COMPENG 4SL4 Assignment 2 Report

Raeed Hassan
hassam41
400188200

November 4, 2022

Data Set Loading and Splitting

The data set was loaded using `sklearn.datasets.load_boston` and split into a training set (containing 80% of the data sample points), and a test set (containing the other 20% of the data sample points). The `sklearn.model_selection.train_test_split` function was used to perform the split, using the last 4 numbers of my student ID (8200) as the random state for the function.

K-Fold Cross-Validation Setup

For K-fold cross-validation in this lab, the value of K was set to 5. Arrays containing the training and test indexes for each fold were created using `sklearn.model_selection.KFold`. These indices are used for all K-Fold cross-validation done throughout the lab. These indices are used for row indexing to generate the feature and target matrices before doing regression.

Feature Selection with no Basis Expansion

K-Fold cross-validation is performed for remaining feature in a loop until all features have exhausted. The error for each fold is calculated, then the errors are averaged to calculate the cross-validation error for each feature tested. The feature will the lowest cross-validation error is selected as the best feature. The model is then retrained with the best feature using the entire training set and the test error is calculated using the test set.

Table 1 shows the best feature for k = 1–13, and the cross-validation and test errors when adding this feature to S. For every k, when using the best feature the cross-validation error was smaller than the test error. This relation also largely scaled with the cross-validation error, with a larger cross-validation error generally corresponding to a larger test error. The lowest cross-validation error was recorded for $k = 11$, which had the third lowest test error. The largest cross-validation error at $k = 1$ also had the largest test error. The cross-validation and test errors are plotted on Figure 1.

Feature Selection with Basis Expansion

When the best feature for any k was chosen, K-fold cross-validation basis expansion was performed against but with basis expansion performed on the feature. The two basis functions chosen were selected to compare two functions that would generally scale down the values of all the features. The two basis functions selected were $\phi_k(x) = \sqrt{x_j}$ and $\phi_k(x) = \ln(x_j + 1)$. The second basis function had modified to include the addition as many of the values in feature $f_{(4)}$ contain zeros, and this addition barely affected the cross-validation error when tested. The cross-validation errors using the two basis functions were compared, and the function that produced the lower cross-validation error was selected for every k, and the model is then retrained with the best feature and basis expansion using the entire training set and the test error is calculated using the test set.

Table 2 shows the best feature for k = 1–13 with basis expansion, and the cross-validation errors with both basis functions, and test error with the best basis function. For all values of k, the basis function, b_2, which is $\phi_k(x) = \ln(x_j + 1)$ outperformed b_1 with a smaller cross-validation error. Similar to the case with no basis expansion, when the model is retrained with basis expansion using the training set and the test error is calculated, the cross-validation error is smaller than the test error for every k. This relation also largely scaled with the cross-validation error, with a larger cross-validation error generally corresponding to a larger test error. The lowest cross-validation error was recorded for $k = 13$, which had the lowest test error. The largest cross-validation error at $k = 1$ also had the largest test error. The cross-validation and test errors are plotted on Figure 1.

Table 1: Errors for Best Feature for k = 1–13

k	Best Feature	Cross-Validation Error	Test Error
1	$f_{(13)}$	37.52981349037695	44.01642965365188
2	$f_{(6)}$	30.232181061635934	34.84114009877059
3	$f_{(11)}$	26.93968194188119	32.20558361910458
4	$f_{(8)}$	26.177187358856212	30.645130159002235
5	$f_{(5)}$	24.912879835977943	28.286173753054985
6	$f_{(12)}$	24.30481665830434	28.388570169930453
7	$f_{(2)}$	24.045578821916497	27.829372049292544
8	$f_{(1)}$	23.93238097885264	27.62940695273575
9	$f_{(9)}$	23.553279587919693	27.145745012675526
10	$f_{(10)}$	22.87754245924457	26.15100337729019
11	$f_{(3)}$	22.936063395709063	26.10522648390962
12	$f_{(7)}$	23.0123559430166	26.203337950256564
13	$f_{(4)}$	23.17317153083585	25.081388777344092

Table 2: Errors for Best Feature for k = 1–13 with Basis Expansion

k	Cross-Validation Error for b_1	Cross-Validation Error for b_2	Best Basis Function	Test Error
1	31.804778028535747	28.78911747132245	b_2	30.26744238122664
2	27.82424066223519	26.18837863475094	b_2	28.382217037889895
3	25.084940264392877	23.804041380557386	b_2	26.5603133022723
4	23.511570217312173	22.09032521586629	b_2	23.83276391329117
5	22.227661019873914	20.738081968392216	b_2	21.77507921513151
6	21.619810645158452	20.130119743750072	b_2	21.813762402597206
7	21.468492327469754	20.088725383807375	b_2	21.900783544904453
8	21.551075676436437	20.192849446212882	b_2	21.897733955789434
9	20.86242873106243	19.297245284435142	b_2	22.02694846079194
10	20.10985039649996	18.68610785258567	b_2	21.07789265155987
11	20.141963029677022	18.653648609589492	b_2	20.91134050704141
12	20.096759819961267	18.54072504399405	b_2	21.115487227577024
13	20.16513360676659	18.531483796143725	b_2	20.077605226890636

Cross-Validation and Test Error Plot

Figure 1 shows the cross-validation and test errors before and after basis expansion, with the errors for no basis expansion marked with circles and errors with basis expansion marked with crosses. We can see that performing basis expansion decreases both the cross-validation and test errors for the model. We can also see that the cross-validation errors are smaller than the test errors in both cases, as previously discussed.

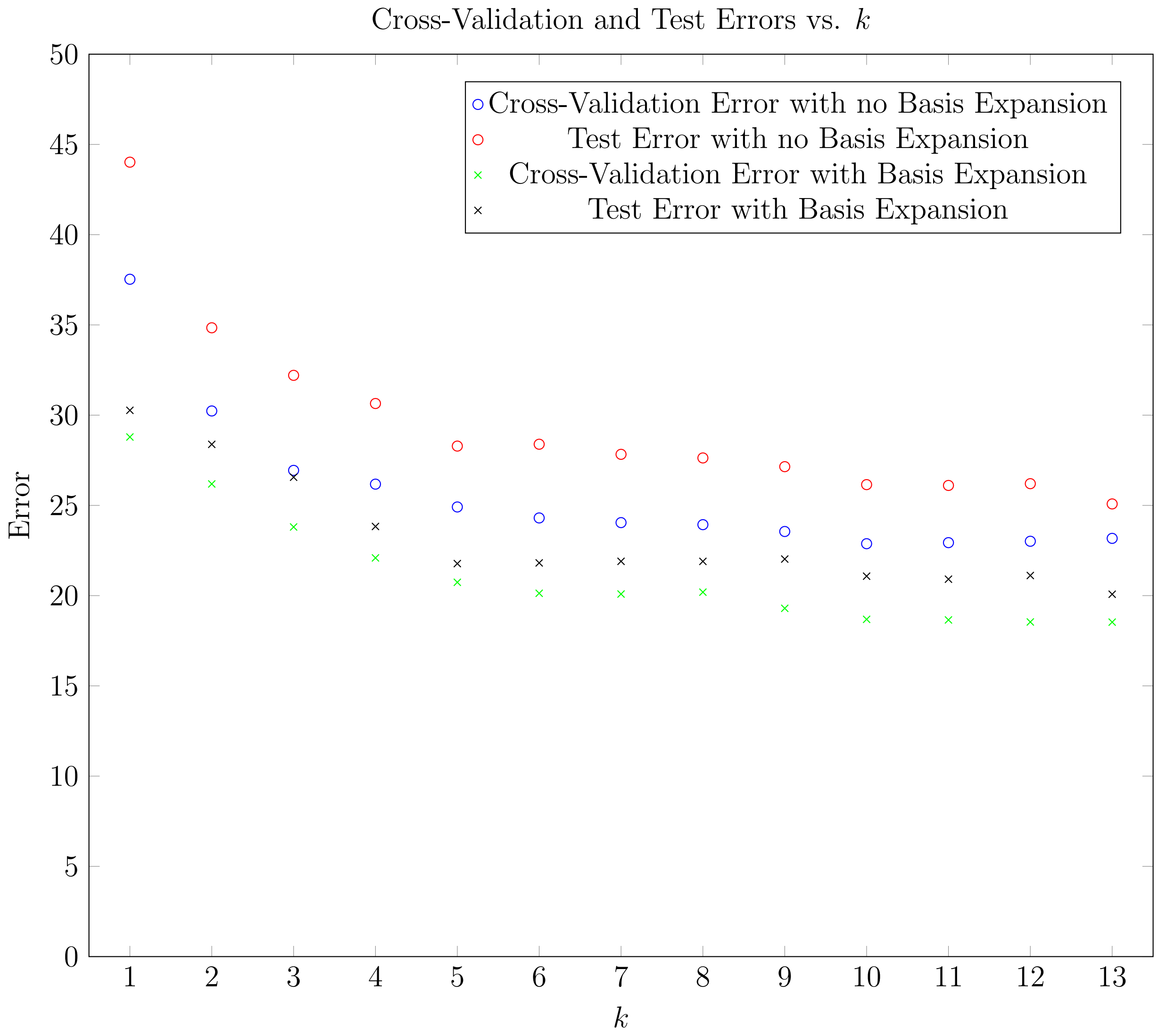

Figure 1: Cross-Validation and Test Errors vs. k