Normalized Hamming Distance

Dekel Viner(s2612925) Andreea Glavan (s3083691)

September 2018

1 Assignment 1

1. The code for this plot, as well as the rest of the exercise, is available in the appendix.

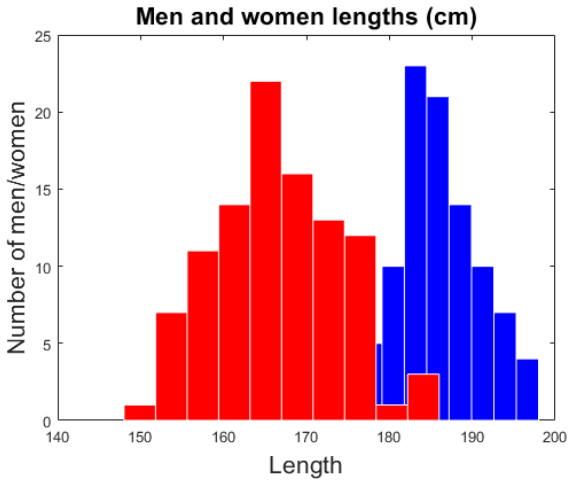

Figure 1: Histogram illustrating men vs women length, showcased in blue and red, respectively

2. With the decision criterion set to 170, there are 0 incorrectly classified men, however there are 35 incorrectly classified women.

3. For the minimal number of mis-classifications, we found 179 to be the decision criterion. With this value, there are 8 men outliers and 4 women outliers. In order to determine this value, we tested the value range (160, 190), as shown in the code which is provided in the appendix.

2 Assignment 2

1. This plot was produced using the code shown in the appendix.

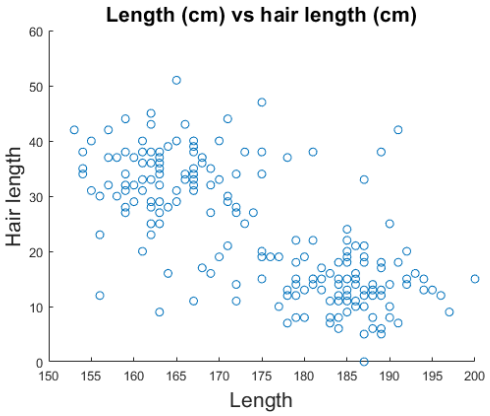

Figure 2: Scatter plot illustrating length vs hair length

2. The figure below illustrates the decision boundary we chose. This decision was made under the assumptions that, generally, men have shorter hair than women and men are taller. However, the curved shape of the line is due to the fact that there are likely to be a few men with long hair and a few women with short hair.

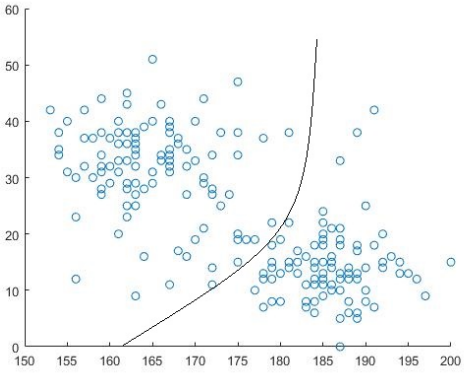

Figure 3: Decision boundary for the previous plot, separating men and women, on the right and left hand sides respectively

3 Assignment 3

1. The iris code difference, for 2 different random rows from different files, is approximately 15.

2. The code for both points a and b are available in the appendix. After running it for both data sets S and D, it was immediate that the normalized hamming distance was greater in the case of two different persons but the same row, as opposed to two different rows of the same person's iriscode.

3. The histogram is illustrated in the figure below. There is little to no overlap between the two, due to the different value ranges. For the value 0.2 there is a clear transition from the first data set to the second. As noted in the previous point, this is because of the difference in the data between two different people iriscodes is significantly higher then the difference between two random rows of the same person. The gaps in the histogram can be attributed to a high frequency of similar values in a data set. This is especially true for the S data set which compares two rows from the same file, the range of different HD values is lower which in turn results in lots of similar values for the normalized hamming distance. Overall from the histogram we can deduce the S data set has higher frequencies and a lower variance.

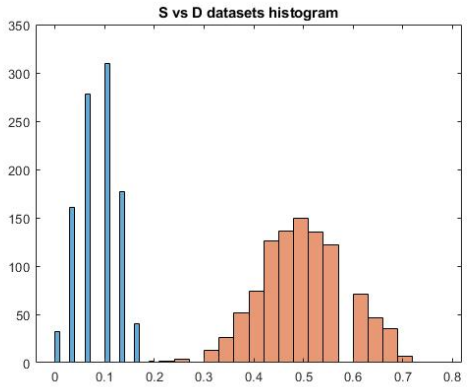

Figure 4: Normalized Hamming distance for S vs D, displayed in blue and orange respectively

4. The normal distributions approximately fit the histograms, however it is notable that, due to the fact that the data sets resulted in small sets of possible values, certain areas in the histogram are empty, as opposed to the normal distributions.

   It is notable that, while the data set D is above the normal distribution by small values, only at points 0.4 and 0.6, data set S peaks(at 0.1) later than

the normal distribution expects it to. What is more, the values following for this data set are also slightly above the predicted normal distribution. This could be due to the random factor and the lower differences between different rows if they belong to the same person's iriscode. If ran multiple times, of course the histogram values would be different, in some cases more like the normal distributions, in others less like it.

Due to the nature of the exercise, for the Gaussian functions used in the normal distribution plot, we used constant values for the peaks. This values were chosen based on the average highest frequency value over multiple runs for each data set. It would have also been possible to use the highest frequency value of the current run, however this estimated values provide equally accurate results.

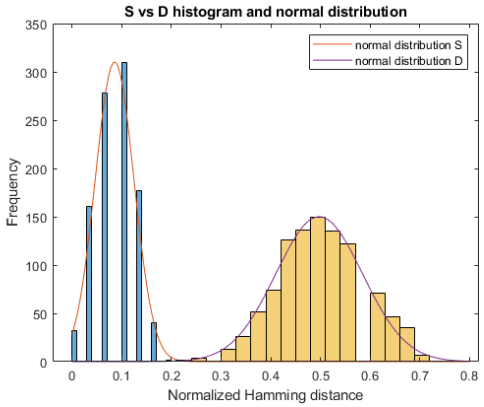

Figure 5: Normalized Hamming distance for S vs D,displayed in blue and yellow respectively, against normal distribution line plots, red and purple respectively

4 Appendix

Listing 1: `Source code of assignment 1`

```
load('lab1_1.mat');

%create histogram
hist(length_men)
hold on
hist(length_women)

%labels
title('Men and women lengths (cm)','fontsize',16);
xlabel('Length','fontsize',16);
```

```
ylabel('Number of men/women','fontsize',16);

%get the handle of the hist bars
h=findobj(gca, 'Type','patch');

%set colors of the 2 sets
set(h(1),'FaceColor','r','EdgeColor','w')
set(h(2),'FaceColor','b','EdgeColor','w')

%set range to search, initialize variable
decision_point = (160:190);
best=400;
best_decision_point=0;
men_cnt=0;
women_cnt=0;
%calculate total outliers for each decision point
for j=1:length(decision_point)
    for i=1:length(length_men)
        if(length_men(i)<decision_point(j))
         men_cnt=men_cnt+1;
        end
        if(length_women(i)>=decision_point(j))
         women_cnt=women_cnt+1;
        end
    end
    total_outliers=women_cnt+men_cnt;
    %set new best outlier val
    if(total_outliers<best)
        best_women_outliers=women_cnt;
        best_men_outliers=men_cnt;
        best=total_outliers;
        best_decision_point=decision_point(j);
    end
    %reset counters
    women_cnt=0;
    men_cnt=0;
    total_outliers=0;
end
```

Listing 2: Source code of assignment 2

```
load('X:\My Desktop\IIS\A2\lab_week2_data\lab1_2.mat')

%create scatter plot
figure
scatter(measurements(:,1),measurements(:,2));
```

```
%labels
title('Length (cm) vs hair length (cm)','fontsize',16);
xlabel('Length','fontsize',16);
ylabel('Hair length','fontsize',16);
```

Listing 3: Source code of assignment 3 point 2a

```
%load('X:\My Desktop\IIS\A2\a3\.mat')
%initialize
file_names = dir('a3');
S=zeros(1000,1);

%repeat process
for trial=1:1000
%pick rand file + rows
i=randi([3 22],1,1);
file=file_names(i).name;
res=load(strcat('a3\',file));
row0=0;
row1=0;
while(row0==row1)
row0=randi([1 20],1,1);
row1=randi([1 20],1,1);
end

%compute hamming dist
HD=sum(abs(res.iriscode(row0,:)-res.iriscode(row1,:)));
%normalize hd
norm_HD=HD/30;
%add to set
S(trial)=norm_HD;
end
```

Listing 4: Source code of assignment 3 point 2b

```
%load('X:\My Desktop\IIS\A2\a3\.mat')
%initialize
file_names = dir('a3');
D=zeros(1000,1);

%repeat process
for trial=1:1000
    i=0;
    j=0;

%get and load rand files + row
while(i==j)
```

```
i=randi([3  22],1,1);
j=randi([3  22],1,1);
end
file0=file_names(i).name;
file1=file_names(j).name;
res0=load(strcat('a3\',file0));
res1=load(strcat('a3\',file1));
row=randi([1  20],1,1);

%compute and normalize hamming diistance
HD=sum(abs(res0.iriscode(row,:)-res1.iriscode(row,:)));

norm_HD=HD/30;

%add to set
D(trial)=norm_HD;
end
```

Listing 5: Source code of assignment 3 point 3

```
figure

histogram(S);
hold on
histogram(D);
title('S vs D datasets histogram', 'textsize',16);
```

Listing 6: Source code of assignment 3 point 4

```
%compute means and variances
mean_S=mean(S);
mean_D=mean(D);
var_S=var(S);
var_D=var(D);

%compute normal dist
x=normpdf(mean_S, sqrt(var_S));
y=normpdf(mean_D, sqrt(var_D));

%plot

figure
histogram(S);
hold on;
p1=fplot(@(x) 310*exp(-((x-mean_S)^2)/(2*var_S)),[0  0.8]);
%plot(S,x);
```

```
histogram(D);
hold on;
%plot(D,y);
p2=fplot(@(x) 150*exp(-((x-mean_D)^2)/(2*var_D)),[0 0.8]);

%labels
title('S vs D histogram and normal distribution');
xlabel('Normalized Hamming distance');
ylabel('Frequency');
legend([p1 p2],{'normal distribution S','normal distribution D'},'Location','nor
```