Asmt 6: Regression

Vai Suliafu, u0742607
Wednesday, April 01

1 Linear Regression & Cross-Validation

A (30 points): Solve for the coefficients `alpha` (or `alphas`) using Least Squares and Ridge Regression with $s \in \{0.2, 0.4, 0.8, 1.0, 1.2, 1.4, 1.6\}$ (i.e. s will take on one of those 7 values each time you try, say obtaining `alpha04` for $s = 0.4$). For each set of coefficients, report the error in the estimate $\hat{y}$ of y as `norm(y - X*alpha,2)`.

Each list of alpha coefficients and their corresponding errors can be viewed below in Figure 1.

B (30 points): Create three row-subsets of `X` and `Y`

- `X1 = X[:66,:]` and `Y1 = Y[:66]`
- `X2 = X[33:,:]` and `Y2 = Y[33:]`
- `X3 = np.vstack((X[:33,:], X[66:,:]))` and `Y3 = np.vstack((Y[:33], Y[66:]))]`

Repeat the above procedure on these subsets and *cross-validate* the solution on the remainder of `X` and `Y`. Specifically, learn the coefficients `alpha` using, say, `X1 and Y1` and then measure `np.norm(Y[66:] - X[66:,:] @ alpha,2)`.

The errors for each model and each round of cross validation can be viewed below in Figure 2.

C (15 points): Which approach works best (averaging the results from the three subsets): Least Squares, or for which value of s using Ridge Regression?

Per the results in Part B, it appears ridge regression with s = 0.6 works best.

D (15 points): Use the same 3 test / train splits, taking their average errors, to estimate the average squared error on each predicted data point.

What is problematic about the above estimate, especially for the best performing parameter value s?

In order to assess a models general performance, a model must be evaluated on unseen data. We have already used all of training data during the cross validation phase. To get a better estimation of the average squared error, we would need to omit some testing data from the parameter selection cross validation phase.

E (10 points): Even circumventing the issue raised in part **D**, what *assumptions* about how the data set (`X,y`) is generated are needed in an assessment based on cross-validation?

The primary assumption of cross validation is that data observations are independently and identically distributed. In reality, this may not be true and is highly dependent on the data collection process. For example, data is often collected in a particular order. Thus, observations are not randomly distributed throughout the indices. One strategy to mitigate this conflict could be random shuffling of the observations before splitting into folds. Still, there is not way to completely eliminate the risk of non iid data without understanding the data collection process.

Alphas									
	OLS	**s0.2**	**s0.4**	**s0.6**	**s0.8**	**s1.0**	**s1.2**	**s1.4**	**s1.6**
0	-0.048955	-0.042409	-0.030717	-0.023488	-0.020366	-0.020138	-0.021741	-0.024400	-0.027580
1	-9.622454	0.228028	0.291355	0.235982	0.204878	0.190021	0.181357	0.174130	0.166660
2	6.969844	2.069162	1.555846	1.301439	1.113732	0.963994	0.841651	0.740130	0.654814
3	-3.998485	-0.465681	-0.270232	-0.265812	-0.275148	-0.276245	-0.268724	-0.255489	-0.239186
4	6.770697	1.149011	0.744397	0.586894	0.493259	0.427561	0.377276	0.336854	0.303297
5	-3.167428	1.242247	1.230335	1.075410	0.918916	0.783919	0.672245	0.580684	0.505355
6	-2.872200	0.598827	0.537451	0.475625	0.410444	0.349096	0.295717	0.251070	0.214336
7	4.203035	-0.016126	-0.134546	-0.083651	-0.044739	-0.023870	-0.014722	-0.012110	-0.012770
8	-11.388664	-0.622410	-0.202053	-0.028716	0.064142	0.118312	0.150033	0.167461	0.175447
9	-12.105919	1.015919	1.024414	0.944805	0.850458	0.759428	0.677094	0.604424	0.540840
10	3.526670	0.801241	0.670730	0.585711	0.514910	0.452607	0.397723	0.349846	0.308452
11	5.453729	0.876264	0.178958	-0.062218	-0.148713	-0.178666	-0.185891	-0.183172	-0.175886
12	-0.390273	0.744895	0.777435	0.701514	0.610510	0.527563	0.456668	0.397197	0.347491
13	-0.093303	1.391399	1.224742	1.056438	0.903828	0.774264	0.667144	0.578956	0.506001
14	-14.457255	0.083549	-0.009490	-0.084994	-0.141313	-0.179038	-0.201762	-0.213473	-0.217464
15	10.256248	0.960683	0.628125	0.501939	0.422456	0.362667	0.314171	0.273629	0.239322
16	-8.419337	0.024727	0.040889	0.021027	-0.008953	-0.033879	-0.051794	-0.063613	-0.070656
17	-18.669280	0.704277	0.667089	0.564588	0.483510	0.417790	0.362673	0.315851	0.275956
18	7.925126	-0.595951	-0.192262	0.016091	0.122355	0.179427	0.208975	0.221721	0.223951
19	14.639789	0.404584	0.378793	0.421108	0.435794	0.431475	0.416508	0.395885	0.372529
20	-0.350306	-0.128553	-0.049127	-0.024024	-0.028250	-0.045358	-0.065112	-0.082511	-0.095752
21	1.177132	-0.010076	0.162853	0.179135	0.151573	0.114697	0.080548	0.052765	0.031641
22	11.750783	1.735937	1.223933	0.930401	0.742968	0.608578	0.505500	0.423732	0.357755
23	-5.426293	-0.256007	-0.139550	-0.087745	-0.071138	-0.066641	-0.065802	-0.065635	-0.065134
24	-2.071996	0.546745	0.484248	0.430189	0.396380	0.369597	0.344489	0.319699	0.295272
25	18.494670	1.408023	1.249826	1.099651	0.973599	0.862170	0.763237	0.676206	0.600273
26	-2.080082	0.198660	0.345423	0.360932	0.348666	0.327517	0.303091	0.277773	0.252901
27	-3.975287	-0.207265	0.107215	0.158012	0.166865	0.162231	0.151685	0.138658	0.125042
28	1.233356	0.507626	0.626148	0.547878	0.455298	0.373083	0.305189	0.250735	0.207556
29	-4.682879	0.204334	0.467736	0.514213	0.508560	0.487090	0.460264	0.431677	0.402899
30	4.344831	-0.880109	-0.462632	-0.265908	-0.166107	-0.109163	-0.073629	-0.050038	-0.033707
31	-10.252558	0.731839	0.600485	0.573403	0.557837	0.536109	0.508570	0.477973	0.446479
32	3.649277	0.150802	0.496767	0.601657	0.599892	0.556696	0.500881	0.444632	0.392731
33	-5.365426	-0.365474	-0.078266	0.056157	0.116421	0.139836	0.145376	0.142418	0.135597
34	-1.273302	0.416901	0.382225	0.329015	0.278593	0.235717	0.201153	0.173878	0.152374
35	9.358160	-0.512866	-0.362454	-0.296263	-0.243644	-0.196185	-0.154545	-0.119636	-0.091383
36	-5.762584	0.378716	0.237732	0.166513	0.122182	0.091293	0.068999	0.052734	0.040769
37	-12.724886	0.327956	0.481052	0.407354	0.306094	0.216030	0.144033	0.089090	0.048266
38	7.830288	1.413594	1.055246	0.856471	0.730873	0.643753	0.577423	0.523132	0.476542
39	-12.978475	1.256007	0.993010	0.833559	0.726095	0.645827	0.581015	0.526018	0.478011
40	-1.366504	0.751464	0.815903	0.883400	0.910269	0.903270	0.873354	0.829705	0.778994
41	-0.539813	-0.916085	-0.397023	-0.249809	-0.178378	-0.131588	-0.096628	-0.069286	-0.047724
42	-8.565273	0.683123	0.414128	0.316983	0.259377	0.220232	0.193137	0.173894	0.159509
43	-2.396308	1.280183	0.521158	0.184087	0.027721	-0.044651	-0.076287	-0.087761	-0.089254
44	-3.941062	-0.687247	-0.326694	-0.147945	-0.082487	-0.060018	-0.052864	-0.050889	-0.050404
45	13.653107	0.664733	0.551659	0.375315	0.231329	0.129043	0.060371	0.015515	-0.013258
46	-13.574655	-0.686947	-0.349981	-0.249781	-0.178518	-0.122303	-0.079660	-0.048598	-0.026687
47	-2.253550	-0.120709	0.184792	0.264737	0.275214	0.259689	0.235252	0.209045	0.184041
48	15.619195	1.406458	1.007889	0.751736	0.587044	0.473734	0.391554	0.329583	0.281445
49	-2.212480	0.028415	0.057272	0.038240	0.014017	-0.005058	-0.017337	-0.024001	-0.026736
50	-5.975303	-0.575643	-0.349152	-0.271706	-0.214511	-0.172020	-0.140677	-0.117322	-0.099603

Errors								
OLS	**s0.2**	**s0.4**	**s0.6**	**s0.8**	**s1.0**	**s1.2**	**s1.4**	**s1.6**
3.456630	3.676513	3.823765	3.995933	4.197474	4.422363	4.660180	4.901718	5.140233

Figure 1: Model Results

Cross Validation Errors				
Model	**Error0**	**Error1**	**Error2**	**Avg**
OLS	6.290761	4.937106	4.443744	5.223870
s0.2	4.094690	3.229181	3.276312	3.533394
s0.4	3.606376	3.083150	2.775584	3.155036
s0.6	3.432230	3.210792	2.539067	3.060696
s0.8	3.400994	3.388169	2.471685	3.086949
s1.0	3.435351	3.571457	2.502003	3.169604
s1.2	3.499165	3.747140	2.586945	3.277750
s1.4	3.575567	3.910420	2.700000	3.395329
s1.6	3.656301	4.059809	2.824578	3.513563

Figure 2: Cross Validation Errors