COMPSCI 589

Project 8 Report

Trung Dang – 33858723

I. Task 1

- Choice of K (used for K-fold cross validation)

I found that using **K = 5** yields the best result. One possible explanation is that since our dataset is small, if K is too large then the number of cross validation data points is small, and there is little difference between the validation accuracy of the hyperparameters, while if K is too small then we are spending too much data on cross validation.

- Choice of normalization method (no normalization/Standard Normalization/PCA sphering) and other preprocessing method used (if any)

I decided to use PCA-sphering to preprocess the data. This helps both standard normalize the data and make the contour rounder (which in turn make the model converges faster)

- Cost function used

Noticing that the problem was two class classification and the labels (y) were -1 and 1, I decided to use the Softmax cost function for two class classification. I also added the L1 norm regularizer, which is multiplied with the hyperparameter lambda.

- Hyperparameter setting of your final model (learning rate, penalty of the regularizer, and other hyperparameters if any)

The only hyperparameter tuning I used for this model was the alpha (learning rate) and the lambda (penalty for L1 normalization).

I trained the model with alpha = 1, 0.1, 0.01, and lambda = 0.1, 0.01, 0.001.

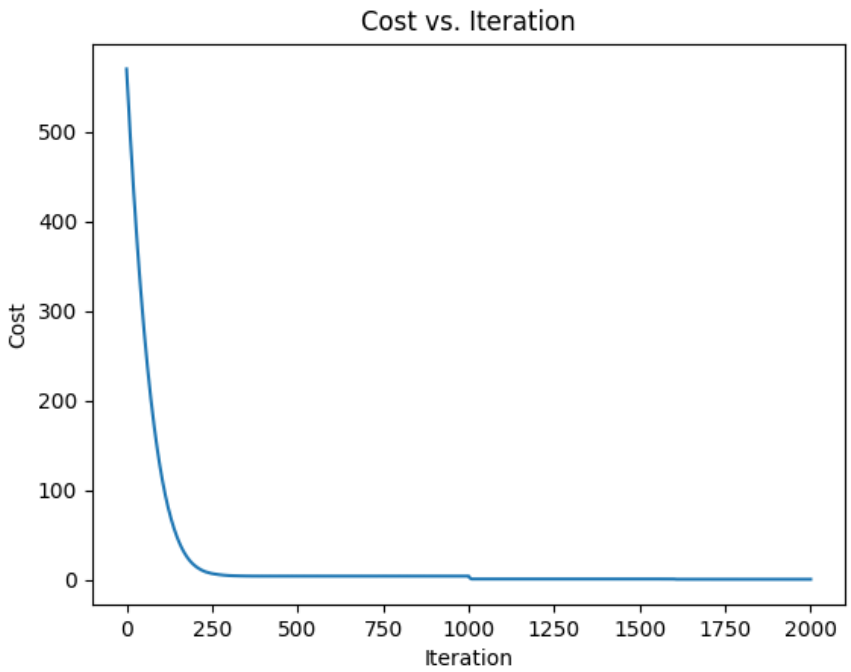
In the end, I found that the best hyperparameters were achieved *simultaneously* at alpha = 1, lambda = 0.1 and alpha = 0.1, lambda = 0.1. I trained both models and observed the result as well as the loss through iterations.

The graphs and statistics provided below was on the model with **alpha = lambda = 0.1**, while the results of model with alpha = 1 and lambda = 0.1 is also reported in appendix.

- Average validation accuracy of the model with the best set of hyperparameters during the K-Fold Cross Validation process.

Average validation accuracy of the model with best set of hyperparameters was: 0.9272728

- Plot of cost vs. iteration of your final model over the entire training set



- Accuracy of your final model on the testing set

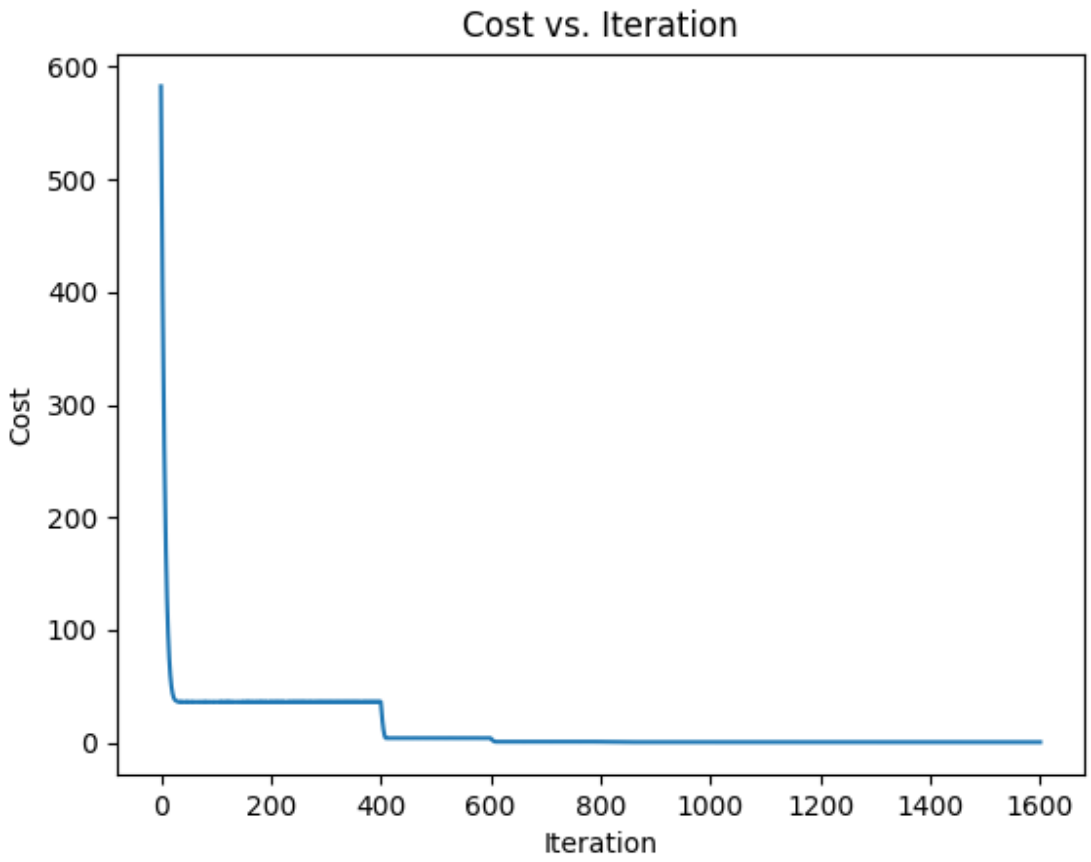
The accuracy of the final model on the testing set was 1.0

- The 5 most influential genes, in descending order

The 5 most influential genes in descending order was (0-indexed) those at index **7126, 7125, 7127, 7095, 7089**

APPENDIX: Result of model with alpha = 1, lambda = 0.1

Graph:



Average validation accuracy of the model with best set of hyperparameters was: 0.9272728

Accuracy when train on entire dataset and eval on test set: 1.0

Most influential genes (0-indexed): **7126, 7125, 7127, 7124, 7095**