability of the fish being a salmon is 0.3403, whereas the probability of it being a seabass is 0.6597. This was calculated in the last lines of the code, and is also visible in the plot due to the seabass plot being greater than the salmon plot at index 71(which has value 8).

   We applied the same method for the value 20, and the resulted probabilities were 0.1342 for salmon and 0.8658 for seabass.

Assignment 2

1. The plot below illustrates the data sets, shown according to the exercise text. The code for this plot, as well as the rest of this assignment, is available in the appendix.

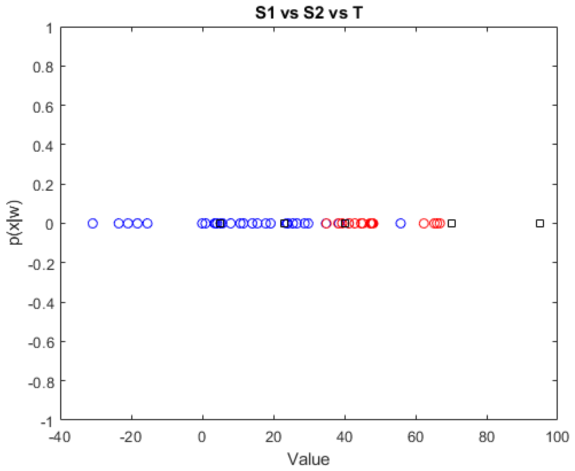

Figure 2: X axis plot of the 3 data sets S1, S2, and T. S1 is plotted as blue circles, S2 as red circles, and T as black squares

2. The plot below was created using the Gaussian function $\frac{1}{2\pi\sigma}exp(\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2})$. μ and σ were calculated using the mean and square root of the variance using maximum estimation, respectively. Using the functions leads to a smoother appearance which illustrates the probabilities of each value in the data set, as opposed to the rugged data which is more difficult to read. What is more, the second plot better illustrates the densities of the data sets. Of course, the two plots(figure 2 and figure 3) do reflect each other value wise.

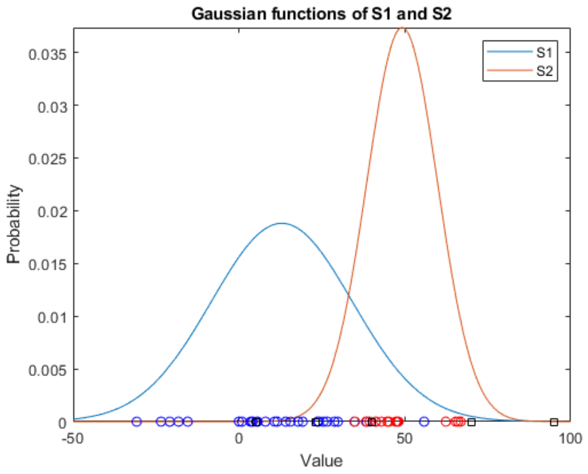

Figure 3: Data sets S1 and S2 plotted using Gaussian functions, in blue and red respectively. Below, the populations is illustrated as circles/squares according to the same legend as used previously

3. We calculated the prior probabilities using $P(w_1) = \frac{|S_1|}{|S_1|+|S_2|}$, and as a result we had the prior probability of dataset 1 equal to 0.6663 and of dataset 2 equal to 0.3337.

4. By multiplying the previous Gaussian function with the new prior probability, we obtained the plot below:

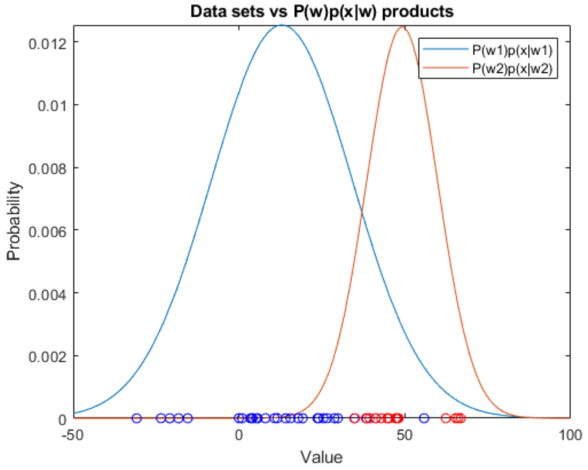

Figure 4: $P(\omega_1)p(x|\omega_1) and P(\omega_2)p(x|\omega_2)$ plots, as well as the population of the 2 data sets, illustrated through blue and red circles, respectively

5. In order to solve the equation, we started from $P(\omega_1)p(x|\omega_1) = P(\omega_2)p(x|\omega_2)$, which is equivalent to: $\frac{1}{2\pi\sigma_1}exp(\frac{-1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}) = \frac{1}{2\pi\sigma_2}exp(\frac{-1}{2}\frac{(x-\mu_2)^2}{\sigma_2^2})$.

   By substituting our values in the formula above, we reached: $-4.3795 - \frac{(x-12.9024)^2}{2*21.2131^2} = -4.3847 - \frac{(x-49.1715)^2}{2*10.6773^2}$. By applying algebra, this simplifies to $0.0033 * x^2 - 0.46 * x + 10.7943 = 0$. Solving this quadratic equations gave us the roots $x_1 = 37.0607$ and $x_2 = 85.8952$, which we used as the decision criterion.

6. The plot of the classified values in the data set T is illustrated below. The two vertical straight lines plotted represent x_1, x_2, illustrating the interval used for the classification. If a value belongs in the interval, then that implies it belongs to S2, otherwise it belongs to S1.

   Figure 5 includes the populations of S1 and S2, while Figure 6 does not, in order to clearly display the chosen categories for the T values.

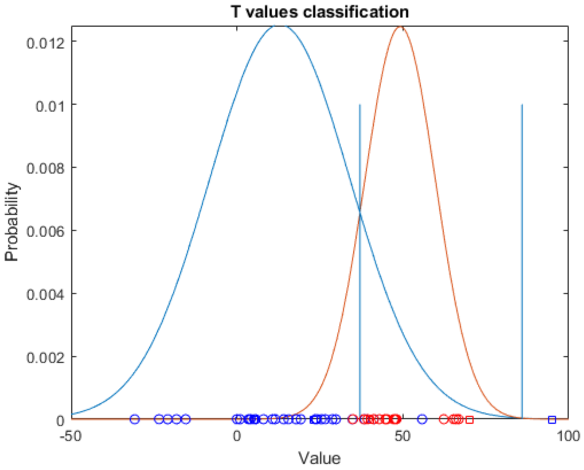

Figure 5: Classification of the values in T (blue squares belonging to S1 and red squares belonging to S2) next to $P(\omega_1)p(x|\omega_1) and P(\omega_2)p(x|\omega_2)$ plots and populations of S1 and S2

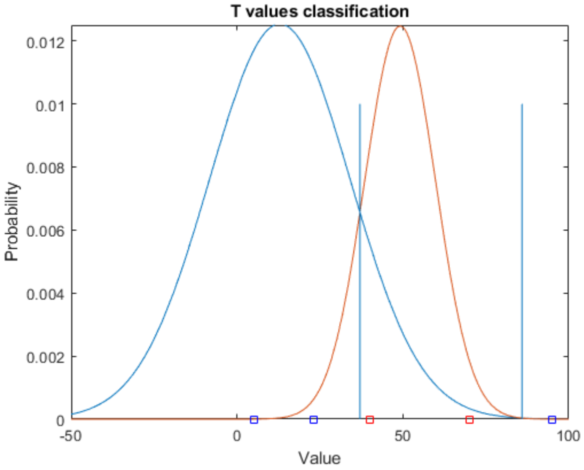

Figure 6: Classification of the values in T (blue squares belonging to S1 and red squares belonging to S2) next to $P(\omega_1)p(x|\omega_1) and P(\omega_2)p(x|\omega_2)$ plots

7. The misclassification rate of each class was calculated using the matlab function erf. While there weren't any misclassifications for dataset S2, there were several for dataset S1. This can be attributed to the slight overlap of the two data sets in between values 25 and 50. Given that this overlap is a part of the classification interval, it is immediate that S1 values would then be incorrectly classified as part of S2. What is more, these values appear to be the only misclassified ones due to,excluding the case mentioned above, there being no overlaps between the two datasets.

Appendix

Listing 1: Source code of assignment 1

```
%posterior probabilities
for i=1:241
    prob_salmon(i)=(p_salmon(i))*0.25/(p_salmon(i)*0.25+p_seabass(i)*0.75);
end

for i=1:241
    prob_seabass(i)=(p_seabass(i))*0.75/(p_salmon(i)*0.25+p_seabass(i)*0.75);
end

%plot
plot(prob_salmon);
hold on;
plot(prob_seabass);
```

```
title('Salmon vs Seabass posterior probabilities');
xlabel('Index');
ylabel('Probability');
legend('salmon', 'seabass');

%point 2, check for 8 and 20
%index of length value 8
i=71
prob_8_salmon = (p_salmon(i))*0.25/(p_salmon(i)*0.25+p_seabass(i)*0.75);
prob_8_seabass=(p_seabass(i))*0.75/(p_salmon(i)*0.25+p_seabass(i)*0.75);

%index of length value 20
i=191
prob_20_salmon = (p_salmon(i))*0.25/(p_salmon(i)*0.25+p_seabass(i)*0.75);
prob_20_seabass=(p_seabass(i))*0.75/(p_salmon(i)*0.25+p_seabass(i)*0.75);
```

Listing 2: Source code of assignment 2

```
load('normdist.mat');

%point 1
%plot datasets
figure;
plot(S1,zeros(1,length(S1)),'bo');
hold on;
plot(S2,zeros(1,length(S2)), 'ro');
hold on;
plot(T,zeros(1,length(T)), 'ks');
title('S1 vs S2 vs T');
xlabel('Value');
ylabel('p(x|w)');

%point 2
%calc mean, sd const
S1_mean = mean(S1);
S1_sd = sqrt(var(S1));
S2_mean = mean(S2);
S2_sd = sqrt(var(S2));

%gaus function plots
figure;
fplot(@(x) ((1/(sqrt(2*pi)*S1_sd))*exp(-1/2*(((x-S1_mean)^2)/(S1_sd^2)))),[-50 1
hold on;
fplot(@(x) ((1/(sqrt(2*pi)*S2_sd))*exp(-1/2*(((x-S2_mean)^2)/(S2_sd^2)))),[-50 1
plot(S1,zeros(1,length(S1)),'bo');
hold on;
plot(S2,zeros(1,length(S2)), 'ro');
```

```
hold on;
plot(T,zeros(1,length(T)), 'ks');
title('Gaussian functions of S1 and S2');
xlabel('Value');
ylabel('Probability');
legend('S1','S2');

%point 3
%prior probability formula
S1_size=size(S1);
S2_size=size(S2);
S1_prior=S1_size/(S1_size + S2_size);
S2_prior=S2_size/(S1_size + S2_size);

%point 4
%plot
figure;
fplot(@(x) S1_prior*(1/(sqrt(2*pi)*S1_sd))*exp(-1/2*(((x-S1_mean)^2)/(S1_sd^2)))
hold on;
fplot(@(x) S2_prior*(1/(sqrt(2*pi)*S2_sd))*exp(-1/2*(((x-S2_mean)^2)/(S2_sd^2)))
hold on;
plot(S1,zeros(1,length(S1)),'bo');
hold on;
plot(S2,zeros(1,length(S2)), 'ro');
title('Data sets vs P(w)p(x|w) products');
xlabel('Value');
ylabel('Probability');
legend('P(w1)p(x|w1)','P(w2)p(x|w2)');

%point 5
%solve P(w1)p(x|w1)=P(w2)p(x|w2)
%eq=0.0033*x^2-0.46*x+10.7943==0;
%sol=solve(eq,x);

x_1=37.0607;
x_2=85.8952;

%point 6
%plot
figure;
fplot(@(x) S1_prior*(1/(sqrt(2*pi)*S1_sd))*exp(-1/2*(((x-S1_mean)^2)/(S1_sd^2)))
hold on;
fplot(@(x) S2_prior*(1/(sqrt(2*pi)*S2_sd))*exp(-1/2*(((x-S2_mean)^2)/(S2_sd^2)))
hold on;
plot(S1,zeros(1,length(S1)),'bo');
line([x_1 x_1],[0 0.01],'LineWidth',0.5)
line([x_2 x_2],[0 0.01],'LineWidth',0.5)
hold on;
plot(S2,zeros(1,length(S2)), 'ro');
hold on;
title('T values classification');
xlabel('Value');
ylabel('Probability');

%classify
for i=1:length(T)
    if(T(i)<x_1 || T(i)>x_2)
        plot(T(i),0, 'bs');
    else
        plot(T(i),0, 'rs');
    end
end

%point 7
S1_er=erf(S1);
S2_er=erf(S2);
```