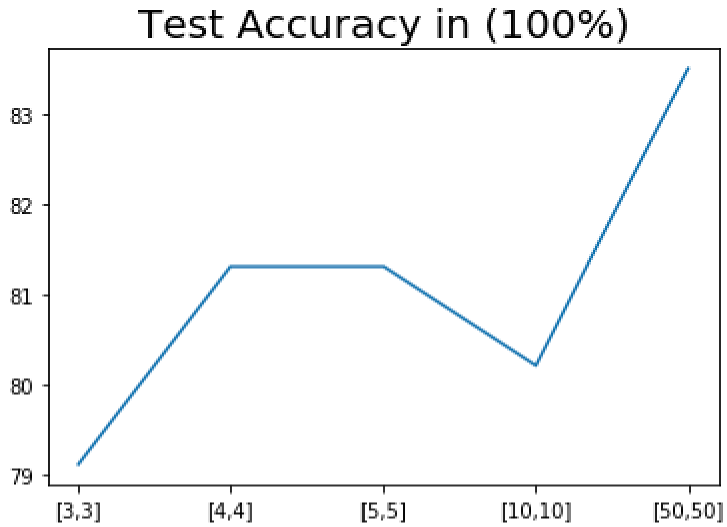
Deep Learning Homework 1

Student ID: 0416106 **Name**: 彭敬樺

1. Below is the accuracy for testing data with the provided hidden layer

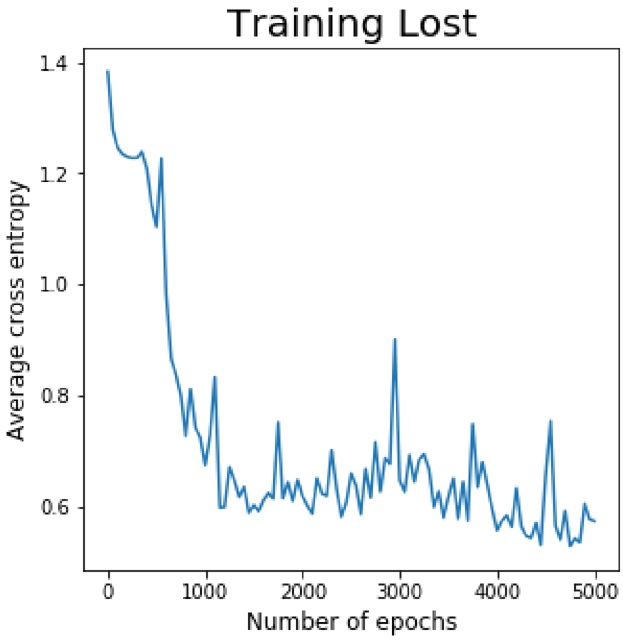


Filename: Number 1.ipynb

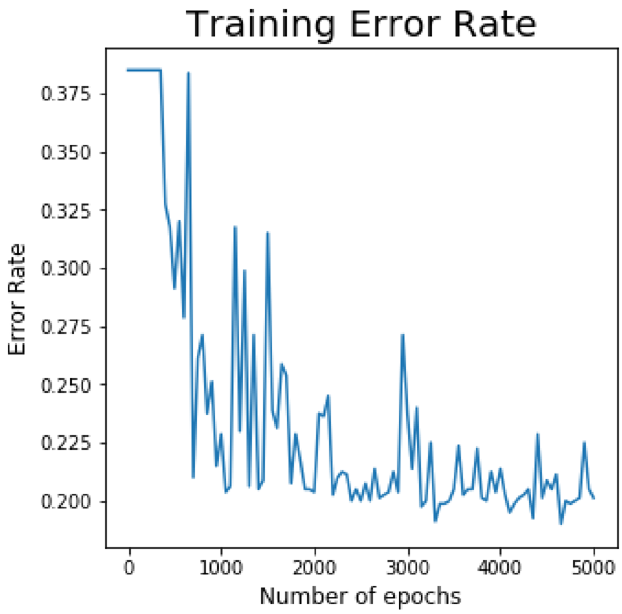
It can be seen that in the configuration where there are 2 layer of hidden layer, with 4 nodes in each layer correspond to the best trade off between computation complexity and accuracy of the result. Therefore, in this problem, DNN in constructed with [6, 4, 4, 2] corresponding to those of input layer, first and second hidden layer, and output layer.

By testing, the best minibatch is approximately 64 and 128. In this model, 128 mini batch is chosen, and the model is trained for 5000 epoch. It can be seen in thee graph, that we do not need 5000 epoch to gain the advantage of accuracy, the model could converge in around 3000 epoch, with learning rate of 0.01

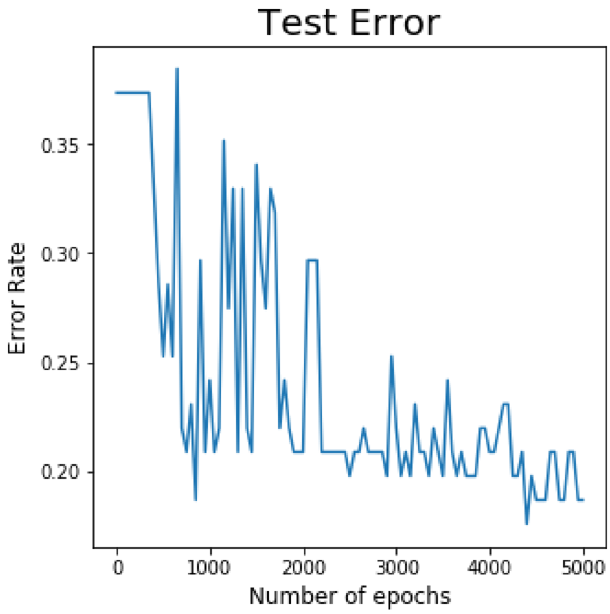
a. Graph of Learning Curve



b. Graph of Training Error Rate

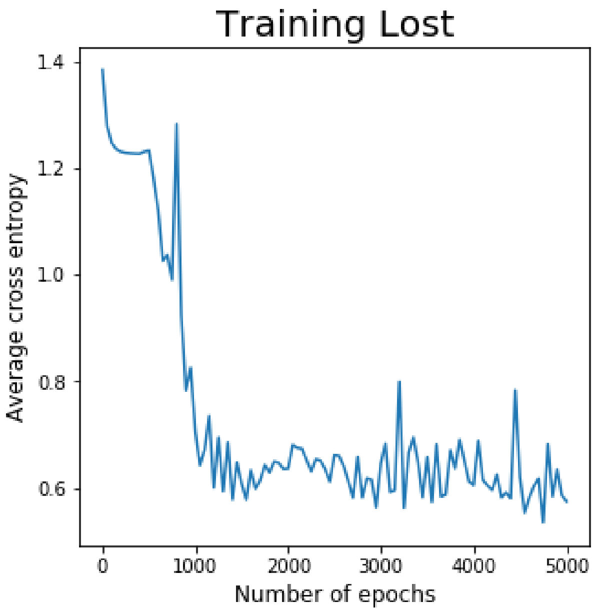


c. Graph of Test Error Rate

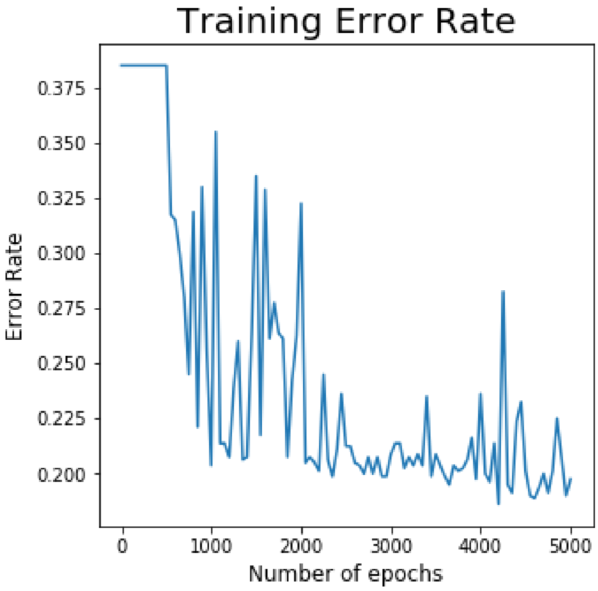


2. The parameter of model used in this problem is the same as (1) with the only difference of [6, 3, 3, 2] layer choice
Filename: Number 2.ipynb

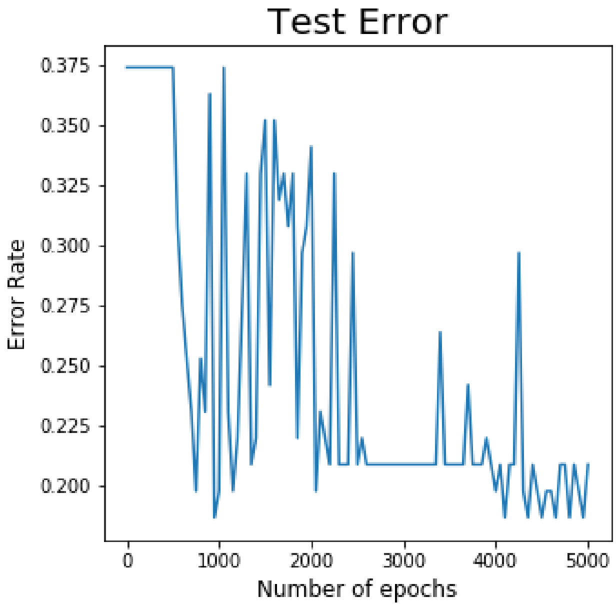
a. Graph of Learning Curve



b. Graph of Training Error Rate

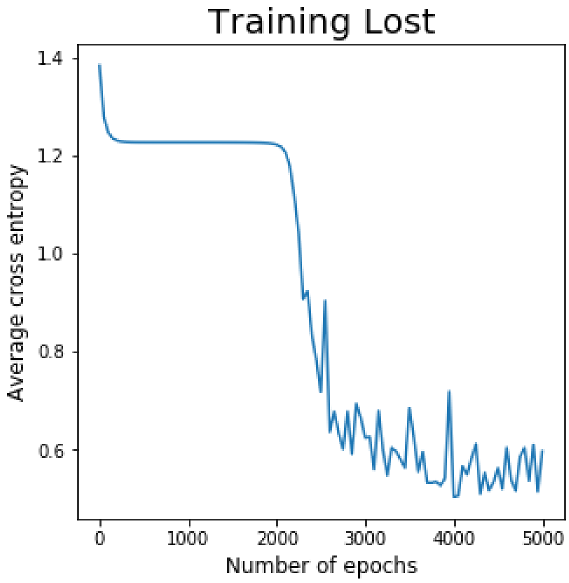


c. Graph of Test Error Rate

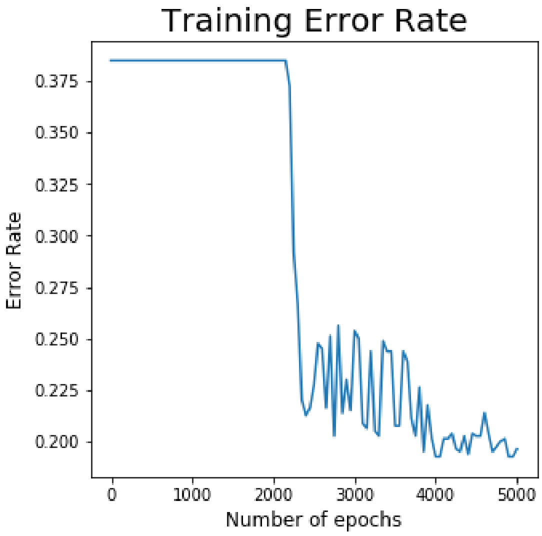


3. The implementation of the normalization done on the Fare feature can be seen in the code.
Filename: Number 1_normalized.ipynb and Number 2_normalized.ipynb
The normalized fare feature does not help with the accuracy of this model between the determined number of iteration, however, it can be seen that both of the error rate has become more stable and rarely overshoot. This happened due to the stabilized value of the fare value, which initially have large range between its maximum value and minimum value. Using the constant learning rate, this large range makes the model hard to approach the convergence if the learning rate is top small, and will overshoot if the learning rate is set too large. Normalization will help to solve that problem, and diminish outlier in the data.
Using the model used in question 1

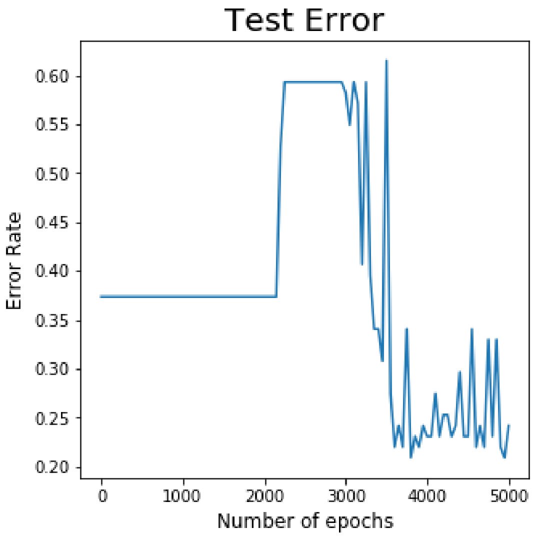
a. Graph of Learning Curve



b. Graph of Training Error Rate

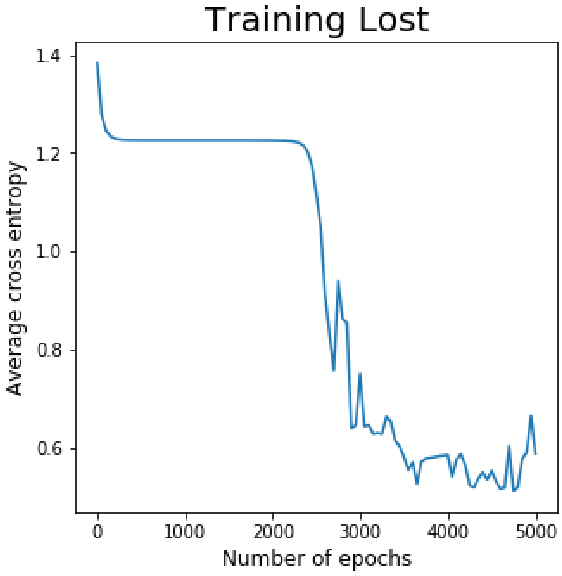


c. Graph of Test Error Rate

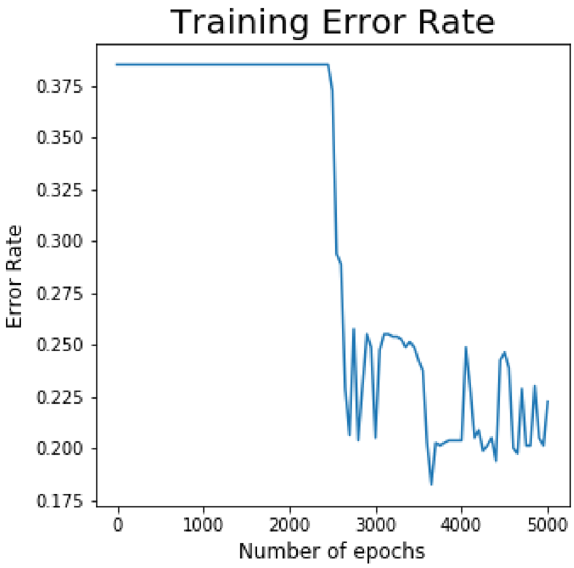


Using the model used in question 2:

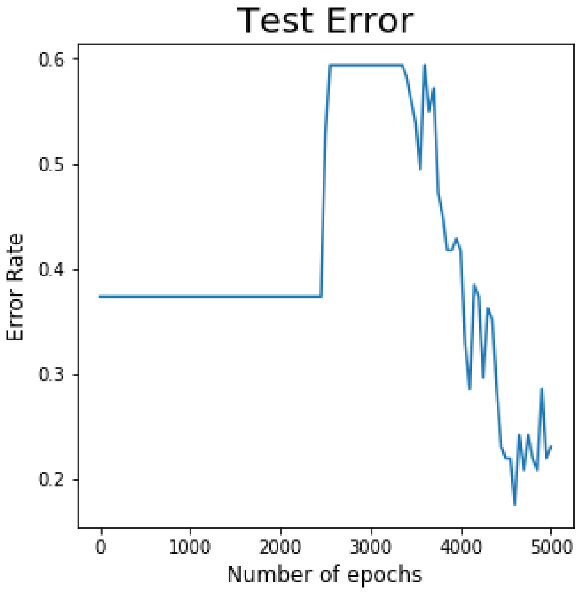
a. Graph of Learning Curve



b. Graph of Training Error Rate



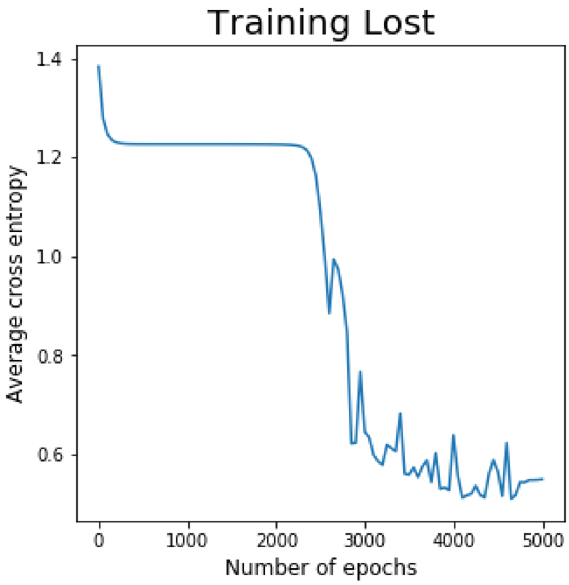
c. Graph of Test Error Rate



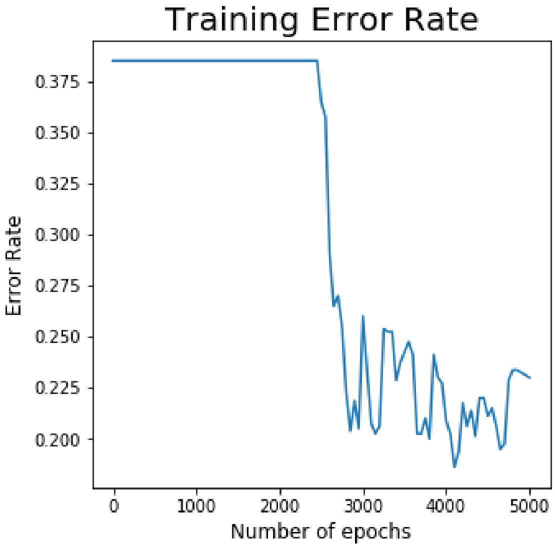
Other than the ‘Fare’ feature, we could also normalize the value of ‘Age’. Other column is not really suitable to be normalized due to its trait of being categorical data. Categorical data in this dataset is represented using integer and span through small range of integer value.

Below will be shown the graph gained from normalizing ‘Fare’ feature and ‘Age’ feature in model used in question 2:

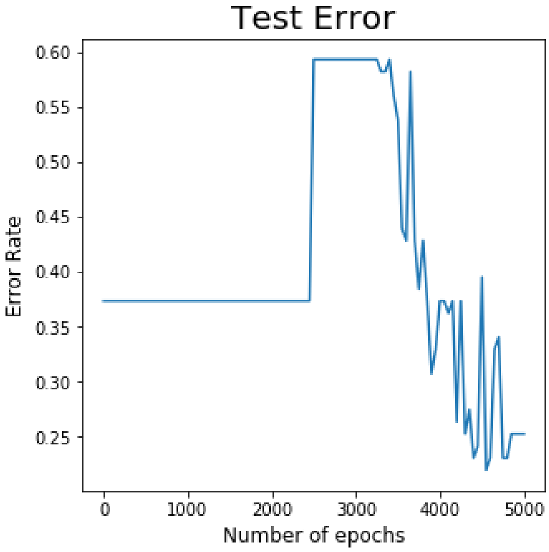
a. Graph of Learning Curve



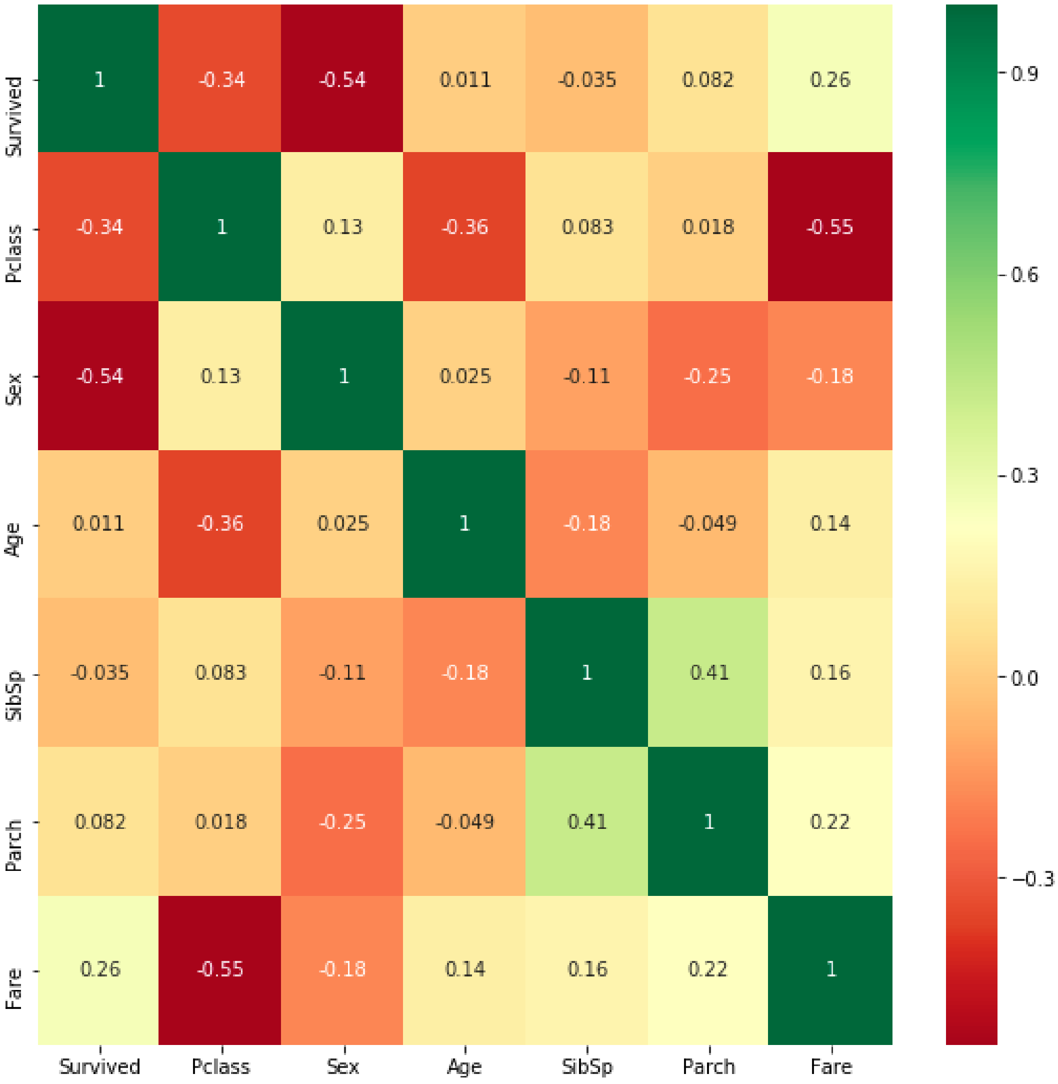
b. Graph of Training Error Rate



c. Graph of Test Error Rate



4. Using the heatmap of correlation value between feature, we can determine feature that affect the model most significantly. It can be deduced from this map, that 'Sex' feature affect the model the most.

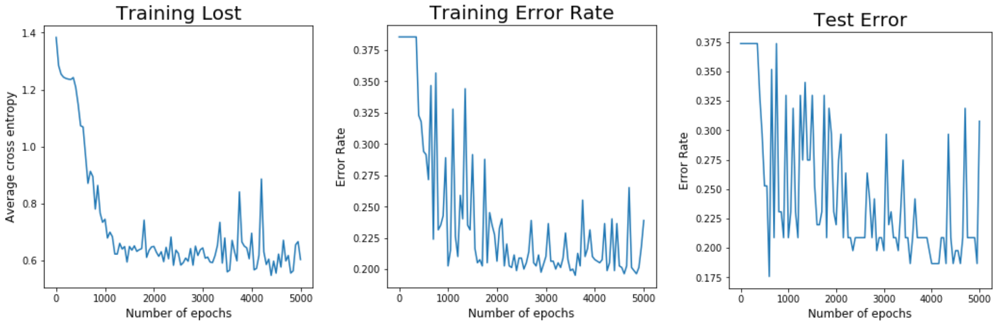


In this question, I also tried to remove each feature and graphing its 3 graph value, in this report it will be only shown the feature removal of the one that has correlation above 0.1 with the target output due to its small difference that is cause in the model after removal, however, complete graph can be seen in the code.

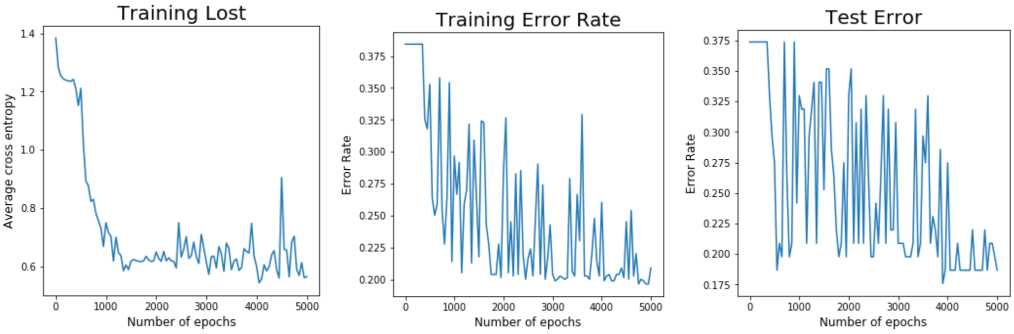
Filename: Number 1_feature.ipynb and Number 2_feature.ipynb

Model Used in Question 1

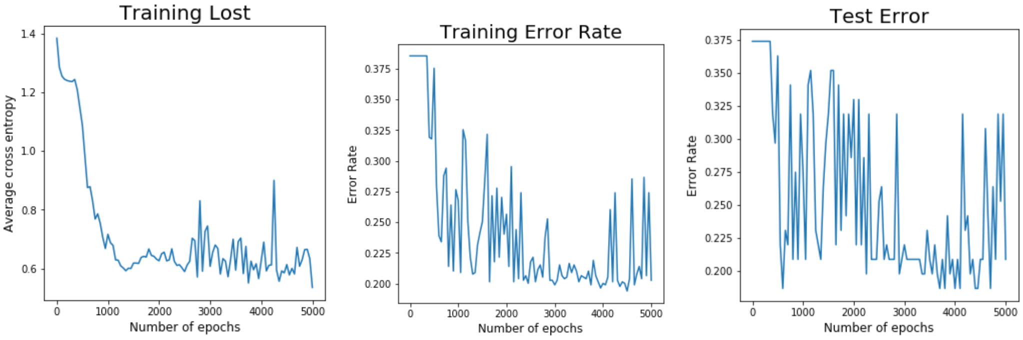
DELETING PCLASS



DELETING SEX

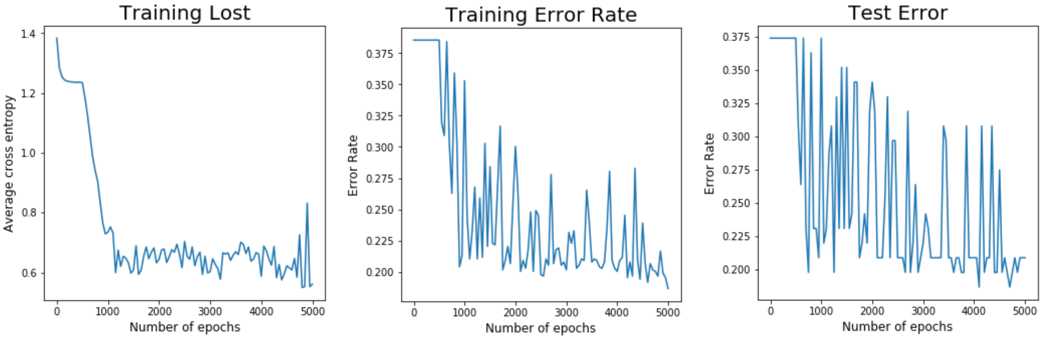


DELETING FARE

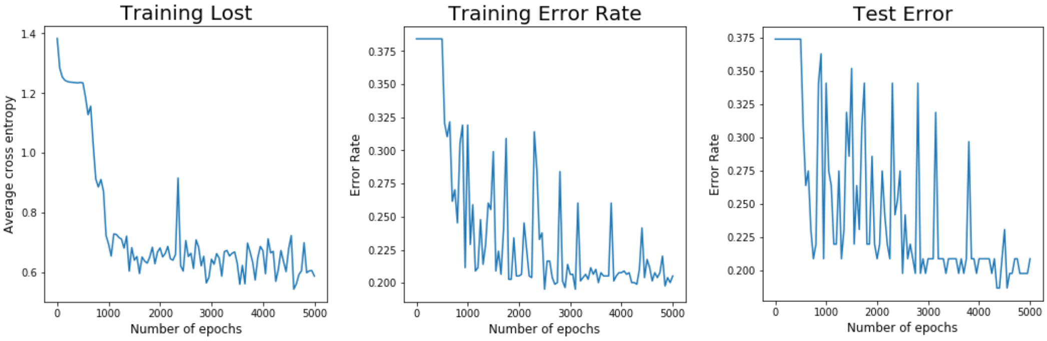


Model Used in Question 2

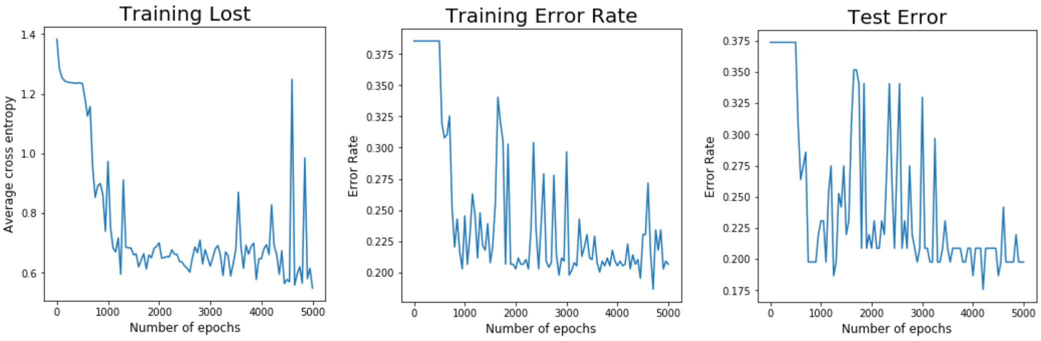
DELETING PCLASS



DELETING SEX



DELETING FARE



In the figure above, we can see that the 3 feature above affect the accuracy of the model. It throws havoc the consistency of testing error rate in the original model.

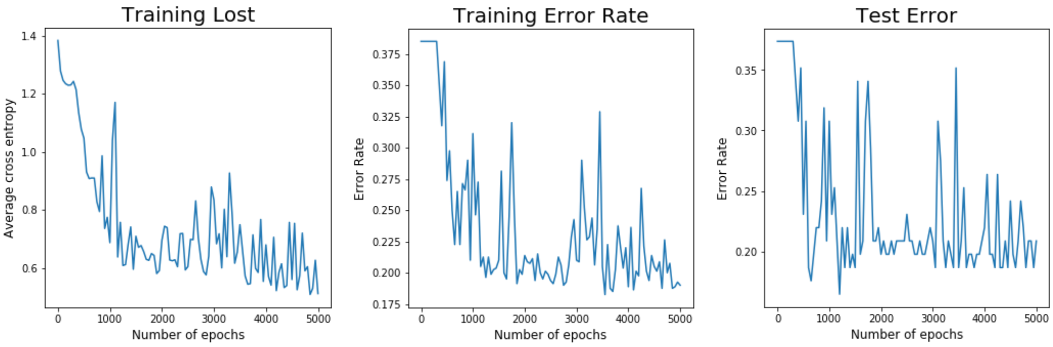
5. A major advantage of the one-hot encoding is that we can use a softmax layer as output which gives a nice interpretation of the output as a probability distribution over the classes. This means that the neurons can express uncertainty and we can randomly sample from the distribution according to the predicted probabilities and explore alternative predictions this way. We can think of a neural network with softmax output layer basically as a categorical probability distribution that is parameterized by the synaptic weights of the neural network.

As an alternative, we could consider encoding the class label by partitioning the value range of an output neuron. **A major problem with that is, however, that the network could not express uncertainty as naturally as with a one-hot encoding.** Suppose the network predicts class 1 and class 3 with 50% probability each. Then the output would be a point in the interval for class 2 which is possibly completely unrelated to the other two classes.

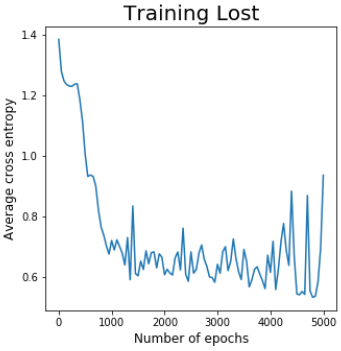
In the case of ticket class, the 3 class has correlation with each other, the one in class 2 has different service and ticket fee compared to the one in class 1 and 3. Class 2 is better than class 2 but worse than class 3. This integer representation could be exploited to represent the gauge of class, in another word, the quality of service and possibly the safety measure and gear provided nearby for higher class individual.

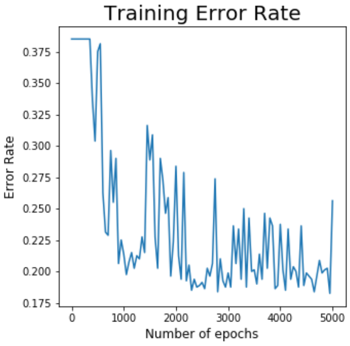
Filename: Number 1_onehot.ipynb and Number 2_onehot.ipynb

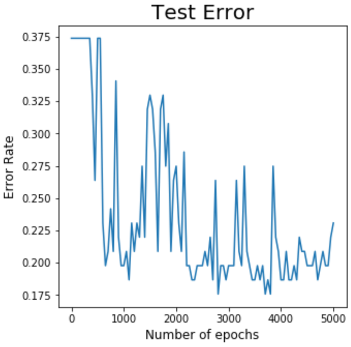
For model used in question 1



For model used in question 2







In the case of this problem, the one hot encoding of class feature does not help the model learn in significant matter. This is not desirable cause accuracy does not increase significantly while the output layer has increased nodes, which means increase in the computation complexity.

6. Provided sample:
Pclass=3, sex=1, age=25, sibsp=2, parch=2, fare=10 -> not survive (for both model)
New sample:
Pclass=1, sex=1, age=50, sibsp=2, parch=2, fare=50 -> not survive (for both model)
Pclass=1, sex=0, age=50, sibsp=2, parch=2, fare=50 -> survive (for both model)

This sample is chosen to generate one survived individual and one not survived individual, the only difference in the input in the value of sex, which has been seen from the previous answer has the largest correlation with the output. The change in fare accompanied by the Pclass would also change the output result due to its high correlation with each another, and moderately good correlation with the output result.