# PS9

Opal Fraser

April 11, 2023

## 1 Questions

7. What is the dimension of your training data (housing train)? it has 937 rows and 7 columns.(14) How many more X variables do you have than in the original housing data? Same as housing, 7. 8. What is the optimal value of l? Df Dev Lambda 1 0 0.00 0.33380 2 1 11.10 0.30420 3 1 20.31 0.27720 4 1 27.96 0.25250 5 1 34.31 0.23010 6 1 39.59 0.20970 7 1 43.96 0.19100 8 1 47.60 0.17410 9 1 50.61 0.15860 10 1 53.12 0.14450 11 2 55.55 0.13170 12 2 57.94 0.12000 13 4 60.43 0.10930 14 5 62.73 0.09961 15 5 64.71 0.09076 16 5 66.35 0.08269 17 5 67.71 0.07535 18 6 69.41 0.06865 19 7 70.92 0.06256 20 8 72.25 0.05700 21 8 73.40 0.05193 22 8 74.36 0.04732 23 10 75.31 0.04312 24 11 76.31 0.03929 25 12 77.34 0.03580 26 13 78.23 0.03262 27 13 79.05 0.02972 28 14 79.74 0.02708 29 16 80.43 0.02467 30 19 81.08 0.02248 31 23 81.75 0.02048 32 23 82.40 0.01866 33 25 82.96 0.01701 34 28 83.54 0.01550 35 29 84.12 0.01412 36 28 84.56 0.01286 37 30 84.96 0.01172 38 32 85.31 0.01068 39 36 85.71 0.00973 40 38 86.14 0.00887 41 39 86.49 0.00808 42 42 86.80 0.00736 43 44 87.08 0.00671 44 45 87.35 0.00611 45 46 87.58 0.00557 46 48 87.79 0.00507 47 48 87.98 0.00462 48 51 88.13 0.00421 49 52 88.28 0.00384 50 53 88.41 0.00350 51 53 88.52 0.00319 52 54 88.61 0.00290 53 55 88.69 0.00265 54 55 88.75 0.00241 55 58 88.81 0.00220 56 61 88.89 0.00200 57 62 88.95 0.00182 58 64 89.01 0.00166 59 65 89.06 0.00151 60 65 89.10 0.00138 61 66 89.14 0.00126 62 66 89.16 0.00114 63 66 89.19 0.00104 64 65 89.21 0.00095 65 65 89.22 0.00087 66 66 89.23 0.00079 67 67 89.25 0.00072 68 67 89.25 0.00066 69 68 89.26 0.00060 70 69 89.27 0.00054 71 69 89.28 0.00050 72 69 89.28 0.00045 73 69 89.29 0.00041 74 70 89.29 0.00038 75 70 89.29 0.00034 76 70 89.30 0.00031 77 71 89.30 0.00028 78 73 89.30 0.00026 79 74 89.31 0.00024 80 73 89.32 0.00021 81 73 89.32 0.00020 82 74 89.34 0.00018 83 74 89.37 0.00016 84 74 89.38 0.00015 85 72 89.40 0.00013 86 72 89.41 0.00012 87 73 89.42 0.00011 88 73 89.43 0.00010 89 73 89.44 0.00009 90 73 89.44 0.00008 91 73 89.45 0.00008 92 73 89.46 0.00007 93 73 89.46 0.00006 94 72 89.46 0.00006 95 72 89.47 0.00005 96 72 89.47 0.00005 97 72 89.47 0.00004 98 73 89.48 0.00004 99 73 89.48 0.00004 100 73 89.48 0.00003 in-sample RMSE was 0.413 out-of-sample RMSE is 0.390 in-sample Rsq'd was 0 out-of-sample Rsq'd is 0

9. What is the optimal value of l now? What is the out-of-sample RMSE (i.e. the RMSE in the test data)? 10. Would you be able to estimate a simple linear regression model on a data set that had more columns than rows?

.metric .estimator .estimate .config ¡chr¿ ¡chr¿ ¡dbl¿ ¡chr¿ 1 rmse standard 0.170 Preprocessor1 Model1 2 rsq standard 0.810 Preprocessor1 Model1 ¿ top rmse print(n = 1) A tibble: 5 × 7 penalty .metric .estimator mean n std err .config ¡dbl¿ ¡chr¿ ¡chr¿ ¡dbl¿ ¡int¿ ¡dbl¿ ¡chr¿ 1 0.000869 rmse standard 0.0680 10 0.00445 Preproces... The model would be too complex and have to many parameters relative to the data, which can lead to overfitting where the model fits the noise in the data instead of the underlying signal. Using the RMSE values of each of the tuned models in the previous two questions, comment on where your model stands in terms of the bias-variance tradeoff. The second model resulted in a higher rsquared.