UC San Diego Extension
Deep Learning Using TensorFlow

Winter 2019

Homework#5

Date Given: Feb 12, 2019 Due Date: Feb 18, 2019

==========

Problem#1

Dataset:

Generate a synthetic dataset using the following Python code.

```
-------------------------------------------------------------
RANDOM_SEED = 42
tf.set_random_seed(RANDOM_SEED)

import numpy as np
import matplotlib.pyplot as plt

n_samples = 30
train_x = np.linspace(0,20,n_samples)
train_y = 3.7 * train_x + 14 + 4 * np.random.randn(n_samples)

plt.plot(train_x, train_y,'o')
-------------------------------------------------------------
```

The value of RANDOM_SEED can be any integer. However, choose the value of RANDOM_SEED as 42. This will allow all students to get identical datasets. The 'train_x' is the predictor variable and 'train_y' is the response variable.

Model Building:

First, compute the regression equation between 'train_x' and 'train_y' variables using Scikit-Learn package. Next use TensorFlow software to compute the regression equation by minimizing the cost function. Use the following 'cost' and 'optimization' functions.

- $Cost\ Function = \sum_1^n (y_{computed} - y_{train})^2$
- $Optimization\ Function = Gradient\ Descent$

Iterate the optimization operation till the cost is minimized. Make sure that the regression equation computed by TensorFlow matches with the regression equation computed by Scikit-Learn. (Note: the value of the slope and intercept will not match exactly – it will be approximately equal)

During the iteration, print the value of 'cost', 'slope', and 'intercept' frequently. Make sure that the value of 'cost' variable is decreasing as the number of iterations are increasing. Since the synthetic data is generated using the slope value of 3.7, intercept of 14 and some random noise, your answer for slope and intercept values should be approximately 3.7 and 14.

Hyper Parameters:

There are 2 hyper parameters in this problem. The first one is learning rate (which varies from 0.1 to 0.00001), and other one is epochs (which varies from 10 to 100,000). You must vary both the hyper parameters to make sure that the slope and intercept values converge to a stable value.

Problem#2

Data Source

Analyze the data source in 'kc-house-data.csv' file. This data source is a part of databases available in the public domain. This file contains 21,613 observations of real-estate properties of King county in Washington state. The data for the following 21 variables is provided.

1. id
2. date
3. price
4. bedrooms
5. bathrooms
6. sqft_living
7. sqft_lot
8. floors
9. waterfront
10. view
11. condition
12. grade
13. sqft_above
14. sqft_basement
15. yr_built
16. yr_renovated
17. zipcode
18. latitude
19. longitude
20. sqft_living15
21. sqft_lot15

Response Variable: price

Predictor Variables:

1. bedrooms
2. sqft_living

Compute the regression equation between the predictor variables and response variables using Scikit-Learn package. Next use TensorFlow software to compute the regression equation by minimizing the cost function. Use the following 'cost' and 'optimization' functions.

- $Cost\ Function = \sum_1^n \left(y_{computed} - y_{train}\right)^2$
- $Optimization\ Function = Gradient\ Descent$

Make sure that the regression equation computed by TensorFlow matches with the regression equation computed by Scikit-Learn.

If your Neural Network doesn't converge, scale all the predictor variables and response variable between 0 – 1.