Week 2 Assignment Solutions:

1. In regression the output is
   A) Discrete.
   B) Continuous and always lies in a finite range.
   C) Continuous.
   D) May be discrete or continuous.
   Solution: C

2. In linear regression the parameters are
   A) strictly integers
   B) always lies in the range [0,1]
   C) any value in the real space
   D) any value in the complex space
   Solution: C

3. Which of the following is true for a decision tree?
   A) Decision tree is an example of linear classifier.
   B) The entropy of a node typically decreases as we go down a decision tree.
   C) Entropy is a measure of purity
   D) An attribute with lower mutual information should be preferred to other attributes.
   Solution: B

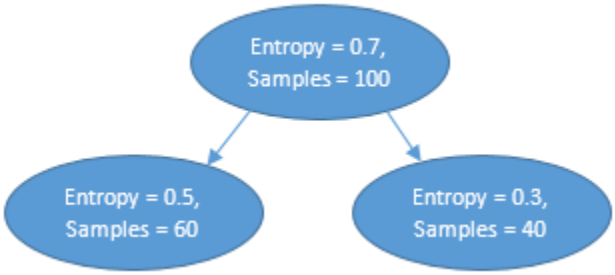
4. Suppose S a collection of objects that belong to two classes. Let the number of elements of class I be p and the number of elements of class two be q. Which of the following is the correct expression for entropy of S with respect to this binary classification.

   A) $E(S) = \frac{p}{p+q} log_2\left(\frac{p}{p+q}\right) - \frac{q}{p+q} log_2\left(\frac{q}{p+q}\right)$
   B) $E(S) = \frac{p}{p+q} log_2\left(\frac{p}{p+q}\right) + \frac{q}{p+q} log_2\left(\frac{q}{p+q}\right)$
   C) $E(S) = -\frac{p}{p+q} log_2\left(\frac{p}{p+q}\right) - \frac{q}{p+q} log_2\left(\frac{q}{p+q}\right)$
   D) $E(S) = -\frac{p}{p+q} log_2\left(\frac{p}{p+q}\right) + \frac{q}{p+q} log_2\left(\frac{q}{p+q}\right)$
   Solution: C [Marks will be given to all students]

5. Given (x1,y1),(x2,y2),…,(xn,yn), best fitting data to y = f(x) by least squares requires minimization of
   A) $\sum_{i=1}^{n}[y_i - f(x_i)]$
   B) $\max(y_i - f(x_i))$
   C) $\min(y_i - f(x_i))$
   D) $\sum_{i=1}^{n}[y_i - f(x_i)]^2$
   Solution: D

6. Given a list of 14 examples including 9 positive and 5 negative examples. The entropy of the dataset with respect to this classification is
   A) 0.940
   B) 0.06
   C) 0.5
   D) 0.22
   Solution: A

7. What is the value of information gain in the following partitioning?



   A) 0.72
   B) 0.42
   C) 0.28
   D) 0.30
   Solution: C

8. The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute back-ache. Find the equation of the best fit line for this data.

Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

   (a) $y = 3.39x + 11.62$
   (b) $y = 4.69x + 12.58$
   (c) $y = 4.59x + 12.58$
   (d) $y = 3.59x + 10.58$

Solution: C

[Hints: For each x calculate the value of y using the given equations. Then calculate error for each equation. Equation with lowest error is the desired answer. For error calculation you may take help from question no 5.]

Programming Question:

A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. Use this dataset to build a decision tree, with Buys as the target variable, to help in buying lip-sticks in the future.

ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10	> 35	Medium	Female	Single	Yes
11	< 21	Medium	Female	Married	Yes
12	21-35	Medium	Male	Married	Yes
13	21-35	High	Female	Single	Yes
14	> 35	Medium	Male	Married	No

You can use sklearn.tree.DecisionTreeClassifier for solving the problem.

Please download the data for this question here (the qualitative fields of the data in the table above has been converted into numbers) and place it in your present working directory. The following python code will load the inputs and targets from the .txt file into the numpy matrices x_train and y-train:

```
import numpy as np
f = open('decision_tree_data.txt','r')
x_train = []
y_train = []

for line in f:
    line = np.asarray(line.split(),dtype=np.float32)
    x_train.append(line[:-1])
    y_train.append(line[-1])

x_train = np.asmatrix(x_train)
y_train = np.reshape(y_train,(len(y_train),1))
```

Now answer the following questions about the decision tree you made:

9. Which of the attributes would be the root node.
   A. Age
   B. Income
   C. Gender
   D. Marital Status
   Solution: C

10. What is the decision for the test data [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?
    A. Yes
    B. No
    Solution: A
    [Hints: construct the decision tree to answer these questions]