Problem 1.1

(a)

In the linear regression model

$y=x^T$ beta

In the closed form:

beta=$(x^T x)^{-1} x^T y$.

Plug the number in, we could obtain:

beta=[26.7808, 0.6438]

(b)

With beta=[26.7808, 0.6438], plug in the x vector with value [95,85,80,70,60]. We could obtain the y vector to be [87.94,81.50,78.29,71.85,65.41]

Therefore, the predicted math scores are 87.94, 81.50, 78.29, 71.85, 65.41.

Problem 1.2

(a)

Beta

ALL the beta is in the beta excel file! It is easy to view in the excel file

MSE:

closed_form MSE:

4.396097860818842

gradient descent MSE:

4.390523165584154

gradient_descent_stochaistic MSE:

4.379961758470524

The beta and MSE are not exactly the same, the reason is due to the computation error.

(b)

Beta

ALL the beta is in the beta excel file! It is easy to view in the excel file

normalized closed_form MSE:

4.396097860818446

normalized gradient descent MSE:

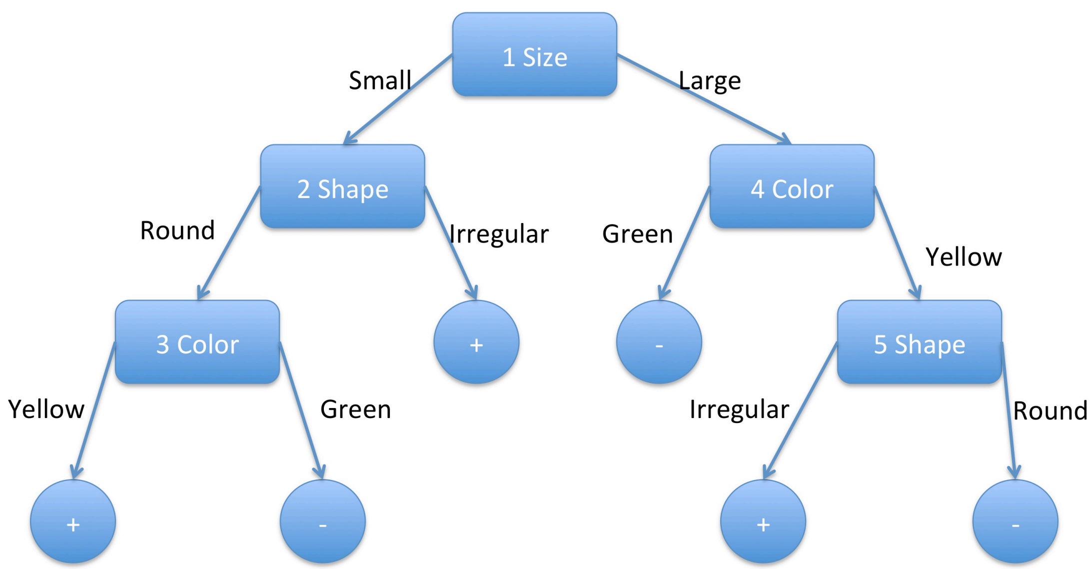
4.393051392930504

normalized gradient_descent_stochaistic MSE:

4.404777368315331

Yes, it affects, but **very slightly**. The reason is that the normalization only change the magnitude and center of the X distribution and distribution's shape is same. Another reason is that the linear regression assumes the sample's residual error has a normal distribution and the sample used has such distribution.

Problem 2.1



1st Node: Info(D)=0.9886
If split on color:
$Info_a(D)=13/16(5/13*log(5/13)+8/13*log(8/13))+3/16(1/3*log(1/3)+2/3*log(2/3))=0.95381$
If split on size:
$Info_a(D)=8/16(6/8*log(6/8)+2/8*log(2/8))+8/16(3/8*log(3/8)+5/8*log(5/8))=0.88$
If split on size:
$Info_a(D)=12/16(6/12*log(6/12)+6/12*log(6/12))+4/16(3/4*log(3/4)+$

1/4*log(1/4))=0.9528
Therefore the info gain is largest with size, split on size

2nd Node:
If split on shape:
$Info_a(D)$=6/8(4/6*log(4/6)+2/6*log(2/6))=0.68
If split on color:
$Info_a(D)$=6/8(1/6*log(1/6)+5/6*log(5/6))+ 2/8(1/2*log(1/2)+1/2*log(1/2))=0.73

So split on shape

3rd node: could only split on color

4nd node:
If split on color:
$Info_a(D)$= 7/8(3/7*log(3/7)+4/7*log(4/7))=0.86
If split on the shape:
$Info_a(D)$= 2/8(1/2*log(1/2)+1/2*log(1/2))+ 2/8(1/2*log(1/2)+1/2*log(1/2))=0.93
Therefor split on the shape

5th node: could only split on shape

Problem 2.2
(a)
InfoGain Accuracy: 0.9310344827586239
(b)
InfoGainRatio Accuracy: 0.9264367816091953

InfoGain Ratio (C4.5) solves the bias towards the attributes with a lot values. In such questions, all the attributes have three values. So it is similar choosing InfoGain or InfoGain Ratio. But InfoGain Ratio's is divided by SplitInfo and SplitInfo may have some fluctuation. So choosing InfoGain.

Problem 3

$$I(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

$$= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)}\right) - \sum_x \sum_y p(x,y) \log\left(p(y)\right)$$

$$= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)}\right) - \sum_x \sum_y p(x,y)\log\left(p(y)\right)$$

$$= \sum_x \sum_y p(x)p(y|x) \log(p(y|x)) - \sum_x \sum_y p(x,y)\log\left(p(y)\right)$$

$$= \sum_x p(x) \sum_y p(y|x) \log(p(y|x)) - \sum_x \sum_y p(x)p(y)\log\left(p(y)\right)$$

$$= \sum_x p(x) \sum_y p(y|x) \log(p(y|x)) - \sum_y p(y)\log\left(p(y)\right) \sum_x p(x)$$

$$= \sum_x p(x) \sum_y p(y|x) \log(p(y|x)) - \sum_y p(y)\log\left(p(y)\right)$$

Let y be the classification label and x be the attribute. The first term is the information needed after splitting the data set with attribute x. The second term is the original information needed to classify it.

Therefore, the results is information gain. That is to say, mutual information is same as information gain.