DATASET

weight	**heartrate**	**>150 situps**
heavy	slow	*yes*
heavy	slow	*no*
heavy	fast	*no*
light	fast	*no*
heavy	slow	*yes*
heavy	fast	*no*
heavy	fast	*no*
light	fast	*no*
light	fast	*yes*
light	fast	*yes*
light	slow	*no*
light	slow	*yes*
light	fast	*yes*
heavy	slow	*no*
heavy	slow	*no*
heavy	fast	*yes*
light	slow	*no*
light	slow	*yes*
light	slow	*yes*
light	fast	*no*

consider the following dataset with 2 attributes (weight, heartrate) and two classes (>150 situps=yes, >150 situps=no)

Question #1 1R

Using the 1R algorithm, answer the following:
a) build the table showing the rules for each attribute-value and error rate of each attribute overall
b) list the two rules given by the attribute with the lowest error rate from (a)

a)

Attribute	Rules	Errors	Total errors
weight	Heavy-> no	3/9	8/20
	Light-> yes	5/11	
heartrate	Fast-> no	4/10	9/20
	Slow-> yes	5/10	

b) if weight=heavy then >150 situps=no
if weight=light then >150 situps=yes

Question #2 Naive Bayes

a) what is the probability >150 situps=yes given weight=heavy and heartrate=slow?
b) show

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

using the symmetry property:

$$\text{P(A,B) = P(B,A)}$$

and the product rule:

$$\text{P(A,B) = P(A|B)P(B)}$$

marking breakdown:
25% question #1
30% question #2
30% question #3
15% question #4

midterm is worth
25% of final grade

a)

$$p(yes|heavy, slow) = \frac{p(heavy|yes)p(slow|yes)p(yes)}{p(heavy|yes)p(slow|yes)p(yes)+p(heavy|no)p(slow|no)p(no)}$$

$$= \frac{\left(\frac{3}{9}\right)\left(\frac{5}{9}\right)\left(\frac{9}{20}\right)}{\left(\frac{3}{9}\right)\left(\frac{5}{9}\right)\left(\frac{9}{20}\right)+\left(\frac{6}{11}\right)\left(\frac{5}{11}\right)\left(\frac{11}{20}\right)} = 0.379$$

b) P(A|B)P(B) = P(B|A)P(A) => P(A|B) = P(B|A)P(A) / P(B)
(symmetry + product rule) (algebra)

Question #3 Decision tree

which attribute should be placed as the root for a decision tree of this dataset (using gini index)

solution) we need to check the gini index for both weight and heartrate
weight: weight=heavy (9 instances total, 3 yes ,6 no)
weight=light (11 instance total, 6 yes, 5 no)

heartrate: heartrate=slow (10 instances,5 yes, 5 no)
heartrate=fast (10 instances, 4 yes, 6 no)

=> gini(weight) =

$$1 - \frac{9}{20}\left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right) - \frac{11}{20}\left(\left(\frac{6}{11}\right)^2 + \left(\frac{5}{11}\right)^2\right) = 0.473$$

=> gini(heartrate) =

$$1 - \frac{10}{20}\left(\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right) - \frac{10}{20}\left(\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right) = 0.49$$

=> weight should be placed as the root

Question #4 linear models

Let $(X^T X)^{-1} X^T = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \\ 3 & 1 & 3 \end{vmatrix}$

Let Y = $\begin{vmatrix} 1 & 2 & 3 \end{vmatrix}^T$

A) find W, where $W = (X^T X)^{-1} X^T Y$

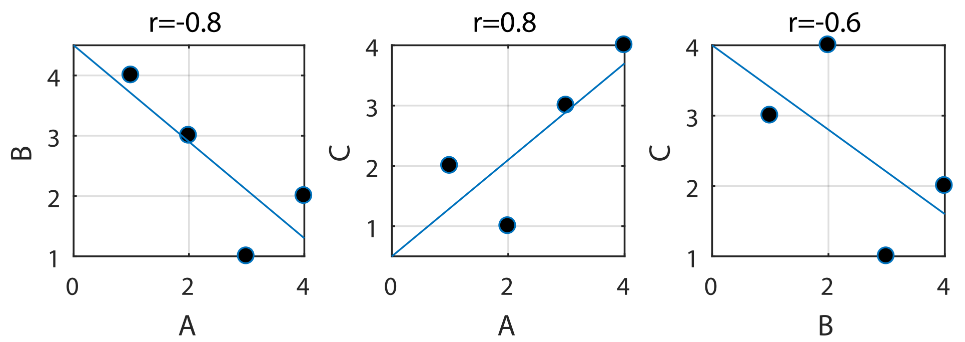
B) Let A = [1 2 3 4], B=[4 3 1 2], C = [2 1 3 4] be row vectors
which pair of vectors has the highest correlation coefficient?
hint - plot a scatter plot for each pair of vectors

A) Using matrix multiplication:

$W = 0 \cdot 1 + 1 \cdot 2 + 1 \cdot 3, \quad 1 \cdot 1 + 2 \cdot 2 + 0 \cdot 3 \quad 3 \cdot 1 + 1 \cdot 2 + 3 \cdot 3$

$W = 5 \quad 5 \quad 14$

B)



clearly, A,C has the highest correlation coefficient (r-value)