CS383 Assignment1

Himanshu Gupta

January 2020

Question 1

1 Average Entropy Calculation for first feature

$$TheMatrixX = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix} \quad TheMatrixY = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Dividing the Matrix X in two classes: positive and negative based on the Label Matrix Y with negative class for 0 and positive class for 1.

$$Positive \quad (p) = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix} \quad Negative \quad (n) = \begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

$$DataLabels \in \begin{Bmatrix} -2 & -5 & -3 & -4 & -1 & 1 & 0 & -8 & 6 & 5 & 11 & 3 \end{Bmatrix}$$

For first feature,
p(-2) = 1 n(-2) = 1
p(-5) = 1 n(-5) = 0
p(-3) = 1 n(-3) = 0
p(-4) = 0 n(-4) = 0
p(-1) = 0 n(-1) = 1
p(1) = 0 n(1) = 1

p(0) = 1   n(0) = 0
p(-8) = 1   n(-8) = 0
p(6) = 0   n(6) = 1
p(5) = 0   n(5) = 1
p(11) = 0   n(11) = 0
p(3) = 0   n(3) = 0

Now calculate the entropy by putting the respective p and n values in entropy formula
Entropy (1st Feature)

$= \{\frac{2}{10} \times (-\frac{1}{2} \times log_2 \frac{1}{2} - \frac{1}{2} \times log_2 \frac{1}{2})\} + \{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} +$

$\{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} + \{0\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} +$

$\{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} + \{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} +$

$\{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} +$

$\{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} + \{0\} + \{0\}$

$= 0.2 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0.2$

2 Average Entropy Calculation for second feature

p(-2) = 0   n(-2) = 1
p(-5) = 0   n(-5) = 0
p(-3) = 0   n(-3) = 1
p(-4) = 1   n(-4) = 0
p(-1) = 0   n(-1) = 1
p(1) = 2   n(1) = 1
p(0) = 0   n(0) = 1
p(-8) = 0   n(-8) = 0
p(6) = 0   n(6) = 0
p(5) = 0   n(5) = 1
p(11) = 1   n(11) = 0
p(3) = 1   n(3) = 0

Entropy (1st Feature)

$= \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - \frac{1}{1} \times log_2 \frac{1}{1})\} + \{0\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} +$

$\{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} +$

$\{\frac{3}{10} \times (-\frac{2}{3} \times log_2 \frac{2}{3} - \frac{1}{3} \times log_2 \frac{1}{3})\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} + \{0\} +$

$\{0\} + \{\frac{1}{10} \times (-\frac{0}{1} \times log_2 \frac{0}{1} - 1 \times log_2 \frac{1}{1})\} + \{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} +$

$\{\frac{1}{10} \times (-\frac{1}{1} \times log_2 \frac{1}{1} - 0 \times log_2 \frac{0}{1})\} +$

$= 0 + 0 + 0 + 0 + 0 + 0.2755 + 0 + 0 + 0 + 0 + 0 + 0 = 0.2755$

3 Which feature is more deterministic ?

From the results of part 1 and part 2, first feature has lower entropy $(0.2 < 0.2755)$ meaning lower randomness which in itself means more deterministic projection of data along that feature. Hence, First feature results in better projection of data than that of second feature.

4 Principal components of Matrix X ?

To calculate principal components of matrix X, I am using svd Matlab to get Eigenvectors matrix (V) after standardising the data by making data having zero mean and divided by its standard deviation. Using svd, columns of Matrix V are already in descending order of eigenvalues so for 1-D projection, taking only the first column of V.

$$[U, S, V] = svd(X)$$

$$The \;\; Matrix \;\; V = \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix}$$

Also, the length of both the eigenvectors is 1

$$\|length(eigenvectors)\| \equiv 1$$

$$\text{P}C1 = \begin{bmatrix} 0.7071 \\ -0.7071 \end{bmatrix} \quad \text{P}C2 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

5 1-D Projection of Matrix X:

Now to get the 1D data projection, multiply the (10 X 2) Matrix X with the first (2 X 1) eigenvector Matrix.

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix} \times \begin{bmatrix} 0.7071 \\ -0.7071 \end{bmatrix}$$

The (10 X 1) result matrix =

$$\begin{bmatrix} -0.1178 \\ 0.2077 \\ -0.2850 \\ -0.1142 \\ -2.7756 \\ -0.7796 \\ 0.5494 \\ 1.3837 \\ 0.7112 \\ 1.2201 \end{bmatrix}$$

Question 2

Source code is in AssgnPart2.m and the 2-D Plot is