CS589 Homework #7

Trung Dang

November 8, 2023

Task 1

Illustrated in table below is the performance cost of the models on different normalized versions of the dataset with different integer values of γ (and correspondingly, $\lambda = 10^{\gamma}$

γ ($\lambda = 10^{\gamma}$)	Raw	Standard-Normalized	PCA-Sphered
-3	2302.251	43.104	36.927
-2	830.353	41.961	36.855
-1	4537.041	32.182	36.143
0	65625.789	8.109	29.408
1	1410006.0	2.909	4.180
2	14402967.0	15.128	0.678
3	143663648.0	155.54	7.335

Table 1: The cost of each dataset when run with different λ

As can be seen from the table, the optimal λ for the raw dataset is 10^{-2}, with a cost of 830.353. The optimal λ for the standard-normalized dataset is 10, with a cost of 2.909. The optimal λ for the PCA-sphered dataset is 10^2, with a cost of 0.678.

Below are the plots of the cost of each model through 10 iterations, as well as the accuracy of the models in comparison to each other.

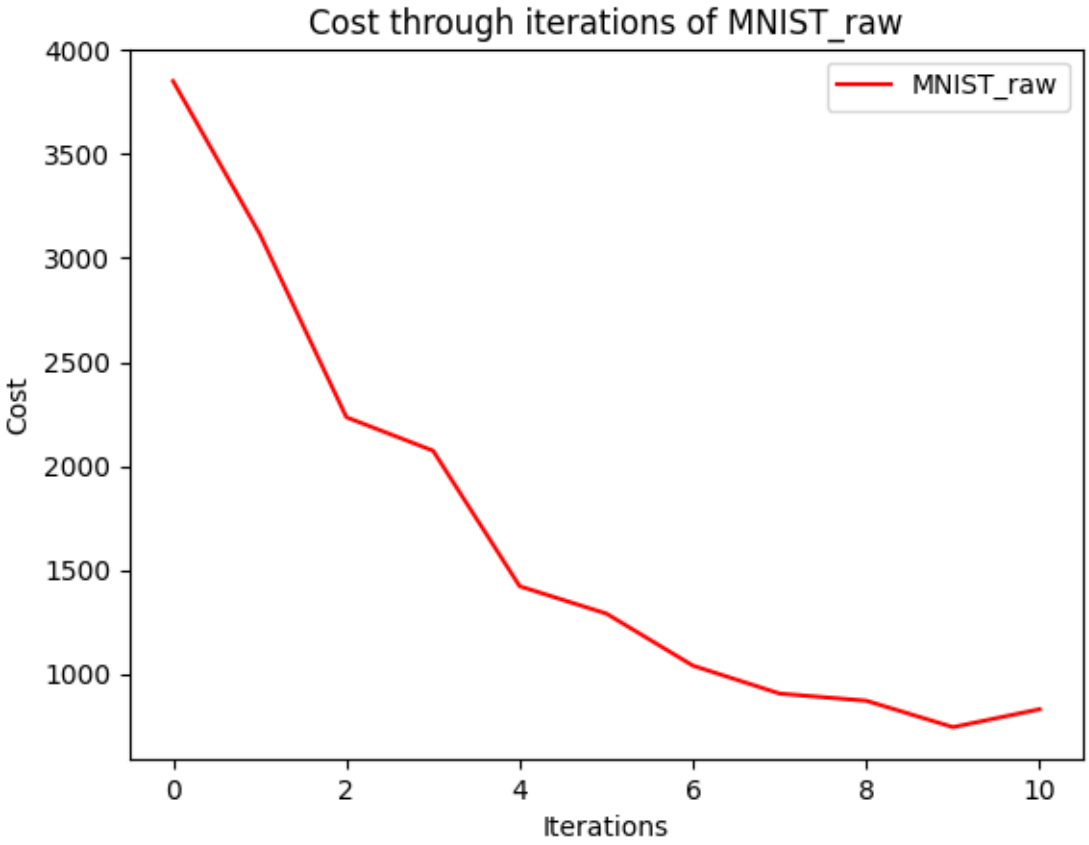

Figure 1: Task 1 cost of raw dataset

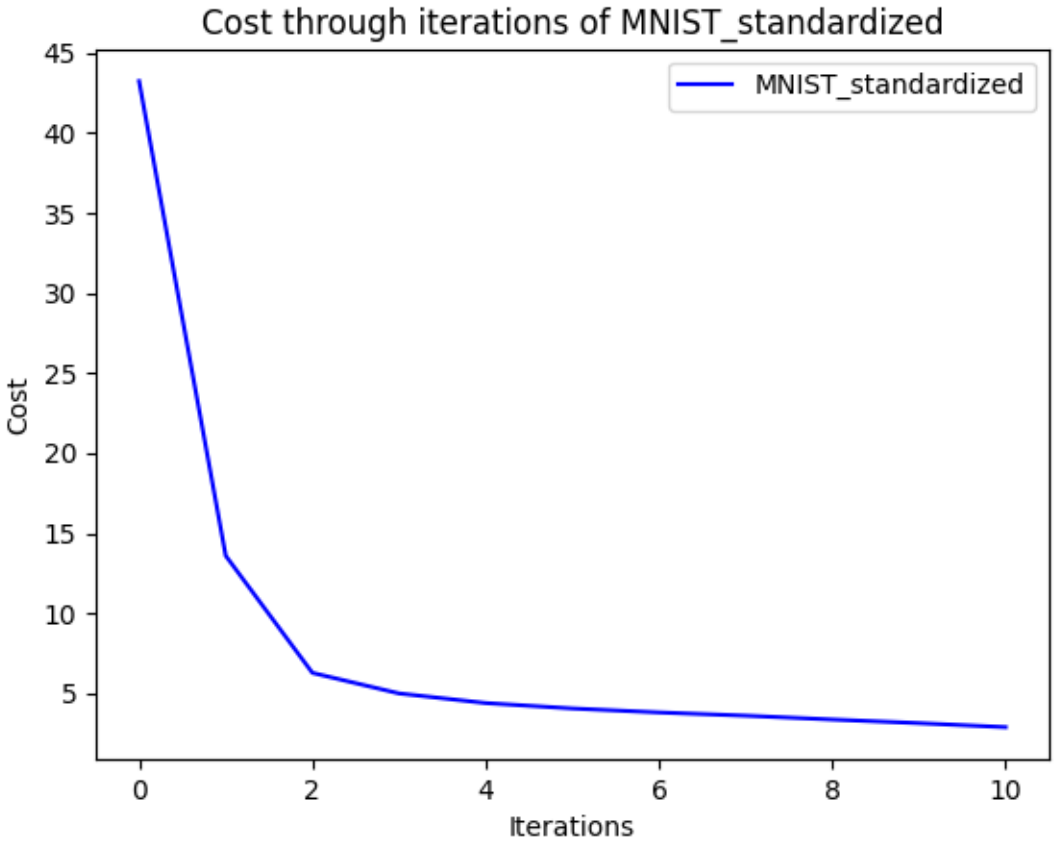

Figure 2: Task 1 cost of standard normalized dataset

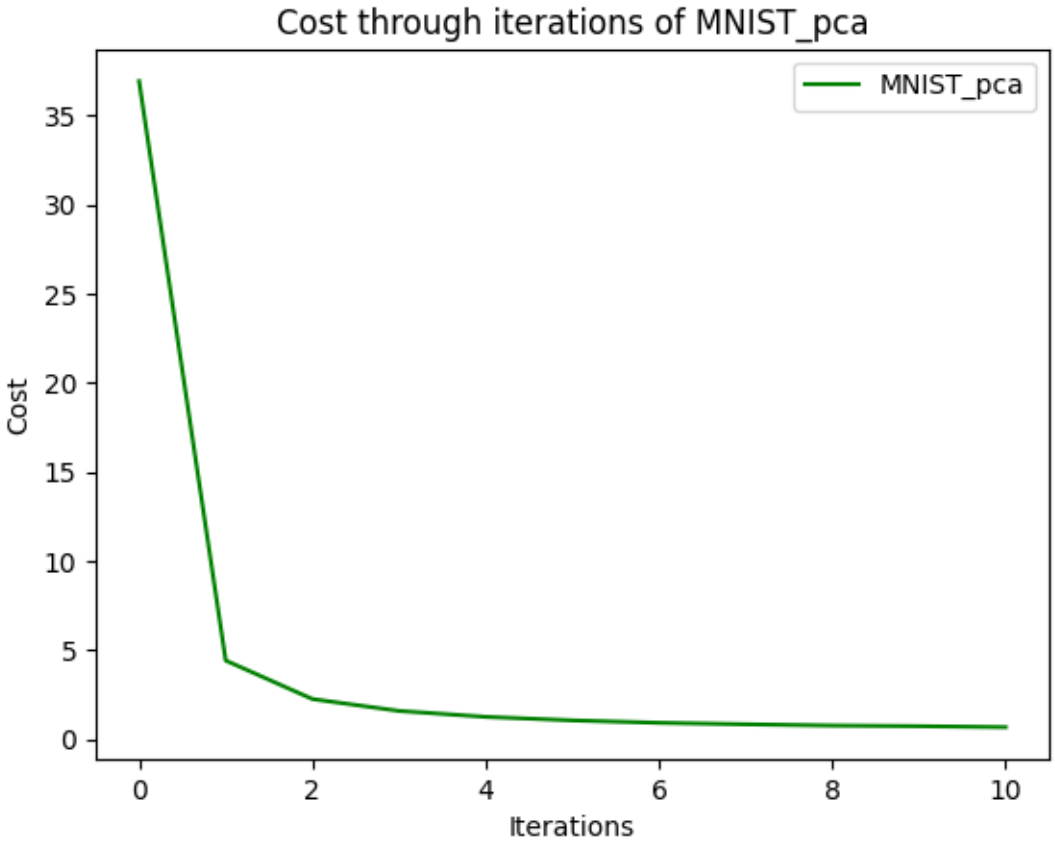

Figure 3: Task 1 cost of PCA-sphered dataset

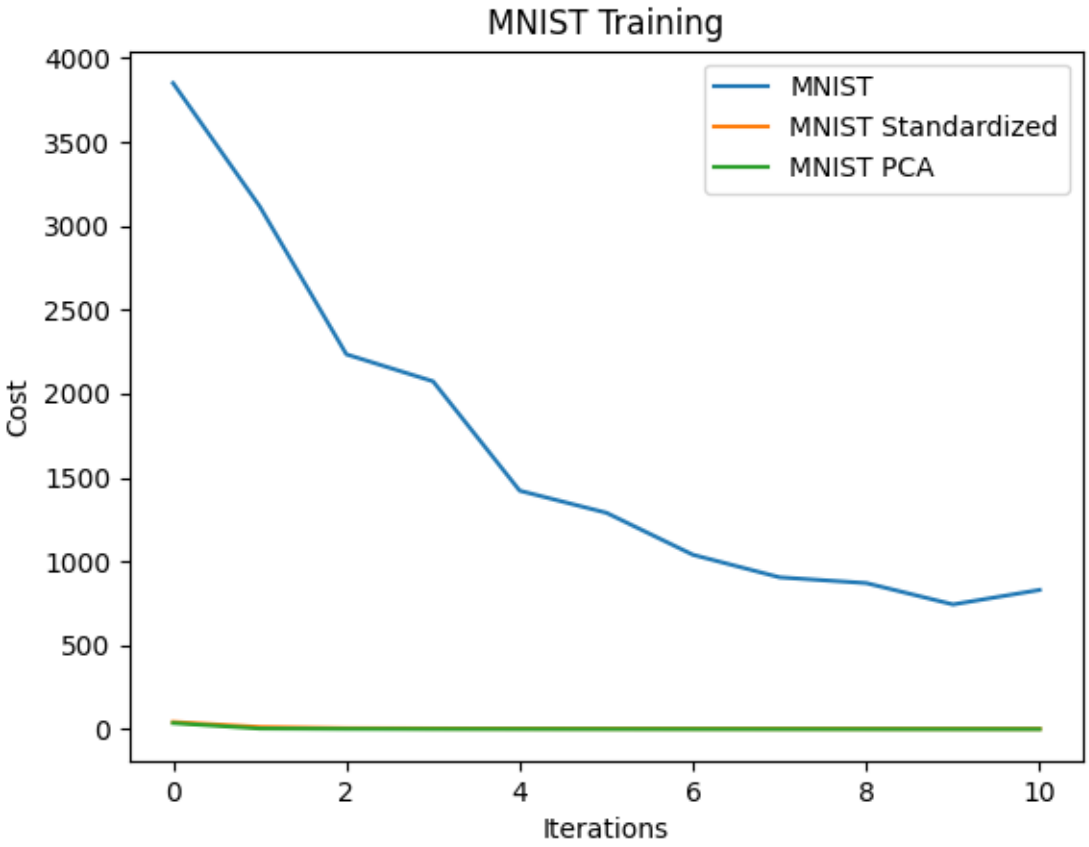

Figure 4: Task 1 costs of the training datasets

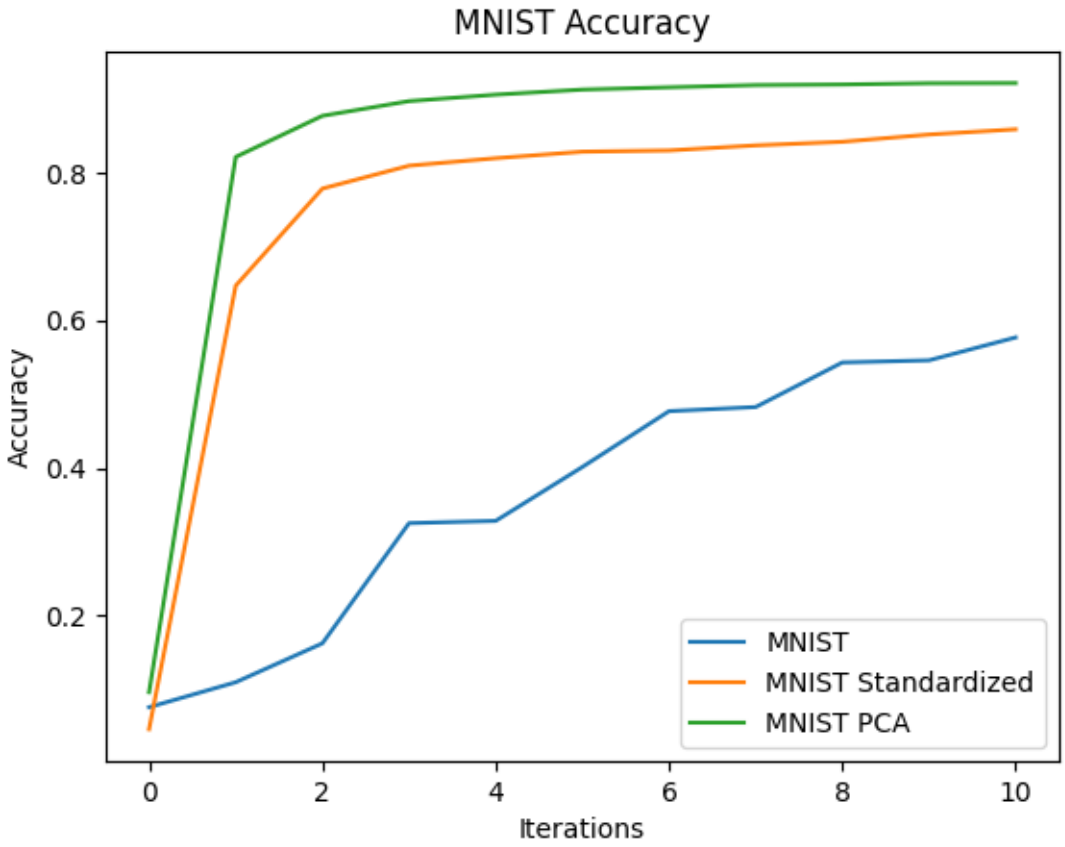

Figure 5: Task 1 Accuracy of the training datasets

Observations: By applying feature engineering techniques such as standard normalization and PCA-sphering, we have enabled the model to converge faster as well as produce better accuracy. The engineered datasets also imply better initialization, as shown by a relatively small cost compared to the raw data. Moreover, the cost also decreases more steeply when the normalization techniques are applied. Likewise, the accuracy of both the models trained on standard normalized and the PCA-sphered datasets also reach higher values (more than 0.8) while the model trained on the raw dataset only achieve roughly 60% accuracy

Task 2

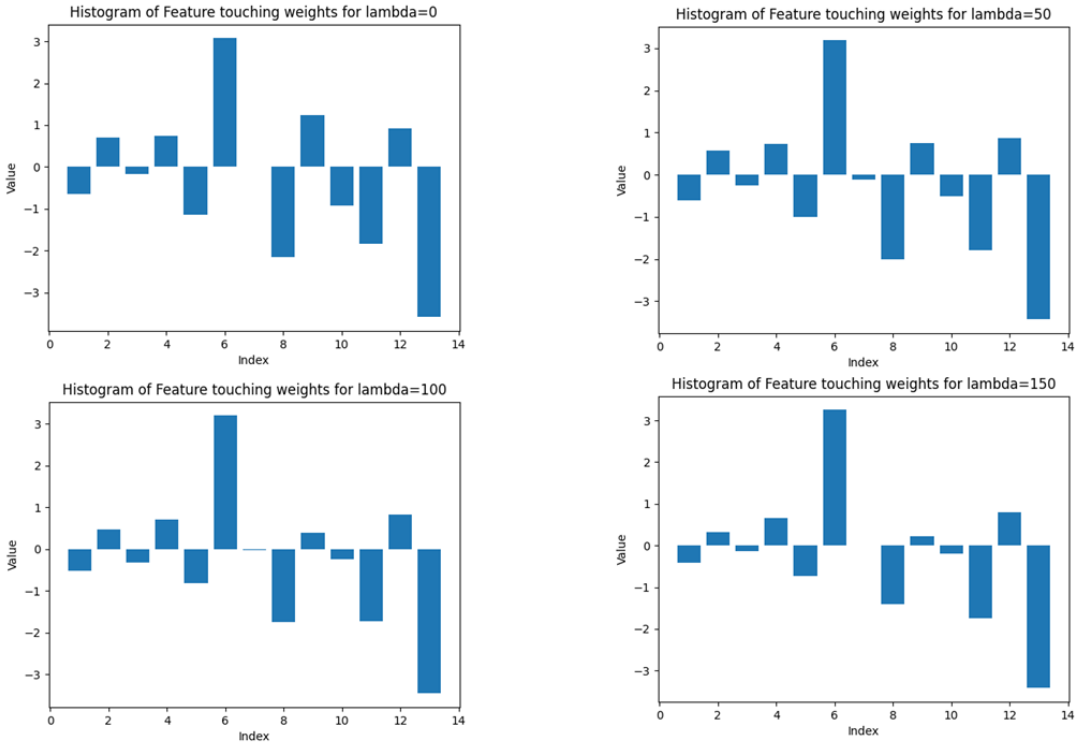

Figure 6: Histogram of the feature touching weights with different λ

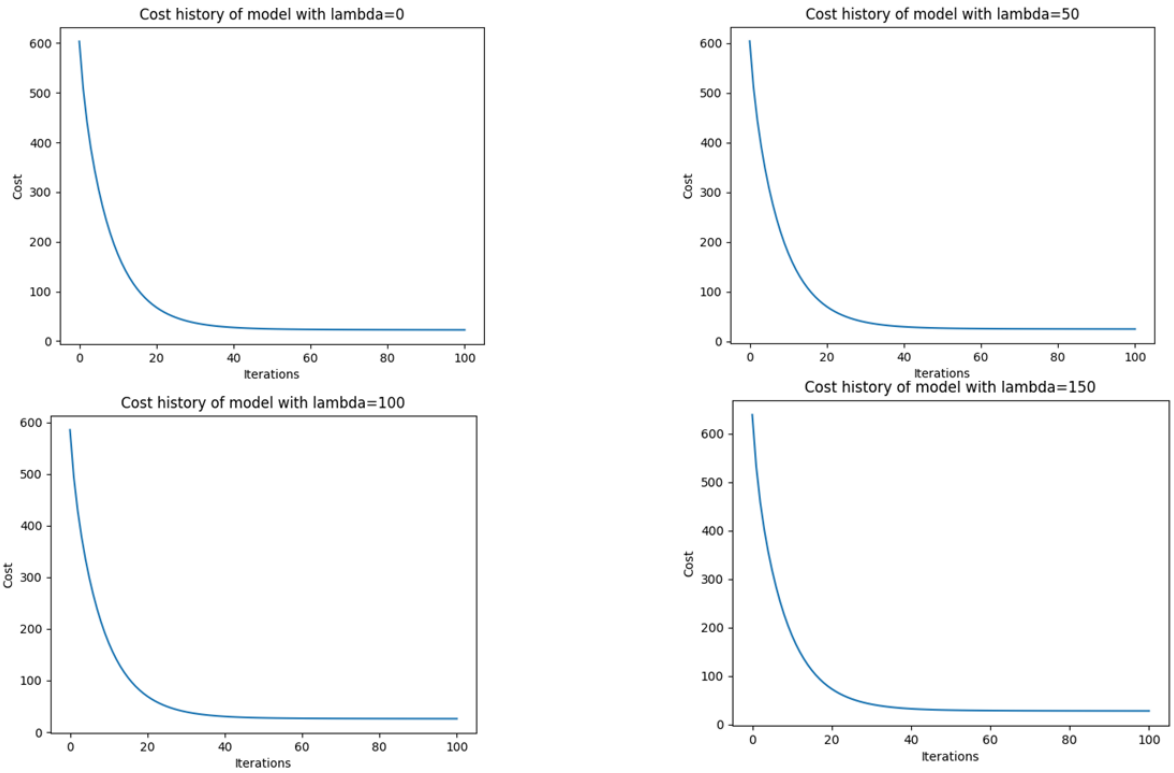

Figure 7: Histogram of the costs with different λ

The step length of my model is selected to be **0.03**, the number of iterations is **100**. The final cost of the models are:

- $\lambda = 0$, final cost=22.48855972290039
- $\lambda = 50$, final cost = 24.392953872680664
- $\lambda = 100$, final cost = 26.08637046813965
- $\lambda = 150$, final cost = 27.589982986450195

As the plots of the cost history for each lambda shows, my model converged.]

Observations: As λ increases, the distinction between significant feature touching weights also increases. In other words, "important weights" remain relatively large, while less important weights are reduced to lower the penalty. Therefore, the resulting graph becomes sparser.