CSE4088
Intro. to Machine Learning
Homework #3

Part 1 - Gradient Descent

1. Question 4

I have calculated partial derivative with sympy library and its "diff" method.

```
Question 4: [e] Partial derivative of E(u,v) with respect to u is (u*exp(v) - 2*v*exp(-u))*(4*v*exp(-u) + 2*exp(v))
```

2. Question 5

In this question, I have used "Simultaneous update" approach to find answer. First, I have calculated both derivative values of u and v, then decreased u and v values as muplication of derivative result and learning rate(eta) which is 0.1 for this question

Correct: Simultaneous update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

```
Question 5: [d] It takes 10 iterations
```

3. Question 6

When the error is reached 10^{-14}, last u and v values are printed.

```
Question 6: [e] The closest values among the following choices to the final (u,v) is (0.045, 0.024)
```

4. Question 7

In this question, now, I have used "incorrect" implementation of gradient descent. Updated u value is used to calculate derivative value of v and this gives us incorrect result. I have reached only 10^{-1} after 30 iterations.

Incorrect:

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 := \text{temp1}$$

```
Question 7: [a] The error E(u,v) be closest after 15 full iterations is 0.140
```

Part 2 – Logistic Regression

5. Question 8

I have generated 100 different points with using uniform probability distribution between -1 and 1. I have run Logistic Regression with Stochastic Gradient Descent to _nd g, and estimate Eout (the cross entropy error) by generating a sufficiently large, separate set of points to evaluate the error. Weights are initialized zero initially and when difference between current and previous weight is less than 0.01, the algorithm stops. This experiment is repeated 100 times.

```
Question 8: [d] The closest to E_out for N=100 is: 0.086
```

6. Question 9

It takes 318.5 epochs on average for Logistic Regression to converge for N = 100 using the above initialization and termination rules and the specified learning rate

```
Question 9: [a] It takes 318.5 epochs on average
```

Part 3 - Regularization with weight decay

7. Question 2

I have read data from “in.dta” and “out.dta” and applied Linear Regression with a non-linear transformation for classiffication the data. The nonlinear transformation is given in the question. The closest (in Euclidean distance) to the in-sample and out-of-sample classiffication errors, respectively are calculate with E() function.

```
Question 2: [a] The closest values among the following choices to the final (u,v) is (0.03, 0.08)
```

8. Question 3

I have added decay rate to the akgorithm and run again. The answer is similar to the previous tecnique.

```
Question 3: [d] The closest values among the following choices to the final (u,v) is (0.03, 0.08)
```

9. Question 4

I have increase the decay rate and errors increased simultaneously.

```
Question 4: [e] The closest values among the following choices to the final (u,v) is (0.57, 0.53)
```

10. Question 5

I have calculated E_{out} error with assingning k as all different integer values in the options of the question.

```
Question 5: [d] The value of k achieves the smallest out-of-sample classification error is: -1
```

11. Question 6

I have calculated E_{out} error with assingning k as all integer values from -20 to +20.

```
Question 6: [b] The closest value to the minimum out-of-sample classification error is 0.06
```

Part 4 – Neural Networks

12. Question 8

We have two layers and the first layer has 5 node and the second has 3. The result is calculated with counted products of the forms which are given in the question as operations.

```
Question 8: [d] Total number of operations required in a single iteration of backpropagation is 45
```

13. Question 9

In order to get minimum possible number of weights that such a network can have can be calculated as putting only 2 node including x_0 also in hidden layers. The result is 46.

```
Question 9: [a] Minimum possible number of weights that such a network can have 46
```

14. Question 10

I have started 3 layers and calculated each possible node numbers in the second and third layers. Luckily, I have found the answer with 3 layers. If I cannot founded, I may try other possible layer numbers from 4 to 18. Layers have 10, 23 and 13 nodes respectively.

```
Question 10: [e] Maximum possible number of weights that such a network can have is 510
```