Problem:

Download the training data set "musicdata.txt" and the test data set "musictestdata.txt" from Avenue. The data consist of 91 variables concerning commercial songs released between 1922 and 2011. The first variable (the song's release year) is the response variable that we are trying to predict based on the remaining 90 input variables which consist of various measures of timbre. There are 1000 songs in the training data and 300 in the test data. As there are a considerable number of input variables over fitting is a serious issue. In order to avoid this, implement the regularized ML estimation with lasso regularizer. Your function should accept a scalar value of λ, a vector-valued response variable (y) and a matrix of input variables (X) and it should output a vector of coefficient values w. Once you have implemented a lasso solver function run the solver using 100 different values of λ (try different range of values for λ). In your analysis, include:

1. A plot of log(λ) against the squared error in the training data.
2. A plot of log(λ) against the squared error in the test data.
3. A plot of λ against the number of nonzero coefficients.

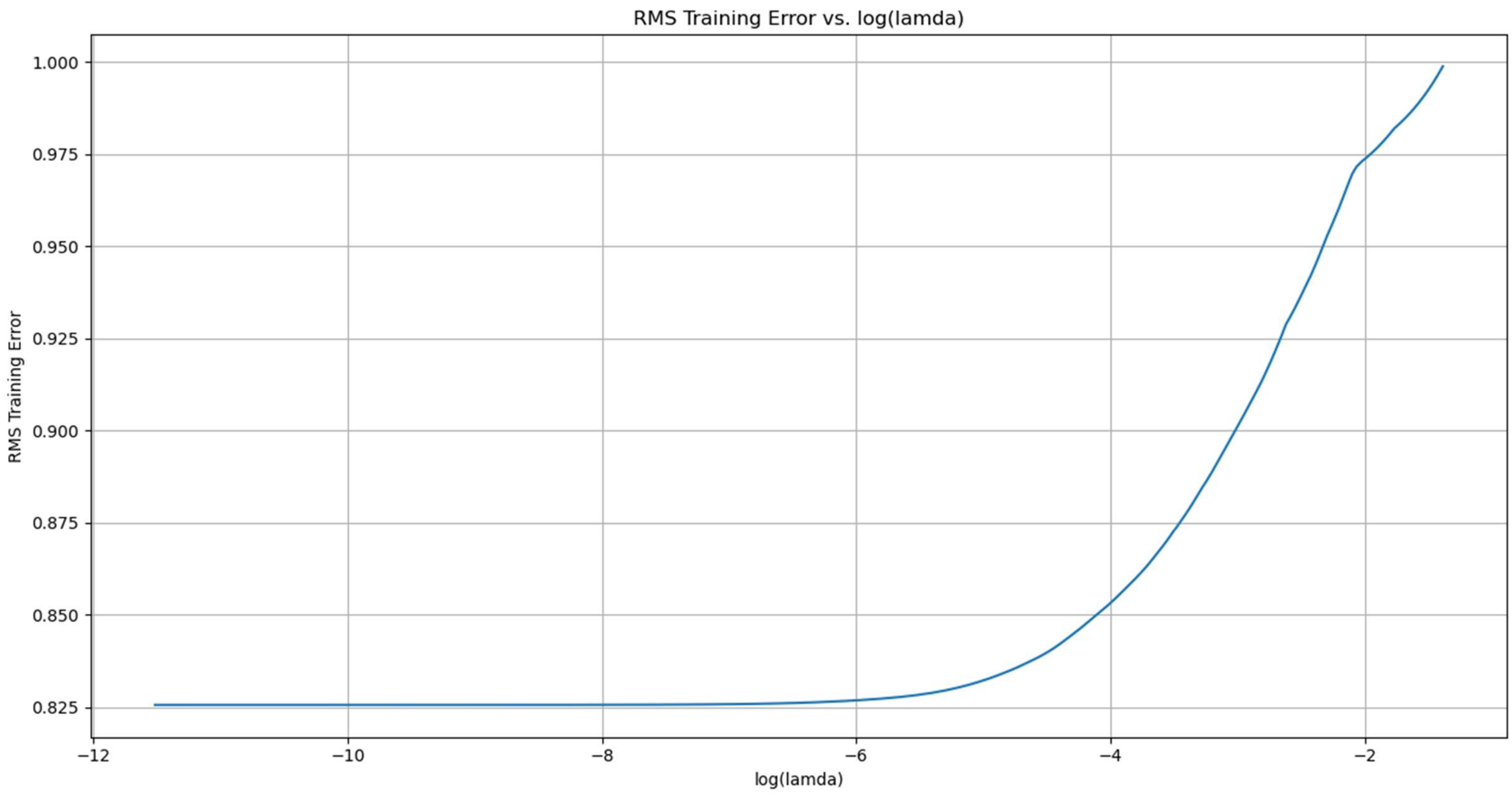
Solution:

question2.py is the python script for this question.

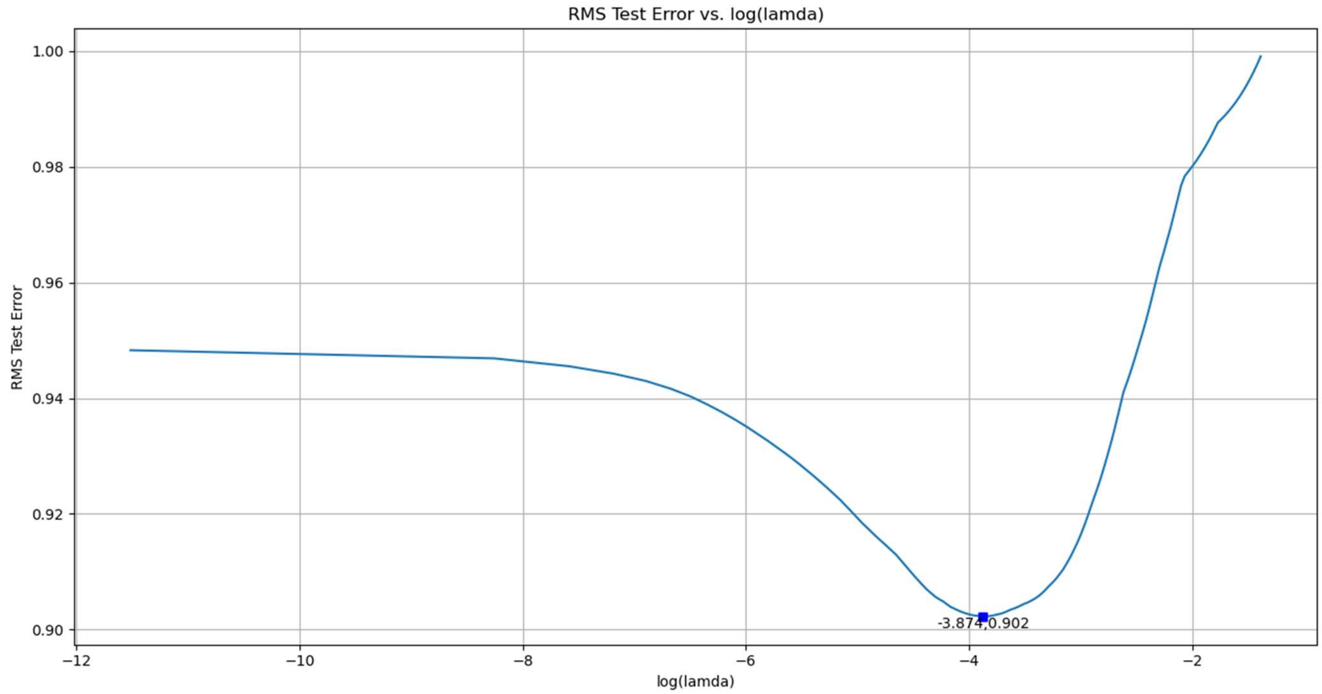
The musicdata dataset has 90 measures of timbre for each of the songs. Its not necessary that all of the 90 features be relevant for predicting which year a new song may have been released, therefore lasso regression can be used to determine which of the 90 features are most useful for the prediction.

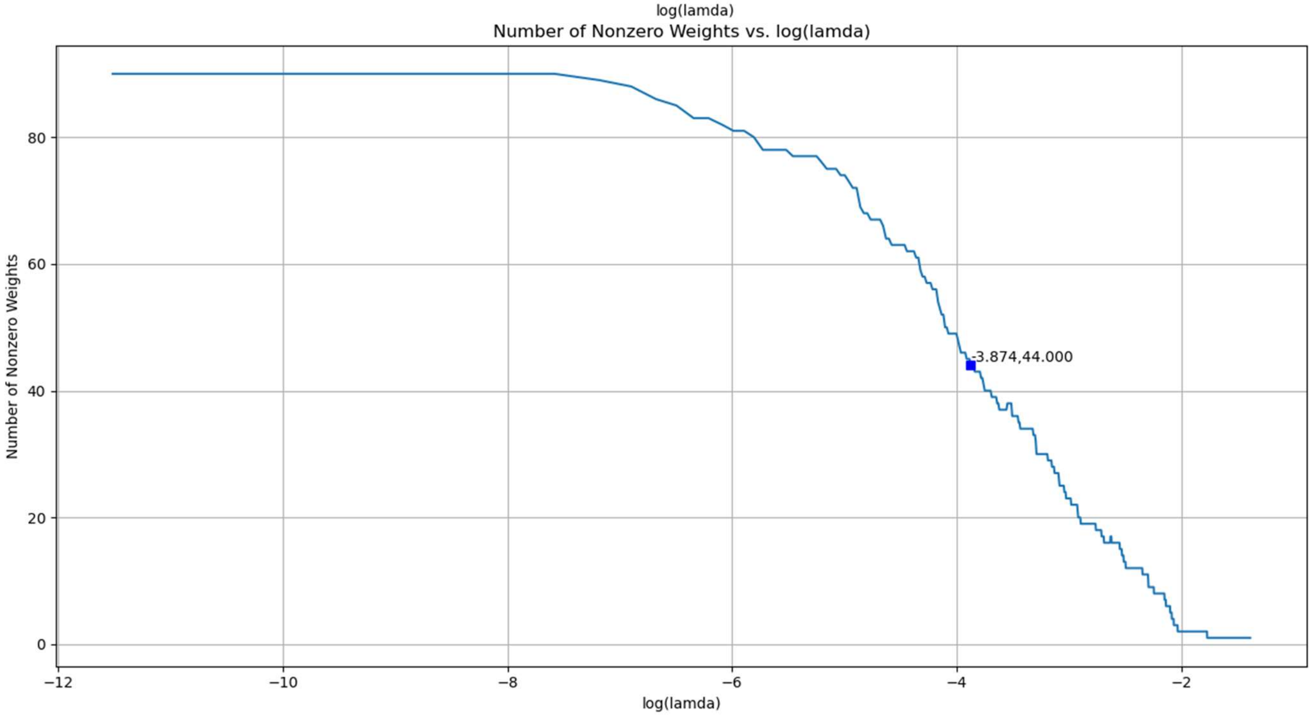
Lasso (L1) and Ridge (L2) Regression are similar in the sense that they force the coefficients of the estimated model to be smaller in magnitude and prevent overfitting the data by providing some value for lambda, the regularizer value. But where they differ is that Lasso regularization will set the coefficients of irrelevant features to zero, whereas Ridge Regression will set the coefficients to very low value but not zero.

In this exercise we will see that by increasing the value of lambda from zero to a higher value in the lasso regularizer, the number of nonzero terms in the coefficient vector will reduce. The range of values for lambda I have chosen are 1e-5 to 0.25. On the next page there are three plots for log(lambda) vs. RMS test error, log(lambda) vs. RMS training error, and log(lambda) vs. Number of Nonzero coefficients.



As we can see the RMS training error increases as we penalize larger values for the coefficients by increasing the lambda value in lasso regularizer. This is because by penalizing large values for the coefficients we are creating a model that does not fit the training data as well, but the goal of any regularization technique is to get more accurate predictions for the test set and other new inputs not already encountered in the training of the model.





The previous two plots show us that a log(lambda) value of -3.874 is related to a local minimum in the RMS test error, which is related to 44 nonzero coefficients.

Next steps could be to do L2 regularization on the 44 features related to the 44 nonzero coefficients to further reduce the RMS test error.