Project Report 3

Question1:
The table below shows the accuracy and epochs to converge of different batch_size given the CONV_THRESHOLD=0.001.

	Batch_size	Accuracy(%)	Epochs to converge
1	1	89.3	16
2	5	83.8	23
3	10	83.2	33
4	25	82.1	70
5	75	74.5	69

The table below shows the accuracy and epochs to converge of different batch_size given the CONV_THRESHOLD=0.01.

	Batch_size	Accuracy(%)	Epochs to converge
1	1	81.2	4
2	5	77.8	11
3	10	76.2	11
4	25	71.9	10
5	75	67.4	11

(a)From the experiment results, bigger batch_size leads lower accuracy and more epochs to converge. However, the actual runtime decreased as the batch_size becomes bigger. The conclusion is somehow contrary to what I learnt in class. Typically, using fewer examples in a batch results in a less accurate estimate of the error gradient that is highly dependent on the specific training examples used. The batch_size of 1 may be most noisy and take more epochs to converge. Anyway, there should be a tradeoff between accuracy and epochs to converge, which may be a batch_size larger than 1.

(b)Even though bigger batch_size takes more epochs to converge, the runtime actually decreases. I think numpy must utilize multithread computing to deal with matrix operation. It may take similar time to calculate the gradient matrix of a relatively large batch or a batch with only one example. Therefore, larger batch_size leads to less computation in a single epoch. That's why the runtime decreases as the batch_size become bigger.

Question2:
The categories including 'workclass', 'marital-status', 'occupation', 'relationship', 'race', 'native-country' were one-hot encoded. Because these categories has multiple values. And if we use normal label encoding, we are assuming that higher values lead to better category, which is contrary to the fact.

Question3:
On average, using non-normalized data has a much lower accuracy and more epochs to converge. For example, by setting BATCH_SIZE = 1 and CONV_THRESHOLD = 0.01, the average accuracy for normalized data is 81.2%, while the average accuracy for non-normalized data is 34.1%.

By utilizing normalized, we can overcome the undershooting and overshooting problems on some attributes and improve the accuracy.

Question4:
On average, using data that removes ”race” and ”sex” attributes has better accuracy. For example, by setting BATCH_SIZE = 1 and CONV_THRESHOLD = 0.01, using data that removes ”race” and ”sex” attributes has an average accuracy of 87.0%, while the accuracy of regular normalized data is 81.2%.

For the experiment, the education level is not correlated with race and sex. Removing these attributes, which are not correlate to target value, can decrease the noise and improve the accuracy of the model.