

## Solutions to Problem set 1

### 1 Problem 1.

Thank you.

### 2 Problem 2. Exploratory data analysis

(a)

Column	Attribute	Minimum	Maximum
1	Number of times pregnant	0	17
2	Plasma glucose concentration	0	199
3	Diastolic blood pressure	0	122
4	Triceps skin fold thickness	0	99
5	2-Hour serum insulin	0	846
6	Body mass index	0	67.1
7	Diabetes pedigree function	0.078	2.42
8	Age	21	81

(b)

Column	Attribute	Mean	Variances
1	Number of times pregnant	3.8451	11.3541
2	Plasma glucose concentration	120.8945	1.0222e+03
3	Diastolic blood pressure	69.1055	374.6473
4	Triceps skin fold thickness	20.5365	254.4732
5	2-Hour serum insulin	79.7995	1.3281e+04
6	Body mass index	31.9926	62.16
7	Diabetes pedigree function	0.4719	0.1098
8	Age	33.2409	138.3030

(c)

Column	Attribute	Correlation
1	Number of times pregnant	0.2219
2	Plasma glucose concentration	0.4666
3	Diastolic blood pressure	0.0651
4	Triceps skin fold thickness	0.0748
5	2-Hour serum insulin	0.1305
6	Body mass index	0.2927
7	Diabetes pedigree function	0.1738
8	Age	0.2384

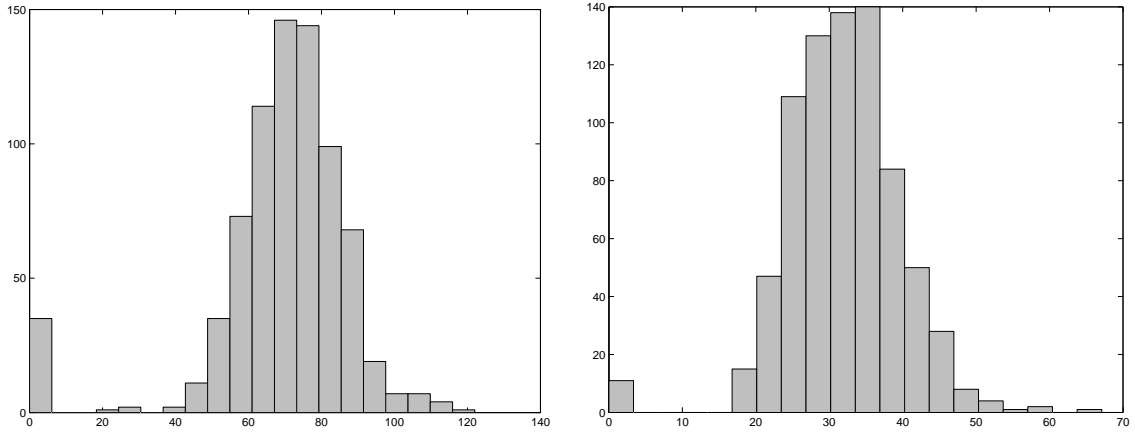


Figure 1: Histogram of the third attribute (Diastolic blood pressure) and the sixth attribute (Body mass index).

The highest positive correlation 0.4666 to the target attribute has attribute Plasma glucose concentration. The attribute is helpful in the prediction of the target attribute because high correlation means linear relationship between the variables. However, only correlation does not necessarily imply causation.

- (d) The largest mutual correlation 0.5443 is between the first (Number of times pregnant) and eight attribute (Age).
- (e) Using two fully correlated attributes to predict a class variable does not help much. The correlation measures the linear dependence between values of attributes. So if the two attributes are fully correlated the value of the first attribute can be computed using the value of the second attribute, the first attribute is not needed anymore.
- (f) There are 2 histograms that resemble normal distribution; the histograms of third (Diastolic blood pressure) and sixth (Body mass index) attribute.
- (g) There is no scatter plot that clearly resembles linear dependencies between two attributes. Probably one of the closest to straight line is the plot between the first (Number of times pregnant) and eight attribute (Age).

### 3 Problem 3. Data preprocessing

(a)

Original value	Normalized value
72	0.1495
66	-0.1604
64	-0.2638
66	-0.1604
40	-1.5037

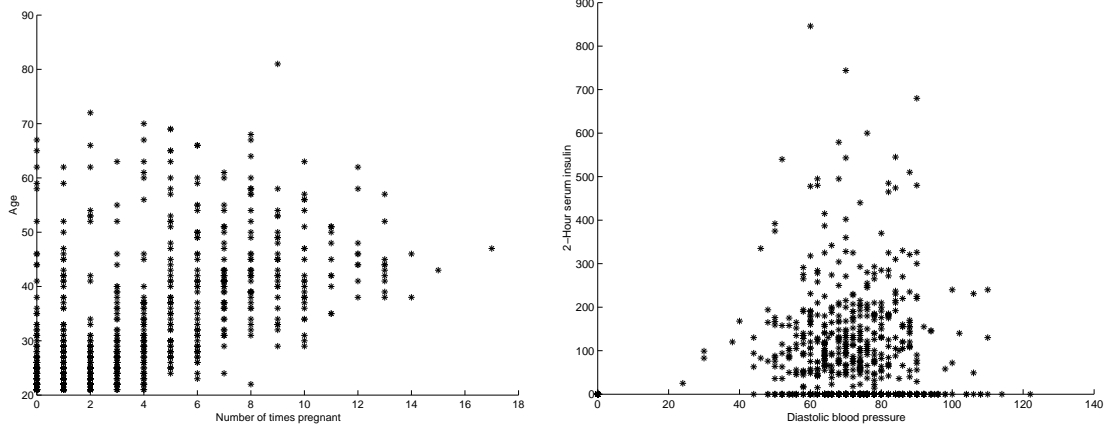


Figure 2: Left. Scatter plot between between the first (Number of times pregnant) and eight attribute (Age). Right. Scatter plot between between the third (Diastolic blood pressure) and fifth attribute (2-Hour serum insulin).

- (b) The attribute was discretized into the following list of bins. The label of every bin is the average between the smallest and the largest value in the bin.

Bin	Label
1	6.1
2	18.3
3	30.5
4	42.7
5	54.9
6	67.1
7	79.3
8	91.5
9	103.7
10	115.9

The first five values of the third attribute in the dataset were discretized on the basis of a closest label as

Original value	Discretized value	Bin
72	67.1	6
66	67.1	6
64	67.1	6
66	67.1	6
40	42.7	4

## 4 Problem 4. Data set splitting

(a) The mean and variance of samples with class label "0" is

Column	Attribute	Mean	Standard deviation
1	Number of times pregnant	3.2980	3.0172
2	Plasma glucose concentration	109.9800	26.1412
3	Diastolic blood pressure	68.1840	18.0631
4	Triceps skin fold thickness	19.6640	14.8899
5	2-Hour serum insulin	68.7920	98.8653
6	Body mass index	30.3042	7.6899
7	Diabetes pedigree function	0.4297	0.2991
8	Age	31.1900	11.6677

The mean and variance of samples with class label "1" is

Column	Attribute	Mean	Standard deviation
1	Number of times pregnant	4.8657	3.7412
2	Plasma glucose concentration	141.2575	31.9396
3	Diastolic blood pressure	70.8246	21.4918
4	Triceps skin fold thickness	22.1642	17.6797
5	2-Hour serum insulin	100.3358	138.6891
6	Body mass index	35.1425	7.2630
7	Diabetes pedigree function	0.5505	0.3724
8	Age	37.0672	10.9683

(b) Probability  $p_{train} = 0.66$  yields the average length of the training set 505.

(c) We generate a random permutation of all samples. Then the first  $0.66 \times 768$  (number of samples) in the permutation are used as training examples and others are used as testing examples.