ECE 521 Assignment 2

# Work Breakdown

|  |  |
| --- | --- |
| Group Member Name | Contribution Percentage |
| Jixong Deng | 33% |
| Jeffrey Kirman | 33% |
| Connor Smith | 33% |

# Part 1: Linear regression

Linear regression is performed using the following minibatch stochastic gradient descent (SGD) algorithm:

|  |
| --- |
| **Init:** **w** = **0** # Weights vector  **xj** = [1 x­j1 xj2 ... xjd]T # Single in feature  **X** = [**x1** ... **xn**] # Matrix of all in-features  **y** = [**y1** ... **yn**] # Vector of all targets  **Step:** for each epoch  Shuffle **X -> X’** # Randomly shuffle the vectors in **X** into **X’**  for each mini-batch: **X’** = [**x’(mini-batch\_size)\*i** ... **x’(mini-batch\_size)\*(i+1)**]  **gt** = ∇Ein(**w(t)**) # Calculate the gradient of the loss  **w(t+1)** = **w(t)** – η**gt** # Update the weights, where η is the learning factor  **Result:** return **w** |

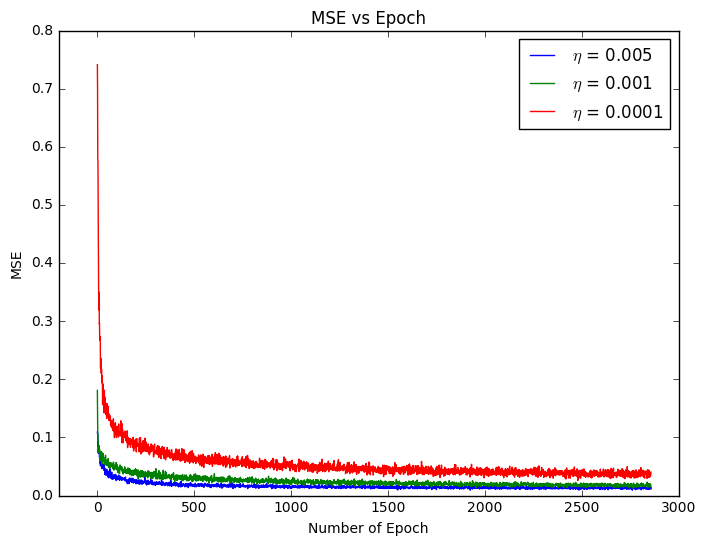
The **SGD** function (**linear\_regression2.py** in *Appendix A*) implements this. This function creates a graph in which with placeholders for input matrix **X** and target vector **y**. It then calculates the MSE loss according to the following equation:

Where λ is the decay constant. The gradients are then calculated using a function in the tensorflow API, tf.gradients. Finally, the weights are updated.

After the graph is created the function calculates the number of epochs based on the iterations and mini-batch size. It then shuffles an array of indices and then uses them to index the matrix of in-features into a mini-batch. feed\_dict is then used to feed the inputs into the placeholders and run the graph for one iteration.

## Part 1 – 1

The results of changing the learning rate (η) can be seen in the figure below. Here we can see that as the learning rate is increased the quicker it converges and the lower final value it converges to. Therefore, **the best η in this case is 0.005.**



## Part 1 – 2: Effect of the mini-batch size

The final values of the MSE loss as well as computation time are summarized in the table below:

|  |  |  |
| --- | --- | --- |
| **Mini-batch Size** | **MSE loss** | **Computation time (seconds)** |
| 500 | 0.0149 | 206 |
| 1500 | 0.0176 | 433 |
| 3500 | 0.0171 | 1243 |

The best mini-batch size in terms of training time is 500. This makes sense since each iteration contains less operations (i.e. less MSE loss and less gradients to compute), hence each update will happen faster.

*Note: for mini-batches that do not equally divide the number of training examples, the remainder of unused examples are used in a shortened mini-batch at the end of each epoch.*

## Part 1 – 3: Generalization

The final values of the accuracy on the validation and test set, as well as computation time are summarized in the table below. When calculating accuracy, if a prediction was above 0.5 it was said to be classified as 1, and below 0.5, classified as 0.

|  |  |  |  |
| --- | --- | --- | --- |
| **λ** | **Validation Accuracy** | **Test Accuracy** | **Computation time (seconds)** |
| **0** | **0.972** | **0.99** | **209** |
| 0.001 | 0.959 | 0.99 | 210 |
| 0.1 | 0.952 | 0.95 | 211 |
| 1 | 0.938 | 0.94 | 211 |

The best decay coefficient in terms of validation accuracy is λ = 0. Here the validation accuracy is 0.972 and the test accuracy is 0.99. Using weight decay can help the model by regularizing the data. In our case, however, the model was quite accurate with 0 decay coefficient that it did not help much. We need to choose the hyper-parameter λ using the validation set instead of the training set as to not overfit the data. It gives a tunable parameter to change how the model predicts values outside of the training set.

## Part 1 – 4: Comparing SGD with the normal equation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MSE loss** | **Validation Accuracy** | **Test Accuracy** | **Computation Time** |
| **SGD** | 0.0121 | 0.966 | 1.0 |  |
| **Normal Equation** | 0.0104 | 0.959 | 0.98 |  |

# Appendix A – Python code

## Linear\_regression2.py

|  |
| --- |
|  |