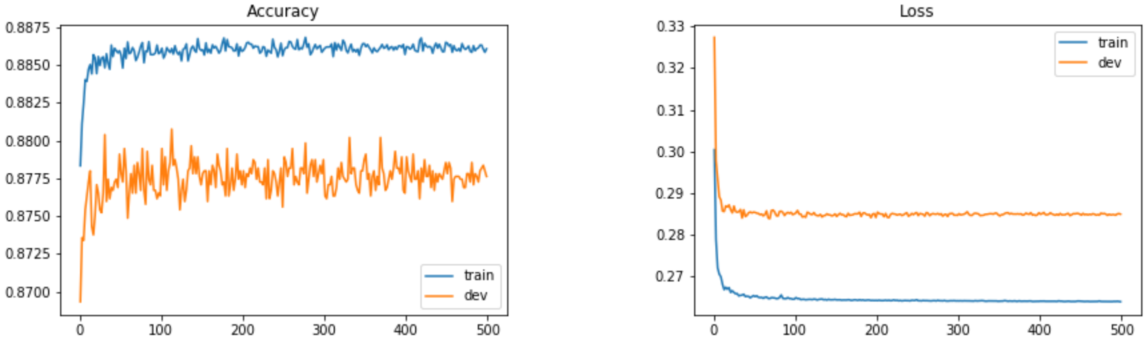
Machine Learning Homework 2 Report

R06221012 數學所 李岳洲

March 26, 2020

1. 請比較實作的 ***generative model*** 及 ***logistic regression*** 的準確率，何者較佳？請解釋為何有這種情況？

   For the training set and the development set, both accuracy scores of logistic regression model are higher than generative model. The reasons should be (1) this dataset is linear classifiable and (2) the number of the dataset is large enough, so that the performance of logistic regression will be better. However, if this dataset is from a probability model, the generative model might be more suitable.



(a) Accuracy

(b) Loss

Figure 1: Logistic model

	Training accuracy	Training loss	Development accuracy	Development loss
Logistic model	0.88615	0.26383	0.87836	0.28576
Generative model	0.87470	–	0.874705	–

Table 1: Comparison table

2. 請實作 ***logistic regression*** 的正規化 ***(regularization)***，並討論其對於你的模型準確率的影響。接著嘗試對正規項使用不同的權重 ***(lambda)***，並討論其影響。

In the experiments of the regularity adjustment, the accuracy of the development set has increased slightly as the increases in the lambda value, but then decreases. It is deduced that the regularity can reduce the scale of the model parameters to increase the degree of generalization of the model, but as the value of the lambda rises, the original model's underfitting will also cause the accuracy of the train set to decrease, and also cause the accuracy of the development set to decrease.

Lambda	0.001	0.01	0.1	1	10
Training accuracy	0.81699	0.83340	0.86587	0.86690	0.86627
Development accuracy	0.80740	0.82602	0.86325	0.86730	0.86483

Table 2: Comparison table (regularization)

3. 請說明你實作的 ***best model***，其訓練方式和準確率為何？

Step 1. Data Preprocessing

a. Normalization

Step 2. Features Extraction

T-test:

We collect all 510 features into the $N \times 510$ matrix and then divide into two matrices, the features in the $N_0 \times 510$ matrix are labeled as 0 and the features in the $N_1 \times 510$ matrix are labeled as 1. Calculate the $p-$value of each feature, if $p-$value is less than 0.05, we will select these features to train and predict.

Step 3. Train

Setting:

a. Loss function: Cross entropy

b. Update of weights: Gradient descent

c. Learning rate: 0.0005

d. Maximum iteration: 2000

e. Early stopping iteration: 20, stop training by judging the validation accuracy.

Step 4. Train again

Setting:

a. Loss function: Cross entropy

b. Update of weights: Gradient descent

c. Learning rate: 0.000001

d. Maximum iteration: 2000

e. Early stopping iteration: 100, stop training by judging the validation accuracy.

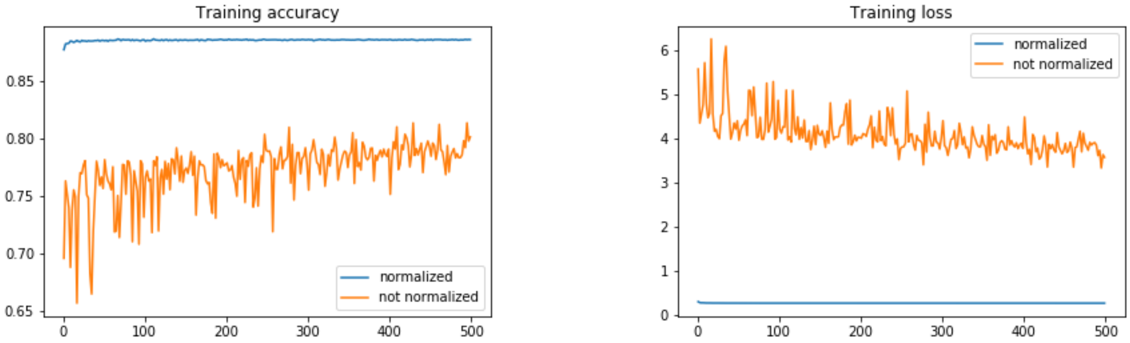
Result .

	Simple baseline	Simple model	Strong baseline	Best model
Accuracy	0.88617	0.88849	0.89052	0.89131

Table 3: Kaggle's public test set accuracy

4. 請實作輸入特徵標準化 ***(feature normalization)***，並比較是否應用此技巧，會對於你的模型有何影響。

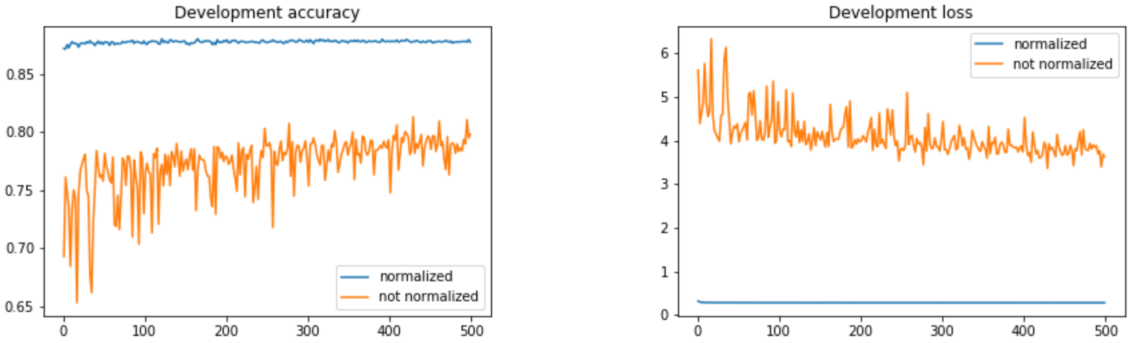
Using normalization will be better than using original features. After normalizing features, both the accuracy curve and the loss curve will smoothly increase and decrease, moreover, both curves fast converge to a local minimum and a local maximum. The reasons are (1) each feature has a different scale and (2) different features have very different effects on the loss function of the gradient descent. If we don't use normalization, it will be quite unstable when the loss drops, that's why we need to take more iterations to reach the local minimum.



(a) Training accuracy

(b) Training loss

Figure 2: Training set



(a) Development accuracy

(b) Development loss

Figure 3: Development set

	Training accuracy	Development accuracy
Normalized	0.88697	0.88002
No normalized	0.81914	0.81865

Table 4: Comparison table (normalized)