Mouhammadou Dabo

February 4, 2021

CS 1675: Intro to Machine Learning

Professor Milos Hauskrecht

Handout 1 – Problem Assignment

Problem 1. Matrix operations

- $u^T * u = 26$
- $u * u^T = \begin{matrix} 16 & 4 & 12 \\ 4 & 1 & 3 \\ 12 & 3 & 9 \end{matrix}$
- $v * u = 71$
- $u + 5 = \begin{matrix} 9 \\ 6 \\ 8 \end{matrix}$
- $A^T = \begin{matrix} 1 & 3 \\ 2 & 4 \\ 5 & 6 \end{matrix}$
- $B * u = \begin{matrix} 56 \\ 19 \\ 42 \end{matrix}$
- $B^{-1} = \begin{matrix} 1.0 & -5.5 & 1.25 \\ 0 & -0.5 & 0.25 \\ -0.6667 & 4.3333 & -1.0 \end{matrix}$
- $B + C = \begin{matrix} 15 & 7 & 14 \\ 3 & -1 & 7 \\ 3 & 6 & 10 \end{matrix}$
- $B - C = \begin{matrix} -1 & -5 & 4 \\ 1 & 5 & -1 \\ 5 & 10 & 2 \end{matrix}$
- $A * B = \begin{matrix} 31 & 45 & 45 \\ 53 & 59 & 75 \end{matrix}$
- $B * C = \begin{matrix} 48 & 21 & 75 \\ 15 & 0 & 30 \\ 34 & -12 & 76 \end{matrix}$
- $B * A$ – Error using * - Incorrect dimensions for matrix multiplication. Check that the number of columns in the first matrix matches the number of rows in the second matrix.

Problem 2. Exploratory data analysis

(a)

- Number of times pregnant
  - Min value: 0
  - Max value: 17
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Min value: 0
  - Max value: 199
- Diastolic blood pressure (mm Hg)
  - Min value: 0
  - Max value: 122
- Triceps skin fold thickness (mm)
  - Min value: 0
  - Max value: 99
- 2-Hour serum insulin (mu U/ml)
  - Min value: 0
  - Max value: 846
- Body mass index (weight in kg/(height in m)^2)
  - Min value: 0
  - Max value: 67.1
- Diabetes pedigree function
  - Min value: 0.078
  - Max value: 2.4
- Age (years)
  - Min value: 21
  - Max value: 81
- Class variable (0 or 1)
  - Min value: 0
  - Maxi value: 1

(b)

- Number of times pregnant
  - Mean value: 3.8451
  - STD value: 3.3696
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Mean value: 120.8945
  - STD value: 31.9726
- Diastolic blood pressure (mm Hg)
  - Mean value: 69.1055
  - STD value: 19.3558
- Triceps skin fold thickness (mm)
  - Mean value: 20.5365
  - STD value: 15.9522

- 2-Hour serum insulin (mu U/ml)
    - Mean value: 79.7995
    - STD value: 115.2440
- Body mass index (weight in kg/(height in m)^2)
    - Mean value: 31.9926
    - STD value: 7.882
- Diabetes pedigree function
    - Mean value: 0.4719
    - STD value: 0.3313
- Age (years)
    - Mean value: 32.2409
    - STD value: 11.7602
- Class variable (0 or 1)
    - Mean value: 0.3490
    - STD value: 0.477

(c)

Label 0

- Number of times pregnant
    - Mean value: 3.298
    - STD value: 3.0172
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
    - Mean value: 109.98
    - STD value: 26.1412
- Diastolic blood pressure (mm Hg)
    - Mean value: 68.184
    - STD value: 18.0631
- Triceps skin fold thickness (mm)
    - Mean value: 19.664
    - STD value: 14.8899
- 2-Hour serum insulin (mu U/ml)
    - Mean value: 68.7920
    - STD value: 98.8653
- Body mass index (weight in kg/(height in m)^2)
    - Mean value: 30.3042
    - STD value: 7.6899
- Diabetes pedigree function
    - Mean value: 0.4297
    - STD value: 0.2991
- Age (years)
    - Mean value: 31.1900
    - STD value: 11.6677

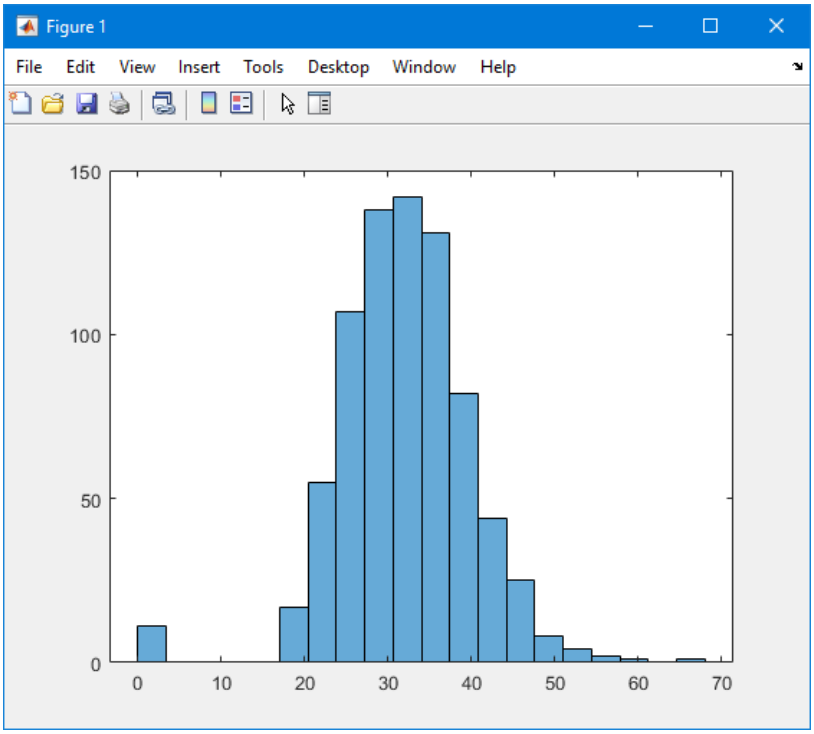
Label 1

- Number of times pregnant
  - Mean value: 4.8657
  - STD value: 3.7412
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Mean value: 141.2575
  - STD value: 31.9396
- Diastolic blood pressure (mm Hg)
  - Mean value: 70.8246
  - STD value: 21.4918
- Triceps skin fold thickness (mm)
  - Mean value: 22.1642
  - STD value: 17.6797
- 2-Hour serum insulin (mu U/ml)
  - Mean value: 100.3358
  - STD value: 138.6891
- Body mass index (weight in kg/(height in m)^2)
  - Mean value: 35.1425
  - STD value: 7.2630
- Diabetes pedigree function
  - Mean value: 0.5505
  - STD value: 0.3724
- Age (years)
  - Mean value: 37.0672
  - STD value: 10.9683

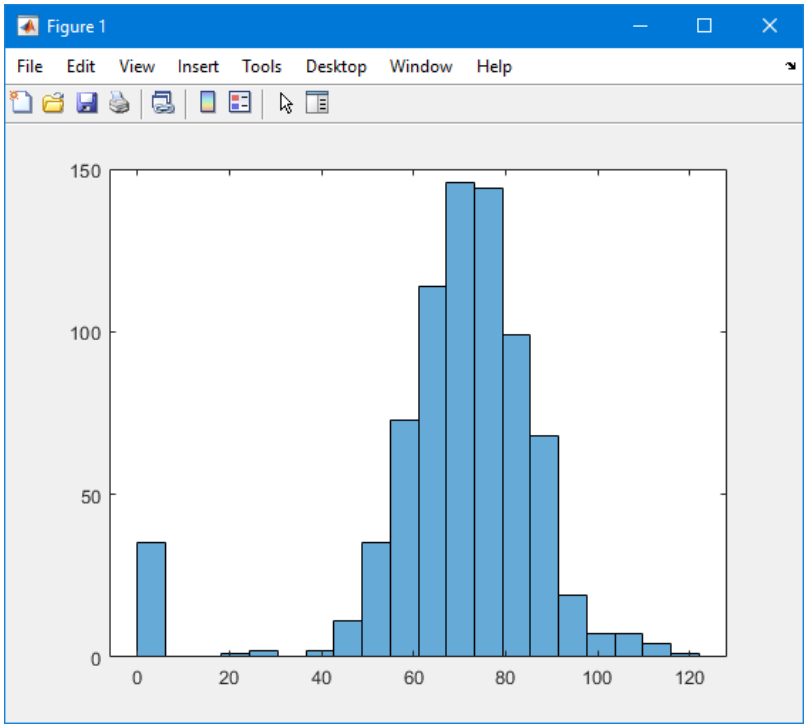
The attribute that would be most helpful in discriminating the two classes is the 2-Hour serum insulin, since it is the only attribute with a big variance between the two classes.

(g) The histogram that resembles most the normal distribution is the histogram of attribute 6. The other histogram that resembles a normal distribution is the histogram of attribute 3.

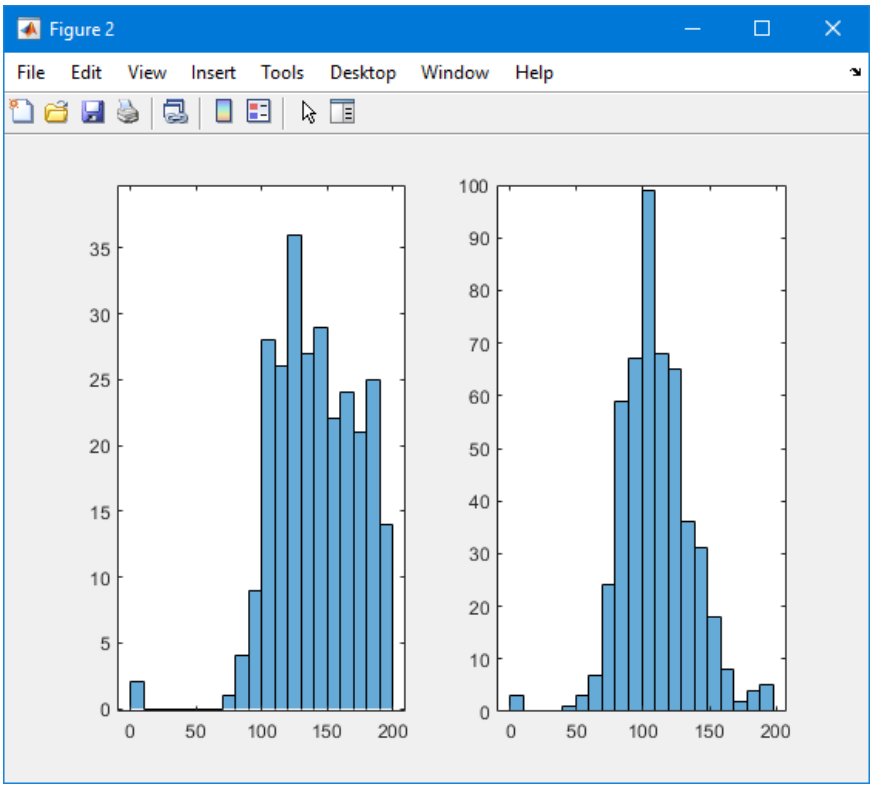
Histogram for Attribute 6



Histogram for Attribute 3



(h) Based on the pairs of histograms, the attribute I think is the most helpful in discriminating the two classes would be Attribute 2 (Plasma glucose concentration a 2 hours in an oral glucose tolerance test), since the range between the two classes are very different.



Problem 3. Data preprocessing

(a)

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

- Brown: (1, 0, 0, 0, 0, 0, 0, 0);
- Blue: (0, 1, 0, 0, 0, 0, 0, 0);
- White: (0, 0, 1, 0, 0, 0, 0, 0);
- Red: (0, 0, 0, 1, 0, 0, 0, 0);
- Yellow: (0, 0, 0, 0, 1, 0, 0, 0);
- Orange: (0, 0, 0, 0, 0, 1, 0, 0);
- Green: (0, 0, 0, 0, 0, 0, 1, 0);
- Black: (0, 0, 0, 0, 0, 0, 0, 1);

Since there are 8 different categories, I used a vector of size 8 with binary values

$$\begin{bmatrix} red \\ black \\ yellow \\ red \\ green \\ blue \\ blue \end{bmatrix}$$

$$=$$

$$\begin{bmatrix} 0,0,0,1,0,0,0,0 \\ 0,0,0,0,0,0,0,1 \\ 0,0,0,0,1,0,0,0 \\ 0,0,0,1,0,0,0,0 \\ 0,0,0,0,0,0,1,0 \\ 0,1,0,0,0,0,0,0 \\ 0,1,0,0,0,0,0,0 \end{bmatrix}$$

(b)

Attribute 3 - Diastolic blood pressure (mm Hg)

- Mean value: 69.1055
- STD value: 19.3558
- First Five Normalized Values:
  - 0.1495
  - -0.1604
  - -0.2638
  - -0.1604
  - -1.5037

(c)

Entry 1: Bin 6
Entry 2: Bin 6
Entry 3: Bin 5
Entry 4: Bin 6
Entry 5: Bin 4