CS1675 - Assignment 1

Zachary M. Mattis

January 24, 2019

I. Problem 2 - Data Analysis

a. Max / Min

Max	17	199	122	99	846	67.1	2.42	81
Min	0	0	0	0	0	0	0.078	21

b. Mean / Variance

μ	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926	0.4719	33.2409
σ	11	1021	374	254	13264	62	0	138

c. Subset

Class 0

μ	3.298	109.98	68.184	19.664	68.792	30.3042	0.4207	31.19
σ	3.0172	26.1412	18.0631	14.8889	98.8653	7.6899	0.2991	11.6677

Class 1

μ	4.8657	141.2575	70.8246	22.1642	100.3358	35.1425	0.5505	37.0672
σ	3.7412	31.9396	21.4918	17.6797	138.6891	7.263	0.3724	10.9683

Based on the data from the tables above, attribute 2 appears to show the greatest difference in mean among the differing classes. Attribute 2 corresponds to the plasma glucose concentration after 2 hours in an oral glucose tolerance test. This analysis makes sense considering that the two classes are divided by a diabetes diagnoses, which would have a strong correlation with an attribute associated to glucose levels.

d. Histogram

```
function histogram_analysis( attribute_vector )
    hist(attribute_vector, 20);
end
```

e. Normal Distribution

The two attribute values that most closely match a normal distribution are Diastolic Blood Pressure (attr. 3) and Body Mass Index (attr. 6), as shown in the histograms below.

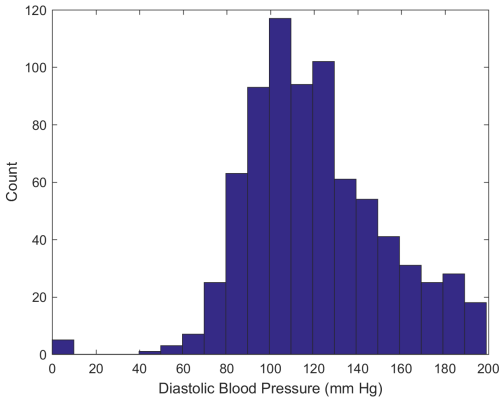

Figure 1

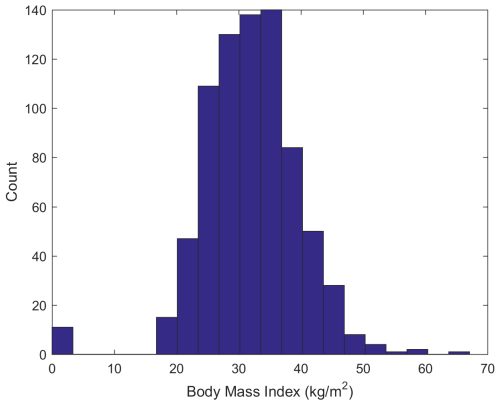

Figure 2

f. Subset Analysis

Attribute 2 of the dataset, plasma glucose concentration after 2 hours in an oral glucose tolerance test, appears to be the most helpful in discriminating between the two sets. Based on the two histograms below, figure 4 represents a normal distribution for the negative diabetes test. However, as seen in figure 3, the distribution is heavily skewed to the right, indicating a correlation between glucose levels and positive test patients.

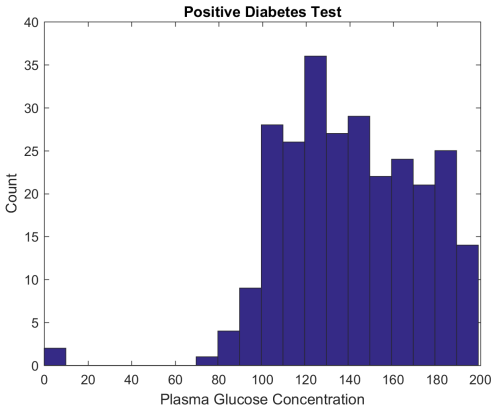

Figure 3

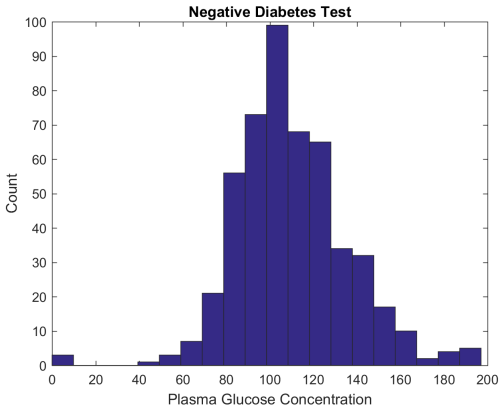

Figure 4

g. Scatter Plots

```
function scatter_plot( vector )
    scatter( vector(:,1), vector(:,2) );
end
```

Given two random variables that are independent, I would predict to see no correlation between the plotted values, which can be seen in figure 6. However, when these random variables are not quite independent, a correlation among the data points can be observed, as in figure 5. This is quite predictable in this example, as it is expected that those with

a higher BMI would also tend toward a higher skin fold thickness due to the overlapping nature of the variables.

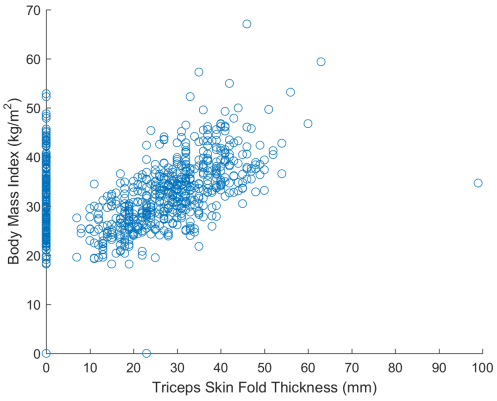

Figure 5

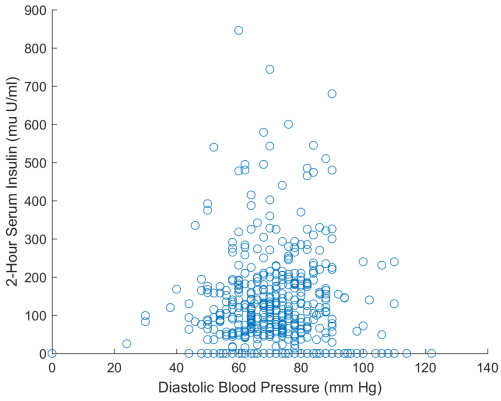

Figure 6

II. Problem 3 - Data Preprocessing

a. One's Hot Encoding

One's hot encoding is a common scheme for encoding information utilizing standard binary values in a vector whereby the index of the '1' indicates the value of the scalar. Each subsequent scalar can be converted into its corresponding one's hot vector given the following encoding:

`{brown, blue, white, red, yellow, orange, green, black}`

red	[0, 0, 0, 1, 0, 0, 0, 0]
black	[0, 0, 0, 0, 0, 0, 0, 1]
yellow	[0, 0, 0, 0, 1, 0, 0, 0]
red	[0, 0, 0, 1, 0, 0, 0, 0]
green	[0, 0, 0, 0, 0, 0, 1, 0]
blue	[0, 1, 0, 0, 0, 0, 0, 0]
blue	[0, 1, 0, 0, 0, 0, 0, 0]

b. Normalization

```
function [ normalized, mu, sigma ] = normalize( attribute )
    mu = mean(attribute);
    sigma = std(attribute);
    normalized = (attribute - mu)/sigma;
end
```

First five values of Attribute 3 (Diastolic Blood Pressure) after being normalized.

0.1495	-0.1604	-0.2638	-0.1604	-1.5037

c. Discretization

```
function [ discrete ] = discretize_attribute( attribute, k )
    min_val = min(attribute);
    max_val = max(attribute);
    bin_div = (max_val - min_val)/k;
    discrete = fix((attribute-min_val)/bin_div);
end
```

First five values of Attribute 3 (Diastolic Blood Pressure) after being discretized.

5	5	5	5	3

III. Problem 4 - Data Training

```
function [ training_set, testing_set ] = divideset( dataset, p_train )

    training_count = round(p_train * length(dataset));
    indices = randperm(length(dataset), training_count);
    t = zeros([length(dataset) 1]);
    t(indices) = 1;

    training_set = dataset(t(:) == 1,:);
    testing_set = dataset(t(:) == 0,:);

end
```

IV. Problem 5 - Matrix Operations

a. A^T

1	3
2	4
5	6

b. B^{-1}

1	-5.5	1.25
0	-0.5	0.25
-0.667	4.333	-1

c. $B + C$

15	7	14
3	-1	7
3	6	10

d. $B - C$

-1	-5	4
1	5	-1
5	10	2

e. $A * B$

31	45	45
53	59	75

f. $B * C$

48	21	75
15	0	30
34	-12	76

g. $B * A$

Cannot compute matrix operation because inner dimensions do no match.