Deep Learning HW1 Report

NTHU Wei-Cheng Tseng A072045

Problem 1

- Grid Search

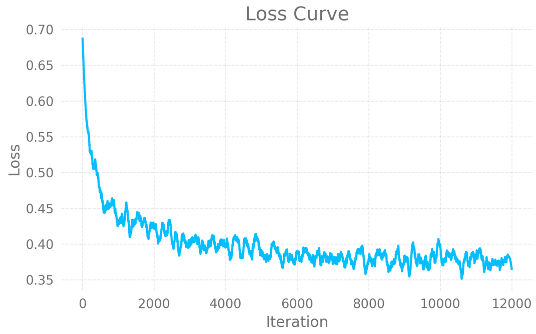
Since there are too many hyperparameters to search. We apply grid search to get relative better hyperparameters. Our hyperparameters are shown as below.

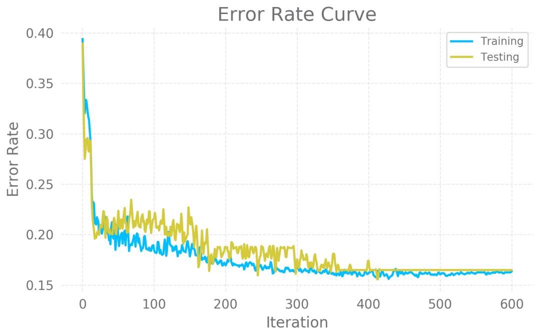
- learning rate (lr): The initial learning rate of the optimizers. Our search range is [0.1, 0.0372, 0.0138, 0.0051, 0.0019, 0.00071, 0.00026, 0.0001].
- learning rate decay rate (lr_dec): We apply exponential decay on the learning every epoch, i.e., $lr_{new} = lr_{original} * lr_{decay\ rate}$ .Our search range is [0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 1]
- batch size (bs): The size of a mini-batch during the training. Our search range is [5, 28, 52, 76, 100].
- weight scale (ws): We initialize the weight with a normal distribution. The normal distribution has mean 0 and variance **ws.** Our search range is [1, 0.1, 0.01, 0.001]

For each of the following experiment, we run the grid search to find the best hyperparameters combination. Note that the search range may be slightly different for each experiments.

- Network Architecture

- layers: [6, 16, 16, 8, 4, 2]
- learning rate: 0.005
- learning rate decay rate: 0.99
- regularization weight: 1e-8
- weight scale: 0.05
- batch size: 40
- The number of epoch: 600 (12000 training iterations)

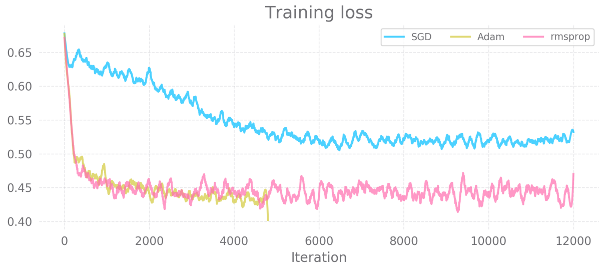


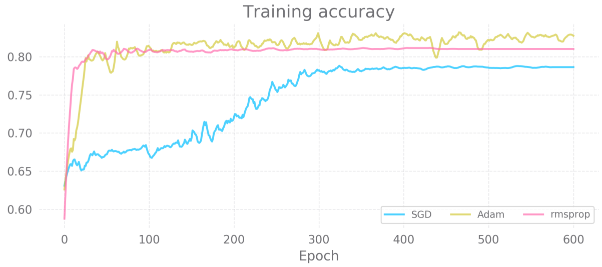


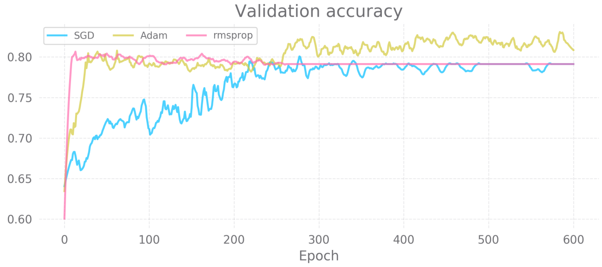
In general, the deep structure has better ability than shallow structure. However, since my computational power is limited, my CPU can afford no more than 6 layers. Therefore, we construct a network with 6 layers. Other hyperparameters are found with grid search. We show the loss curve and error rate here. Note that we smooth the loss curve with Savgol smoothing.

Problem 2

We construct a DNN with the number of neuron [6, 3, 3, 2]. To find the hyperparameters, we apply the grid search to get the best hyperparameters. In general, our hyperparameters include learning rate, learning rate decay rate, batch size, and L2 regularization. Here, we also compare different optimizers. According to the figure shown on the right, it is obvious that Adam and RMSProp are better than normal SGD. Therefore, we will use Adam in most case for later experiments.



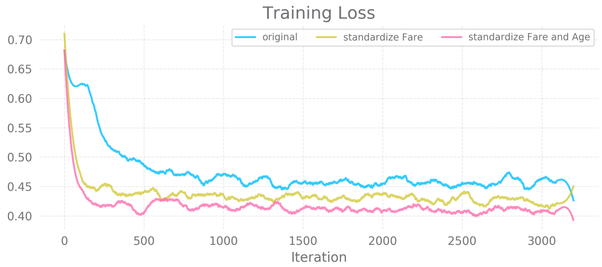


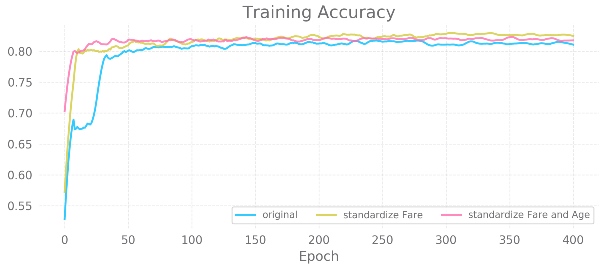


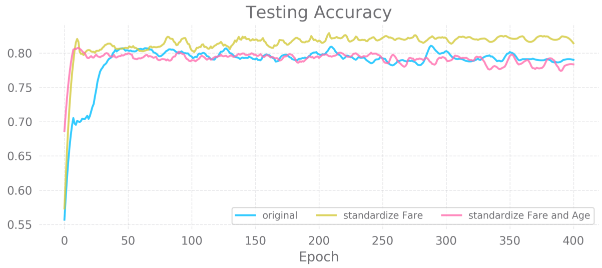
Problem 3

According to the figures shown on the right, it is obvious that standardization on scalar feature is helpful. Model with standardization converges faster than that without standardization.

We also find that we need to initialize weights with a larger variance to model with standardization. One possible reason for this phenomenon is that the value of **Fare** is much larger than other feature. If the weights are too large,

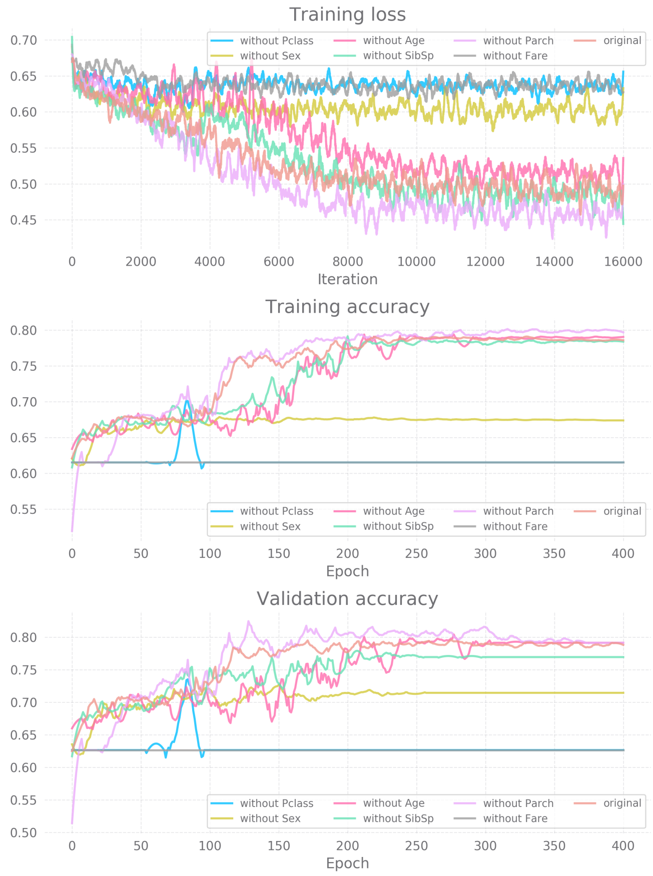






gradient explosion may happen. However, since we standardize the Fare, we can initialize weights with larger variance. Larger weights help to avoid redundant units. They also help to avoid losing signal during forward-propagation or back-propagation. We also try to standardize other scalar features such as Age, SibSp and Parch, but the improvement is not obvious. One possible reason is that these features are not very important for our model.

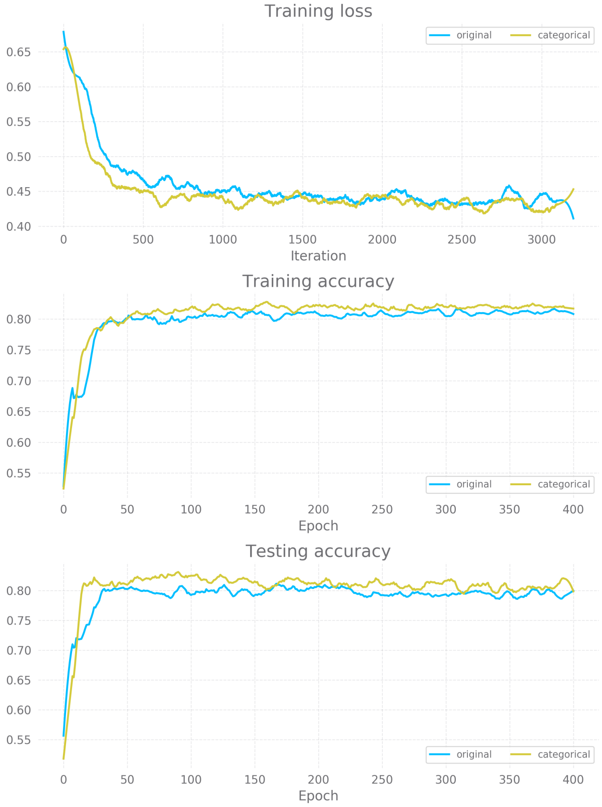
Problem 4



To find the feature that affects the performance the most, we mute one of the feature at a time. Therefore, we can figure out which feature is most important. According to the figures, we can find that **Pclass**, **Fare** and **Sex** are the most important features. This result is consistent with the history of the sinking of the RMS Titanic.

At 00:05 on 15 April 1912, Captain Smith ordered the ship's lifeboats uncovered and the passengers mustered. **The thoroughness of the muster was heavily dependent on the class of the passengers**; the first-class stewards were in charge of only a few cabins, while those responsible for the second- and third-class passengers had to manage large numbers of people. The first-class stewards provided hands-on assistance, helping their charges to get dressed and bringing them out onto the deck. With far more people to deal with, the second- and third-class stewards mostly confined their efforts to throwing open doors and telling passengers to put on lifebelts and come up top. In the third class, passengers were largely left to their own devices after being informed of the need to come on deck. Many passengers and crew were reluctant to comply, either refusing to believe that there was a problem or preferring the warmth of the ship's interior to the bitterly cold night air. The passengers were not told that the ship was sinking, though a few noticed that she was listing. Besides, **women and children have higher priority to board the lifeboats.**

Problem 5



Since **Pclass** is a categorical feature, we encode the **Pclass** to one hot code. According to the figures, one hot encoding is helpful for training. Performance with one hot encoding is also higher than that without one hot encoding. However, we can also find that the training curves of two methods are closed. One possible reason is that **Pclass** is ordinal, so encoding it with 1, 2, 3 still makes sense.

Problem 6

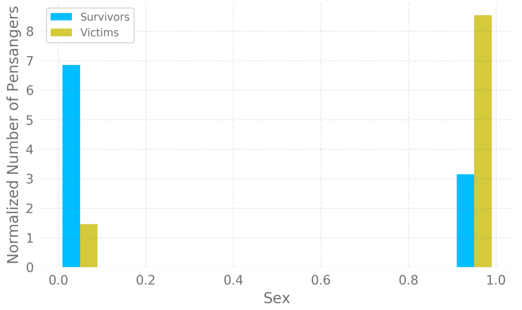
Before we get insight into this problem, let’s take a look at the distributions of these features.

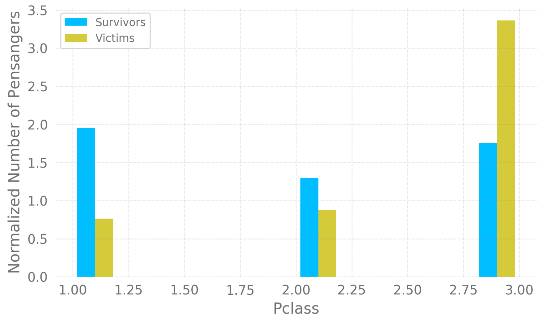
	Pclass	Sex	Age	SibSp	Parch	Fare
count	800	800	800	800	800	800
mean	2.305	0.64625	23.7849	0.51875	0.37375	33.038536
std	0.836869	0.478432	17.700963	1.063514	0.801476	51.52495
min	1	0	0	0	0	0
25%	2	0	5	0	0	7.925
50%	3	1	24	0	0	14.5
75%	3	1	35	1	0	31.275
max	3	1	80	8	6	512.3292

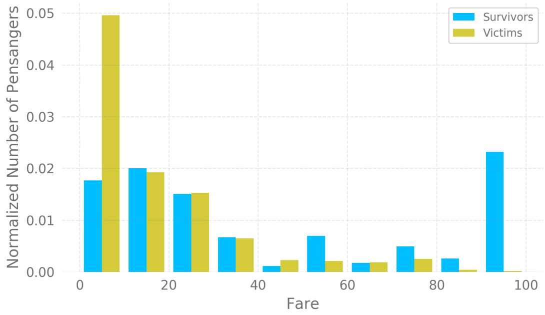
To create fake data of survivor and victim, we need to understand the distribution of the features in the training data. From the figures shown below, we can get three conclusions. (To compare the distribution of survivors and victims fairly, we normalize the data by the number of survivors and that of victims.) First, women (Sex = 0) were more likely to survive. Second, the passengers who paid higher passenger fare were more likely to survive. Third, those who have better ticket class were more likely to survive.

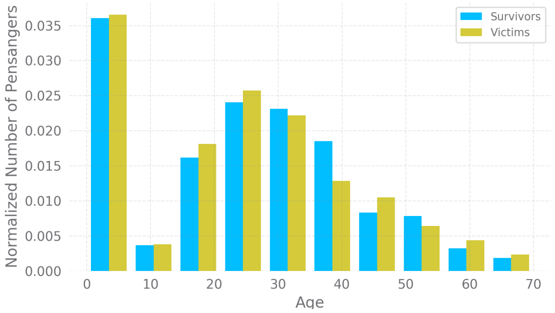
	Pclass	Sex	Age	SibSp	Parch	Fare
Fake survivor	1	0	22	0	0	70
Fake victim	3	1	42	5	0	5

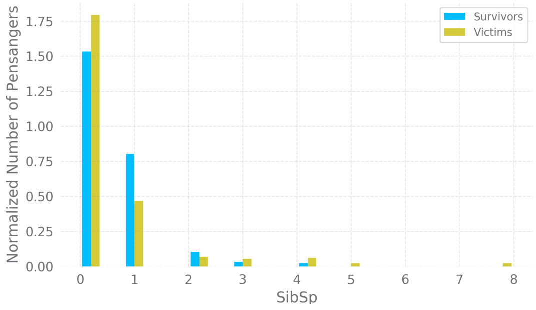
Therefore, we create fake data with the conclusions mentioned above. Our fake data is shown below.

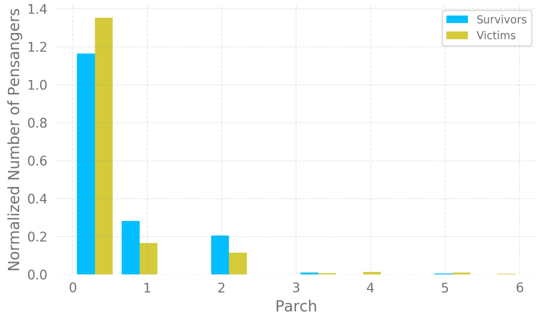

Normalized Number of Pensangers
Survivors
Victims
Sex


Normalized Number of Pensangers
Survivors
Victims
Pclass


Normalized Number of Pensangers
Survivors
Victims
Fare


Normalized Number of Pensangers
Survivors
Victims
Age


Normalized Number of Pensangers
Survivors
Victims
SibSp


Normalized Number of Pensangers
Survivors
Victims
Parch