Deep Learning (Homework 1)

Due date : 04/12/2019

- Any tools for automatic differentiation are forbidden in this homework, such as Tensorflow, Pytorch, Keras, MXNet, et cetera. You should implement the backpropagation by yourself.

- Submitting Homework– Please zip each of your source code and report into a single compress file and name the file using this format : HW1_StudentID_StudentName.zip (rar, 7z, tar.gz, . . . etc are all not acceptable)

Deep Neural Network for Classification

In this exercise, please implement a Deep Neural Network (DNN) model to predict if one can survive in the Titanic tragedy. This dataset is collected from a Kaggle competition and cleaned by TAs. The features include: sex, age of a person; how many parents or children, and how many siblings or spouses are on board related to the person; the ticket class and the passenger fare are also included. Please create a DNN model inputs these features and outputs the survival of passengers. The details of this dataset can be referred to: `https://www.kaggle.com/c/titanic/data`. The cleaned data is provided in titanic.csv, please take the first 800 rows as training data and the last 91 rows as test data. The first column, the survival, should be treated as label, while the other six columns are features. In the sex column, 1 indicates male while 0 indicates female. In the age column, the empty entries are filled with zeros.

1. Please construct a DNN for binary classification. For N samples and K categories, the cross entropy objective function is expressed by

$$E(\mathbf{w}) = - \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_k(\mathbf{x}_n, \mathbf{w}).$$

Please minimize the objective function $E(\mathbf{w})$ by error backpropagation algorithm using the Stochastic Gradient Descent (SGD) where the parameter updating in each mini-batch is yielded by
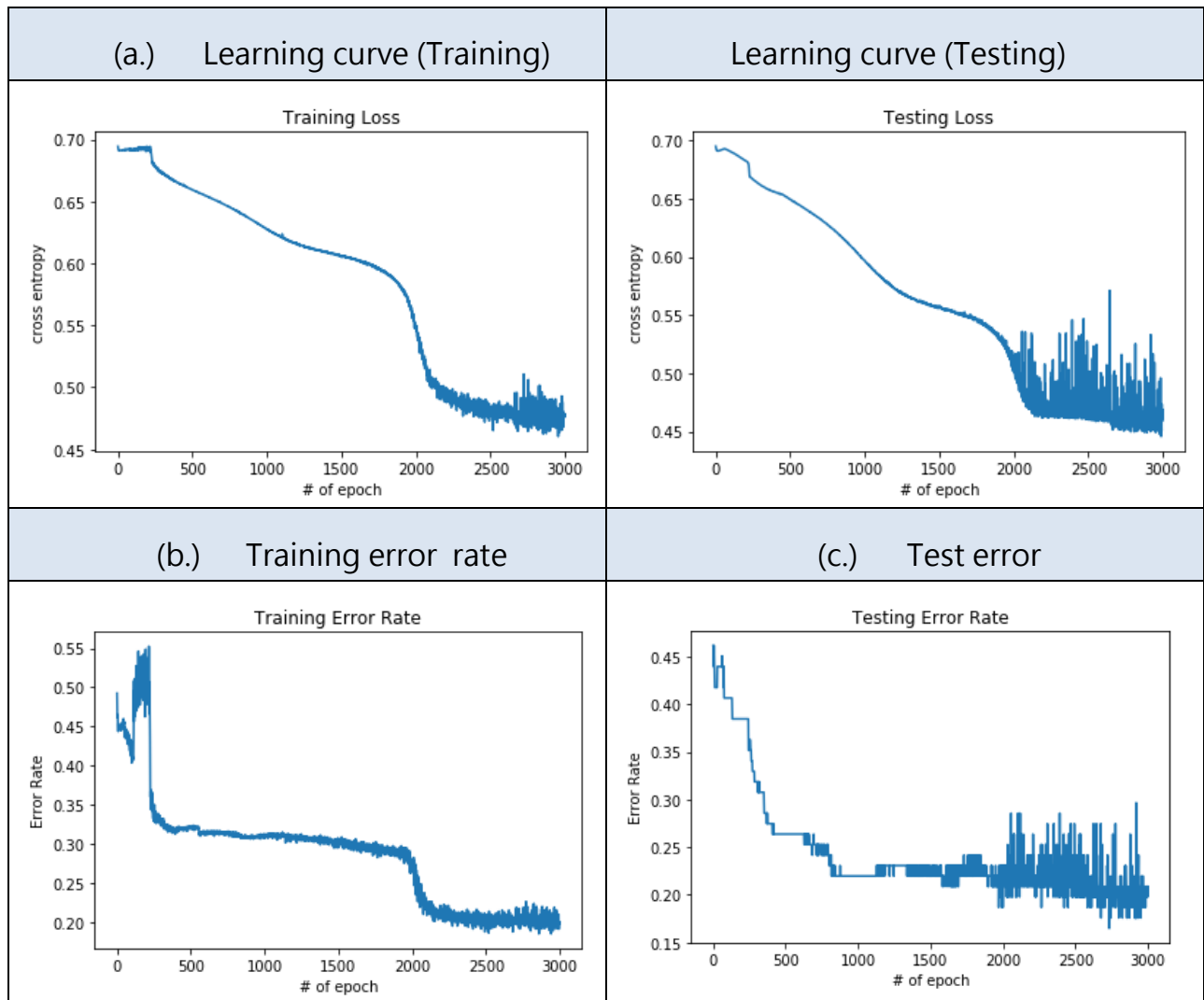
$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)}).$$

You should decide the following variables: number of hidden layers, number of hidden units for each layer, learning rate, number of iterations and mini-batch size. Please show (a) learning curve, (b) training error rate and (c) test error rate in the report. The network architecture should be designed by yourself. Please describe the architecture of your model and explain why you design or choose such an architecture.

2. Please construct a DNN with a specified architecture. The number of units in layers should be arranged as [6 3 3 2], corresponding to those of input layer, first and second hidden layers, and output layer. This network has the same inputs and outputs as provided.

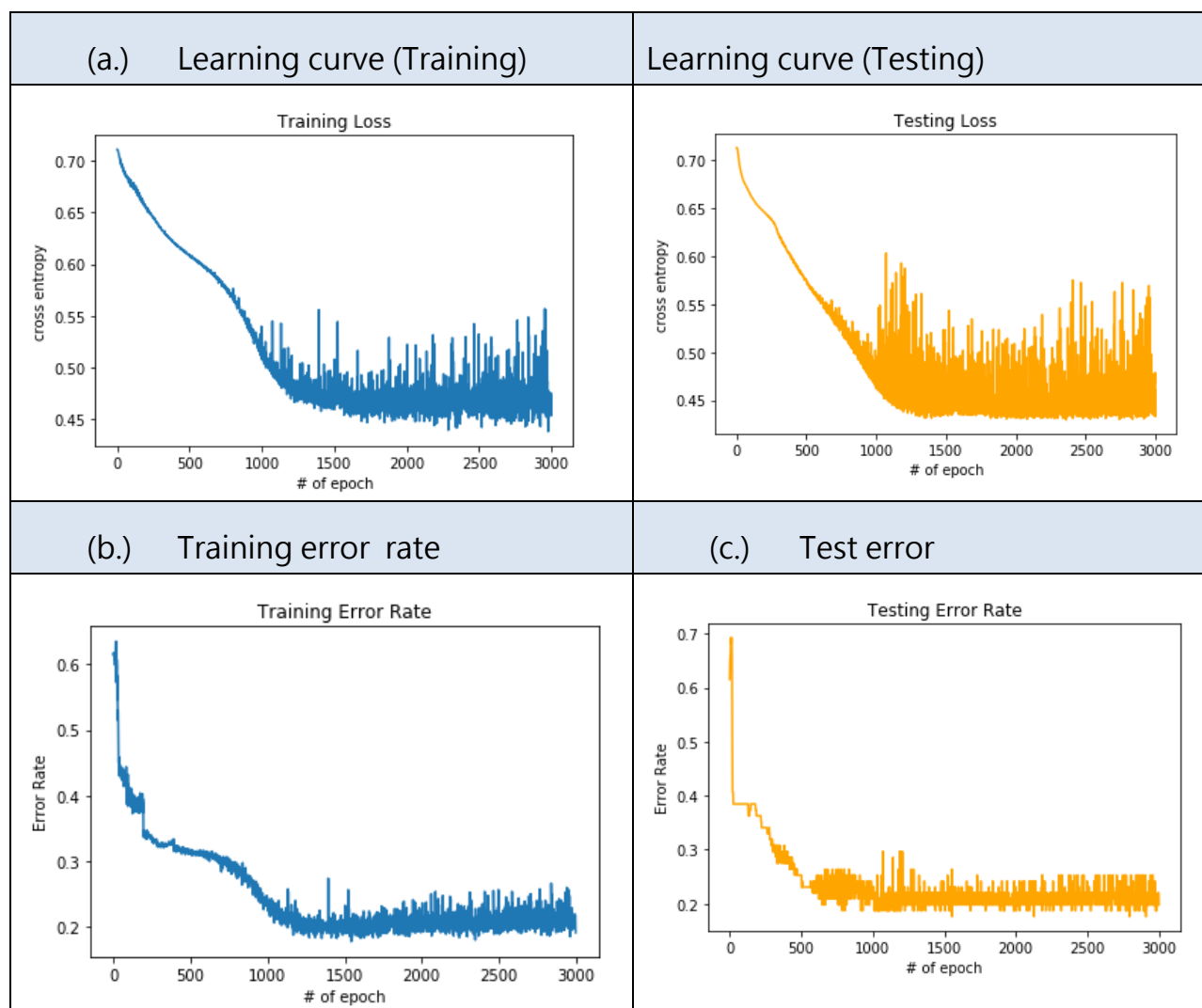當網路參數各層使用第二題的規定[6,3,3,2]，出來的答案如下圖：

(a.) Learning curve (Training)	Learning curve (Testing)
Training Loss	Testing Loss
(b.) Training error rate	(c.) Test error
Training Error Rate	Testing Error Rate

關於網路設計，因為第二題網路規格要求以[6,3,3,2]來呈現，因此直接以此為實驗，在 Learning rate為0.005，batchsize = 40，epoch = 3000，在這樣情況下，最後一層分類結果使用softmax。可以從上圖觀察到震盪非常大，認為原因是在資料部分，Fare的欄位未做過Normalize，其數值的值域從個位數到三位數都有，因此大幅影響結果，將在第三題的時候印證假設。

假設將網路加寬，參數量增加的情況下，欲知道網路的表現是否更好，因此將網路參數設為[6,8,4,1] 在上述其他條件不變的情況下，網路表現如下表所示：

(a.) Learning curve (Training)	Learning curve (Testing)
(b.) Training error rate	(c.) Test error

看了Testing loss 的部分，到後半部**震盪變大**，但是相較前一個初代版本，**這個版本的 Error rate較為穩定**一些 。另外關於在初代版本當中 Testing Error rate 一開始就較Training Error rate低的原因可能有兩個，一是Testing 的資料分佈可能和Training 相近，第二種可能是因為Testing 筆數過少，而Training 的筆數相較多，因此Testing 的資料較容易被猜中。

觀察圖表，看到其震盪劇烈，推測可能原因是learning rate 太大，導致更新的時候每次更新步伐太大而不太穩定，因此嘗試將**Learning rate 改小至0.0003**，做出來以後，單看Testing 的部分有比較好的收斂效果，對Training 整體也比較沒有那麼震盪，結果如下所示：

(a.) Learning curve (Training)	Learning curve (Testing)
(b.) Training error rate	(c.) Test error
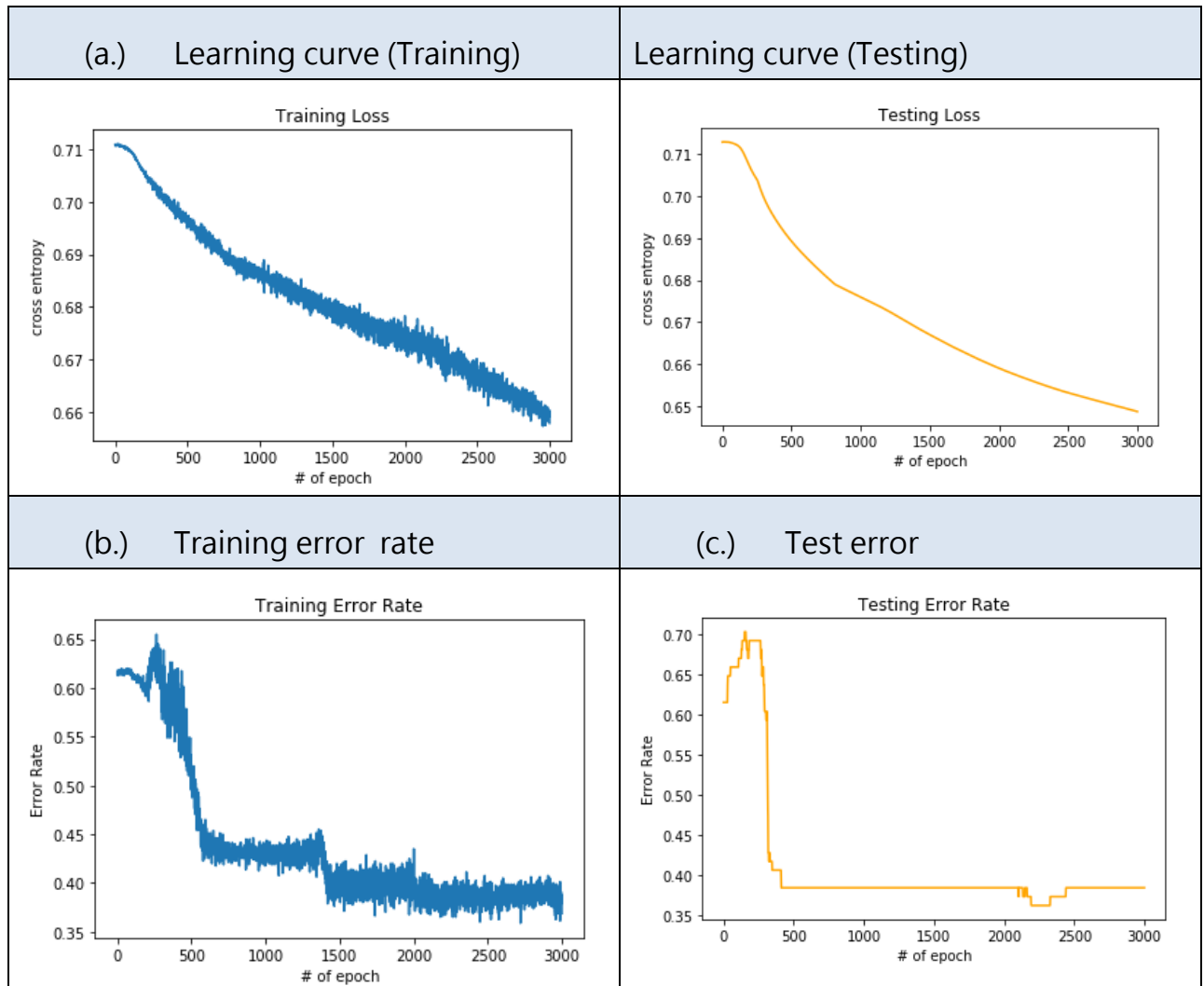
這次作業要求中，為了讓所有修課同學的train 和 test 切割統一，而直接設成前800筆和後91筆，但是在做shuffle的時候，認為應該先shuffle再切割成train跟test比較合理一些。

3. Look into the features. Please perform normalization for the Fare feature and compare the error rate with the original network. For normalization process, you may refer to `https://en.wikipedia.org/wiki/Standard_score`. Are there any other features that need to be normalized? Please answer in the report.

接續上一題的限制將網路結構設定為 [6,3,3,2]，且將其他設定更改回 Learning rate = 0.005，batch size = 40，epoch = 3000，實驗結果相比較以後，認為有做正規化的資料的確有讓網路更加穩定，因為在Fare這個欄位的值域差距非常大，導致Fare這項Feature的值會大幅影響預測結果，其實驗結果如下所示：

	Training	Testing
LOSS		
Error Rate		
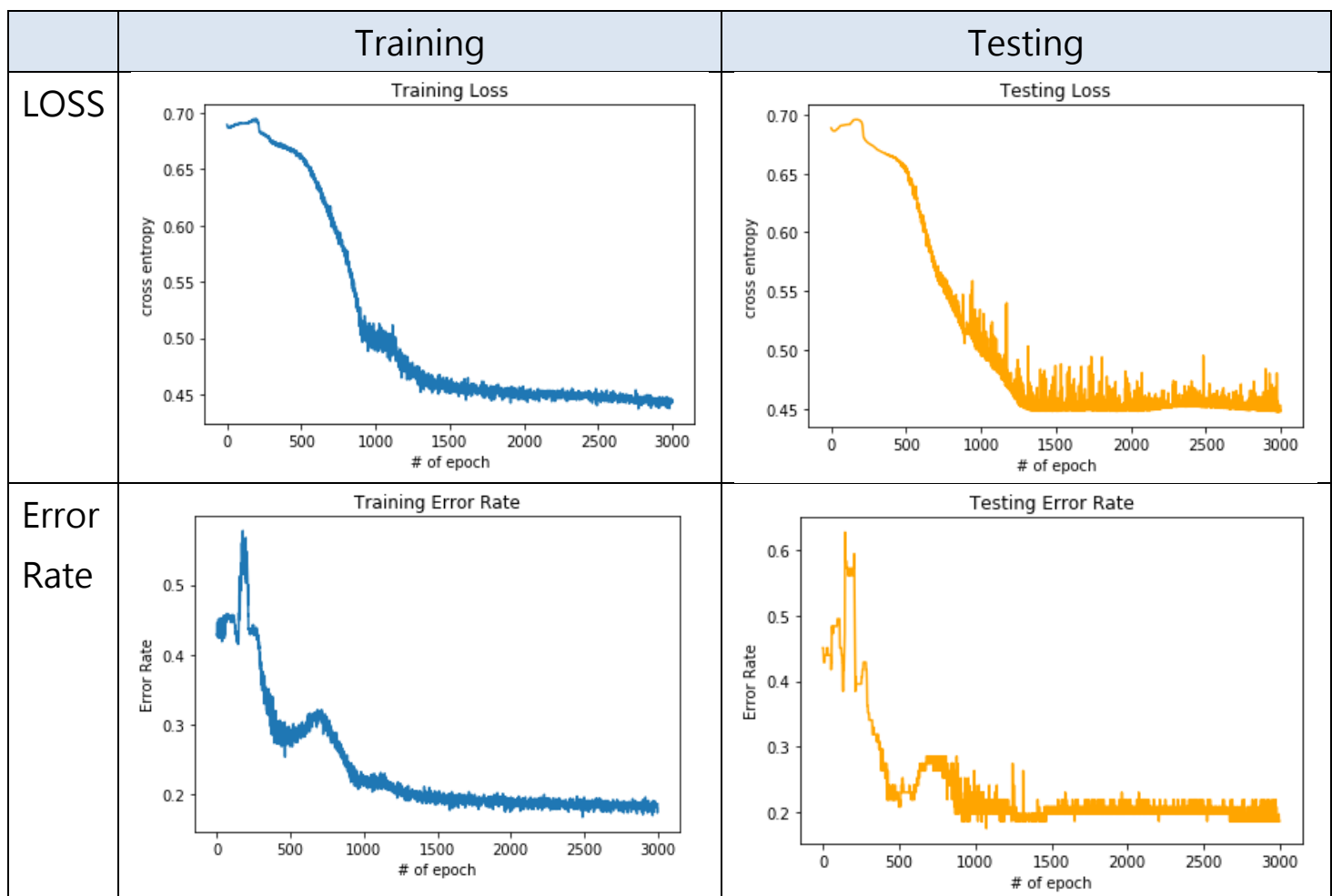
4. Please identify which feature affects the prediction performance the most. Describe the process how you solve this problem in the report.

判斷哪個Feature較有影響力的方法：

1. 先將Fare的部分做正規化處理

2. 將網路架構更改為[5, 3, 3, 2]

3. 觀察缺失了哪項，會讓ACC最低，則該

   項為最有影響力的一項

根據右圖實驗結果顯示，缺少了 **Sex**

對於整體預測結果影響最大，因此判斷這兩

項Feature 是對於Model預測最有影響力的

Feature。

```
0
The Network's structure: [ 5 - 3 - 3 - 2 ]
without Pclass , Acc = 0.813187
1
The Network's structure: [ 5 - 3 - 3 - 2 ]
without Sex , Acc = 0.703297
2
The Network's structure: [ 5 - 3 - 3 - 2 ]
without Age , Acc = 0.802198
3
The Network's structure: [ 5 - 3 - 3 - 2 ]
without SibSp , Acc = 0.802198
4
The Network's structure: [ 5 - 3 - 3 - 2 ]
without Parch , Acc = 0.802198
5
The Network's structure: [ 5 - 3 - 3 - 2 ]
without Fare , Acc = 0.802198
```

5. When the feature is categorical with more than three categories, we often transform the categories into one-hot vectors. Should we perform the same process to the feature of ticket class? Please show your explanations and illustrations in the report.

　　本題欲探討若將Ticket class改為 categorical 餵入網路以後是否能提高準確率，因此跑了兩次實驗，**第一次將資料做categorical的轉換，接著將資料跑過非Categorical的網路，之後查看其準確率**，根據數據顯示，就**結果論而言沒有什麼太大的差別**，但是**依照Loss的收斂效果**來看，對於Training 跟 testing **做Categorical蠻有幫助**的，震盪變小，比較穩定一點，結果如下所示：

一、比對ACC： `Acc with categorical = 0.813187`　`Acc without categorical = 0.813187`

二、觀察圖表：

Categorical	Without Categorical
Training Loss	Training Loss
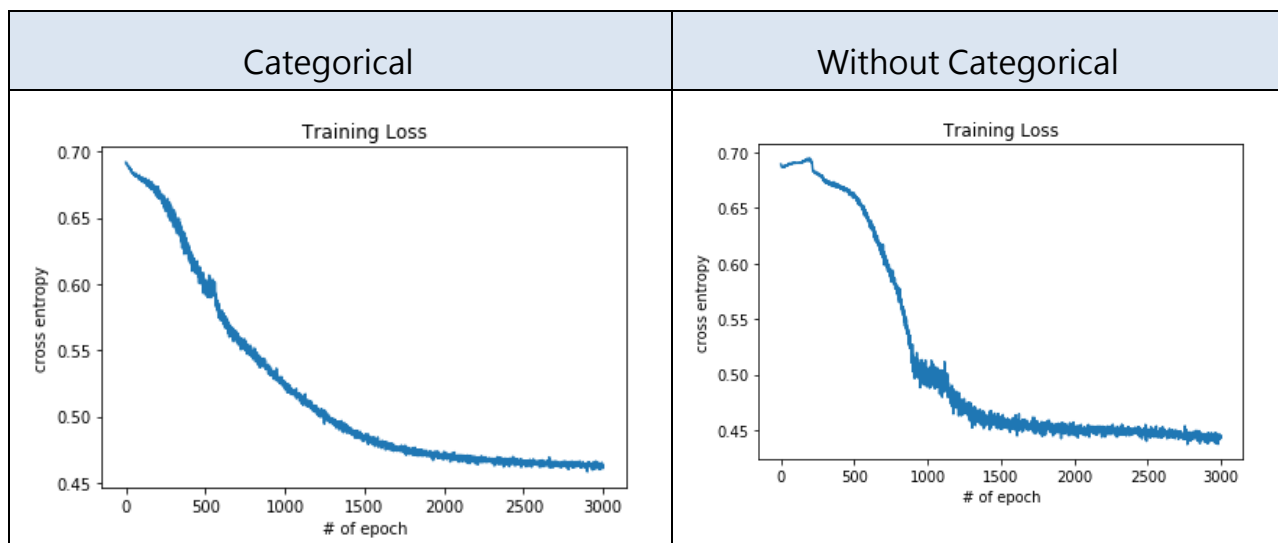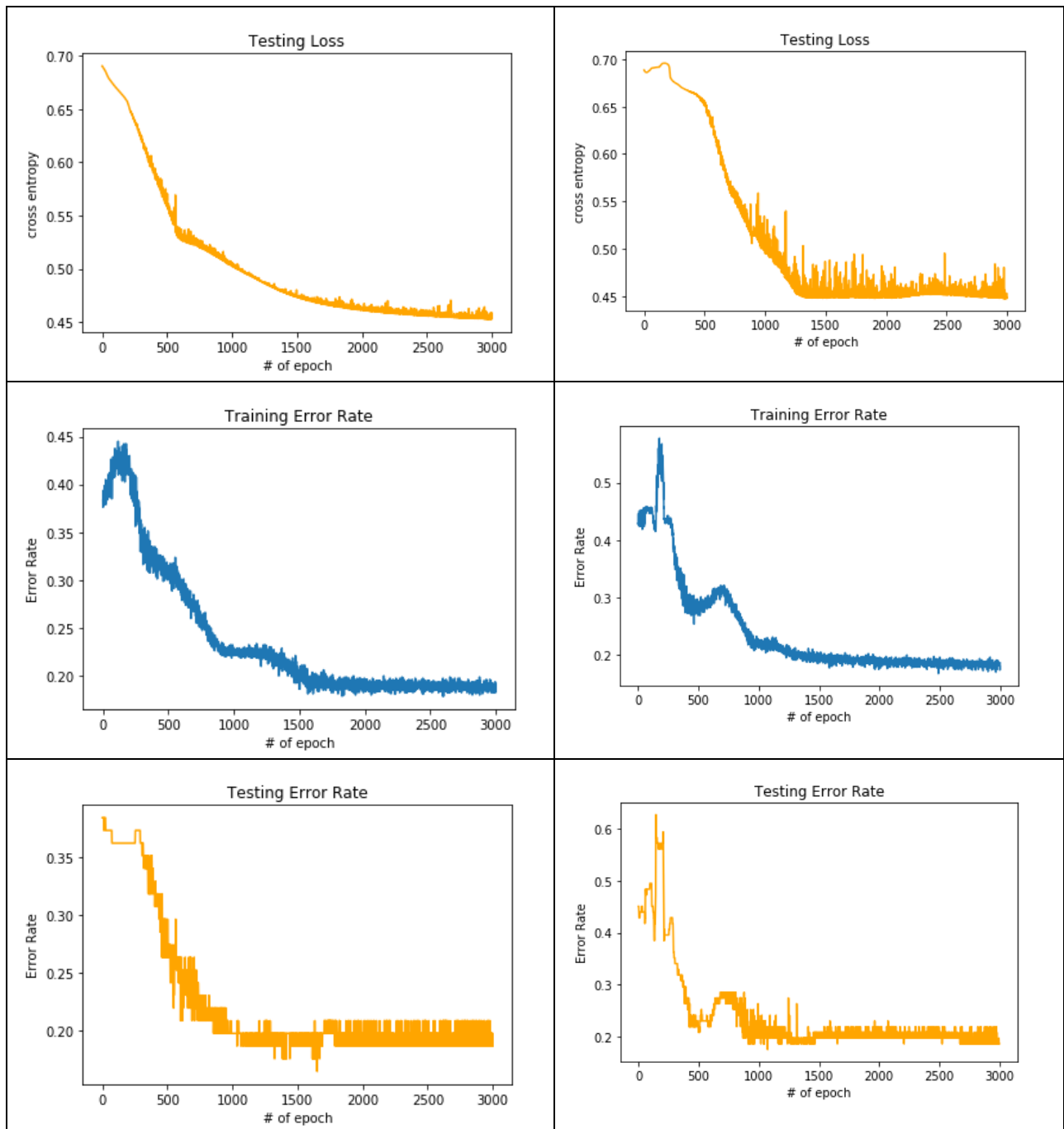
6. Please artificially design two new samples that one can survive and the other can not. For example, a person with the following feature: pclass=3, sex=1, age=25, sibsp=2, parch=2, fare=10. Input this new sample to the model and see the result is survived or not. Explain how you choose these samples to achieve correction classification.

在這題的題目當中生與死的資料和結果如下所示：

```
The Network's structure: [ 6 - 3 - 3 - 2 ]
Test data before normalize:
['Survived', 'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
survived:[3, 0, 23, 5, 0, 99]
dead   :[3, 1, 23, 5, 0, 99]

Predict of survived:yes，dead:yes
```

根據之前使用鐵達尼號資料集的經驗，存活者大部分是女性、老人及年幼的小孩，因此當所有條件都不變，只改變性別這個變數的話，女生會活著而年輕男性會死亡。