Anthony Tummillo

Problem Assignment 1 Report

Problem 1.

Installed MatLab.

Problem 2.

a.)
Attribute 1: minimum = 0, maximum = 17, range = 17
Attribute 2: minimum = 0, maximum = 199, range = 199
Attribute 3: minimum = 0, maximum = 122, range = 122
Attribute 4: minimum = 0, maximum = 99, range = 99
Attribute 5: minimum = 0, maximum = 846, range = 846
Attribute 6: minimum = 0, maximum = 67.1, range = 67.1
Attribute 7: minimum = 0.078, maximum = 2.42, range = 2.342
Attribute 8: minimum = 21, maximum = 81, range = 60
Attribute 9: minimum = 0, maximum = 1, range = 1

b.)
Attribute 1: mean = 3.8451, variance = 11.3541
Attribute 2: mean = 120.8945, variance = 1022.2483
Attribute 3: mean = 69.1055, variance = 374.6473
Attribute 4: mean = 20.5365, variance = 254.4732
Attribute 5: mean = 79.7995, variance = 13281.1801
Attribute 6: mean = 31.9926, variance = 62.16
Attribute 7: mean = 0.4719, variance = 0.1098
Attribute 8: mean = 33.2409, variance = 138.3030
Attribute 9: mean = 0.3490, variance = 0.2275

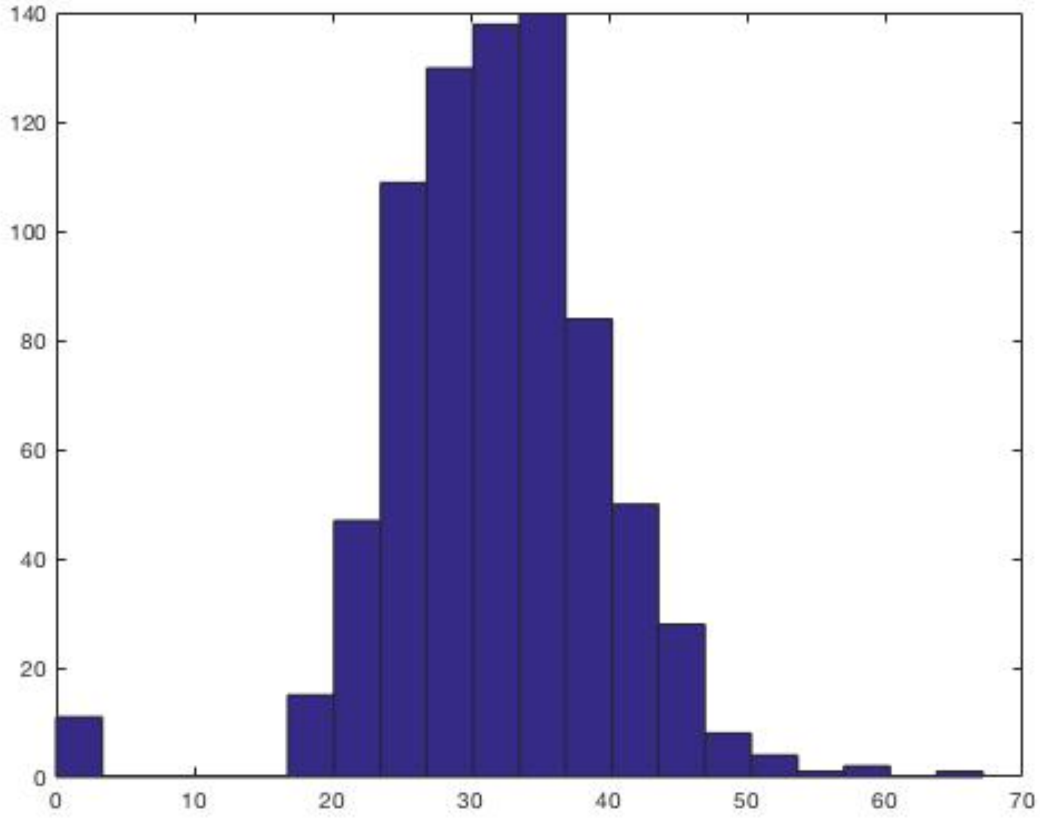
c.)
Attribute 1: correlation = 0.2219
Attribute 2: correlation = 0.4666
Attribute 3: correlation = 0.0651
Attribute 4: correlation = 0.0748
Attribute 5: correlation = 0.1305
Attribute 6: correlation = 0.2927
Attribute 7: correlation = 0.1738
Attribute 8: correlation = 0.2384

The attribute with the highest positive correlation to the target attribute is attribute 2 (Plasma glucose concentration a 2 hours in an oral glucose tolerance test). I would think that this is the most helpful attribute in predicting the target class due to its correlation being significantly higher than any other attribute (the second highest is attribute 6 whose correlation is lower by 0.1739), and given the knowledge that diabetes is a disease which directly affects blood sugar concentrations.

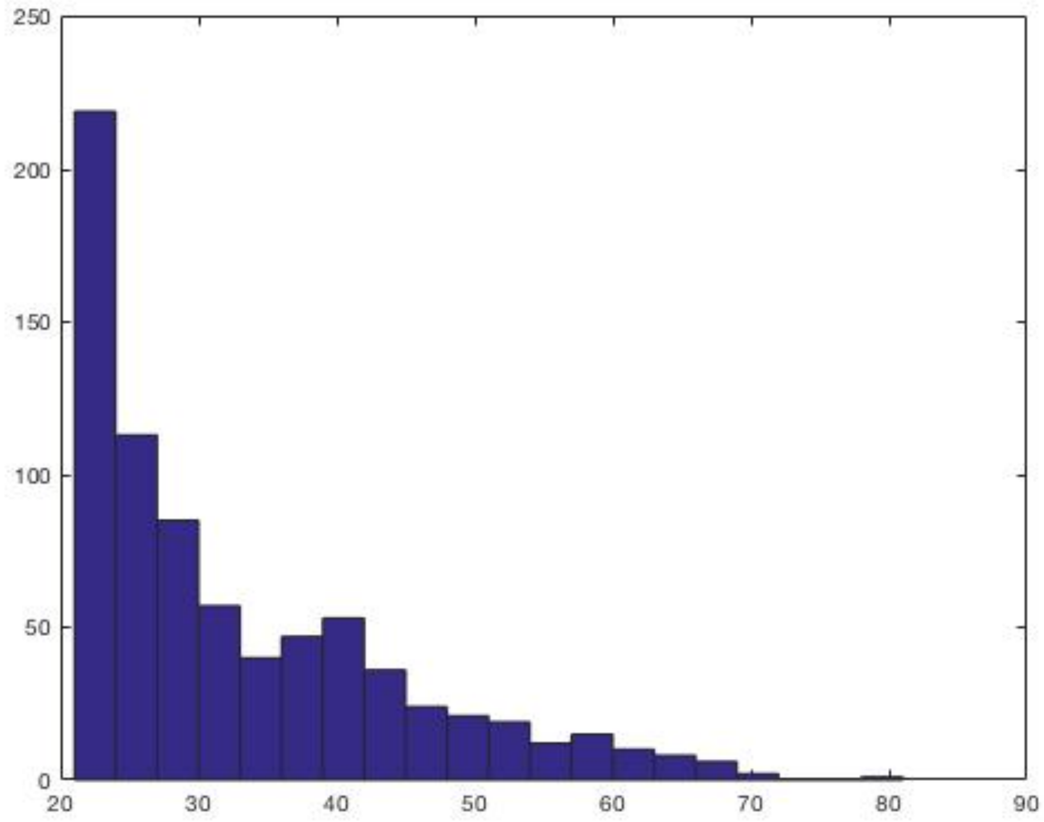
d.) Attribute 1 (Number of times pregnant) and attribute 8 (Age in years) have the largest mutual correlation at 0.5443

e.) I do **not** think having two attributes that are fully correlated would help in the prediction of the target class. While using one to predict the other could be accomplished since the two attributes are fully correlated, I believe this correlation between the two has no effect on the two attributes individual correlation/effect on the target class.

f.) I believe that the histogram representing attribute 6 (Body mass index) resembles the normal distribution the most.

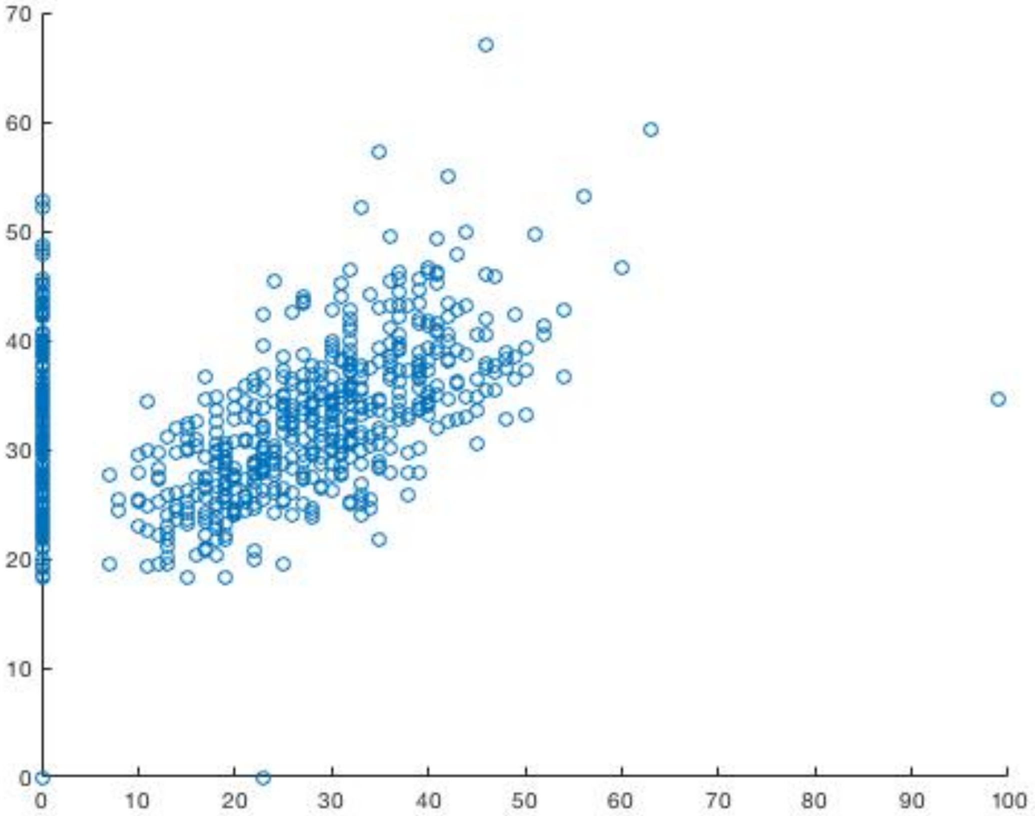


Above is the histogram of attribute 6 (Body mass index), which I believe most resembles the normal distribution.

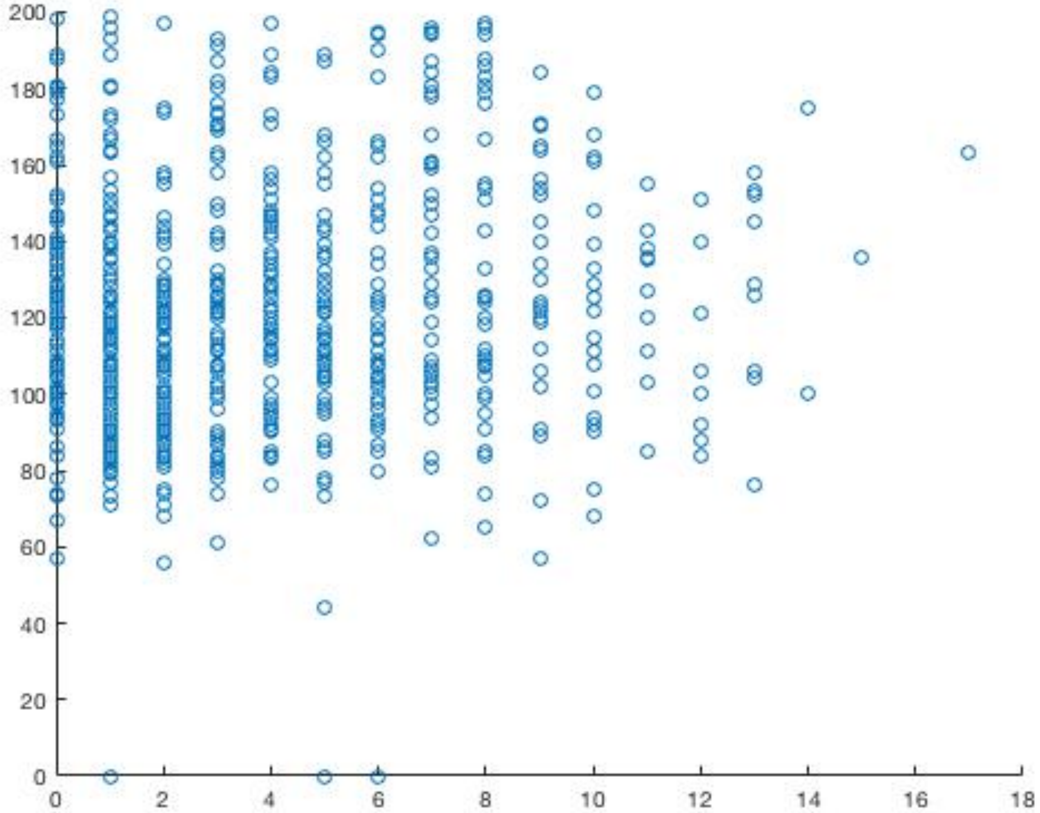


Above is the histogram of attribute 8 (Age in years) which is a good example of a histogram which does not resemble the normal distribution.

g.) I believe that the scatterplot of attribute 4 (Triceps skin fold thickness) against attribute 6 (Body mass index) indicates possible linear dependency between the two variables.



Above is attribute 4 (Triceps skin fold thickness) plotted against attribute 6 (Body mass index). A clear positive linear dependency can be observed between the two variables.



Above is attribute 1 (Number of time pregnant) plotted against attribute 2 (Plasma glucose concentration a 2 hours in an oral glucose tolerance test). It can be observed that there is no clear linear dependency between the two variables.

Problem 3.

a.) Normalized values for the first five entries of attribute 3 are listed below:

Entry1: 0.1495
Entry2: -0.1604
Entry3: -0.2638
Entry4: -0.1604
Entry5: -1.5037

b.) Discretized values for the first five entries of attribute 3 are listed below:

Entry1: 6
Entry2: 6
Entry3: 6

Entry4: 6
Entry5: 4

Problem 4.

a.)
Class label “0”
Attribute 1: mean = 3.298, standard deviation = 3.0172
Attribute 2: mean = 109.98, standard deviation = 26.1412
Attribute 3: mean = 68.184, standard deviation = 18.0631
Attribute 4: mean = 19.664, standard deviation = 14.8899
Attribute 5: mean = 68.792, standard deviation = 98.8653
Attribute 6: mean = 30.3042, standard deviation = 7.6899
Attribute 7: mean = 0.4297, standard deviation = 0.2991
Attribute 8: mean = 31.19, standard deviation = 11.6677

Class label “1”
Attribute 1: mean = 4.8657, standard deviation = 3.7412
Attribute 2: mean = 141.2575, standard deviation = 31.9396
Attribute 3: mean = 70.8246, standard deviation = 21.4918
Attribute 4: mean = 22.1642, standard deviation = 17.6797
Attribute 5: mean = 100.3358, standard deviation = 138.6891
Attribute 6: mean = 35.1425, standard deviation = 7.263
Attribute 7: mean = 0.5505, standard deviation = 0.3724
Attribute 8: mean = 37.0672, standard deviation = 10.9683

b.)
The average length of the training dataset: 504

c.)
When running divideset2 with p_train = 0.66 I get unique, random data sets in which my training set has 507 rows and my testing set has 261 rows.

507/768 = 0.6602 which is very close to 0.66 so I believe this function is working properly.