# Assignment 6 *

Ishaan Kapoor

March 21, 2022

## 1 Linear Regression & Cross Validation

### A

Least Square coefficients: [-3.713, 5.038, -1.71, 4.745, -0.364, 1.532, 0.842, -3.754, -1.279, 2.774, 4.341, 0.01, 1.016, -0.627, 2.842, 0.138, -4.136, -0.026, 3.7, -1.375, -1.439, 1.1, -0.791, -0.316, 6.028, -0.243, -1.397, -0.426, -1.013, 0.894, -2.37, -0.472, -4.243, -1.154, 0.19, -0.793, -1.883, 2.843, 1.814, 0.482, -3.23, 0.557, 1.333, -0.835, 0.212, -4.27, 0.232, 3.082, 0.447, -2.795]
error= 3.452
s= 0.1 coefficients: [0.334, 1.731, -0.328, 0.943, 1.317, 0.645, -0.24, -0.365, 1.054, 0.708, 0.409, 0.751, 1.265, 0.162, 0.616, -0.018, 0.829, -0.312, 0.454, -0.108, 0.081, 1.207, -0.168, 0.503, 1.507, 0.261, 0.047, 0.697, 0.277, -0.577, 0.399, 0.388, -0.279, 0.4, -0.357, 0.148, 0.416, 1.175, 1.158, 0.851, -0.731, 0.554, 0.732, -0.646, 0.569, -0.51, 0.162, 1.1, 0.054, -0.337]
error= 3.697
s= 0.3 coefficients: [0.273, 1.361, -0.281, 0.637, 1.172, 0.516, -0.137, -0.083, 0.967, 0.591, -0.034, 0.714, 1.08, -0.008, 0.464, -0.001, 0.651, -0.011, 0.465, -0.045, 0.179, 0.911, -0.076, 0.451, 1.209, 0.337, 0.145, 0.595, 0.427, -0.312, 0.472, 0.585, -0.023, 0.359, -0.275, 0.12, 0.419, 0.917, 0.897, 0.913, -0.369, 0.361, 0.229, -0.292, 0.392, -0.319, 0.284, 0.775, 0.021, -0.279]
error= 3.9
s= 0.7 coefficients: [0.22, 1.083, -0.287, 0.476, 0.928, 0.403, -0.058, 0.071, 0.832, 0.491, -0.176, 0.588, 0.868, -0.117, 0.374, -0.029, 0.504, 0.141, 0.463, -0.043, 0.154, 0.679, -0.054, 0.4, 0.988, 0.343, 0.16, 0.449, 0.456, -0.163, 0.5, 0.59, 0.091, 0.283, -0.207, 0.084, 0.289, 0.732, 0.721, 0.941, -0.212, 0.261, -0.008, -0.141, 0.193, -0.194, 0.293, 0.546, -0.01, -0.208]
error= 4.199
s= 0.9 coefficients: [0.209, 0.999, -0.287, 0.437, 0.847, 0.366, -0.042, 0.105, 0.782, 0.458, -0.193, 0.541, 0.797, -0.143, 0.347, -0.041, 0.462, 0.172, 0.456, -0.051, 0.134, 0.611, -0.054, 0.384, 0.919, 0.335, 0.158, 0.401, 0.453, -0.131, 0.498, 0.565, 0.109, 0.258, -0.183, 0.072, 0.24, 0.683, 0.672, 0.934, -0.177, 0.237, -0.048, -0.118, 0.138, -0.157, 0.281, 0.485, -0.018, -0.186]
error= 4.327
s= 1.1 coefficients: [0.201, 0.932, -0.286, 0.407, 0.781, 0.336, -0.032, 0.128, 0.738, 0.43, -0.2, 0.502, 0.74, -0.16, 0.325, -0.05, 0.429, 0.193, 0.446, -0.059, 0.116, 0.558, -0.054, 0.371, 0.863, 0.326, 0.156, 0.362, 0.446, -0.108, 0.492, 0.539, 0.119, 0.237, -0.163, 0.063, 0.201, 0.645, 0.634, 0.922, -0.151, 0.219, -0.072, -0.104, 0.097, -0.129, 0.268, 0.439, -0.023, -0.169]
error= 4.446
s= 1.3 coefficients: [0.196, 0.876, -0.283, 0.383, 0.727, 0.31, -0.025, 0.145, 0.7, 0.406, -0.203, 0.47,

*CS 6140 Data Mining; Spring 2022 Instructor: Qingyao Ai, University of Utah

0.692, -0.173, 0.306, -0.057, 0.402, 0.206, 0.437, -0.068, 0.1, 0.514, -0.056, 0.359, 0.815, 0.317, 0.152, 0.33, 0.439, -0.092, 0.484, 0.513, 0.124, 0.22, -0.146, 0.056, 0.168, 0.614, 0.602, 0.907, -0.131, 0.205, -0.087, -0.096, 0.066, -0.107, 0.255, 0.402, -0.026, -0.156]
error= 4.556
s= 1.5 coefficients: [0.191, 0.828, -0.278, 0.363, 0.681, 0.288, -0.021, 0.157, 0.667, 0.385, -0.203, 0.442, 0.651, -0.182, 0.289, -0.063, 0.379, 0.215, 0.427, -0.075, 0.087, 0.477, -0.057, 0.348, 0.774, 0.307, 0.148, 0.303, 0.43, -0.079, 0.476, 0.488, 0.127, 0.206, -0.131, 0.049, 0.141, 0.588, 0.575, 0.89, -0.115, 0.195, -0.096, -0.09, 0.043, -0.089, 0.243, 0.373, -0.028, -0.145]
error= 4.659

**B**

| dataset | 1 | 2 | 3 | 4 | Avg |
|---|---|---|---|---|---|
| Least Sq | 4.575 | 3.954 | 4.24 | 3.938 | 4.177 |
| 0.1 | 2.903 | 2.436 | 2.456 | 2.622 | 2.604 |
| 0.3 | 2.812 | 2.436 | 2.39 | 2.351 | 2.497 |
| 0.7 | 2.871 | 2.621 | 2.4 | 2.296 | 2.547 |
| 0.9 | 2.919 | 2.703 | 2.418 | 2.325 | 2.591 |
| 1.1 | 2.967 | 2.777 | 2.439 | 2.366 | 2.637 |
| 1.3 | 3.013 | 2.845 | 2.462 | 2.412 | 2.682 |
| 1.5 | 3.057 | 2.906 | 2.485 | 2.461 | 2.727 |

**C**

S=0.3 gives us the best result all throughout the database.

**D**

The avg error can be found in table above. Taking avg error across all datasets and using that to estimate error for new data points is:

1. Error in unseen data can not be predicted by taking the average of the errors from previously trained data

2. If the dataset is not iid in nature, then linear regression will never give accurate results for where it's not trained

3. Take the case of dataset 2, where only 25% of the samples were used for training. That is not an accurate estimate of the performance of the algorithm as the algorithm is not trained properly, yet, the weight of that error is the same as the weight of other errors in calculating the average.

**E**

The key points to note in this case are:

1. Two rows are independent and can be shuffled randomly, i.e. it's not time series data

2. The test and train split is judicious, i.e, at least 70% of data should be training.

3. The test set should be a representation of train set, i.e. iid distribution of the whole dataset would be of great help.