Report-HW1

(Abhishek Singhal)

Q1)

a) The criteria for selecting a variable, as node in the tree, is minimizing the mean square error of the two splitting regions. We choose a variable and try all possible splits in the datum. Then we do this process for all variables. We select the variable(j) and split(s) s.t. the mean square error of the region to be divided, is minimum.
For tree of depth=1, it is optimal as we look at all the possible cases.
For trees of depth>1, we do recursive splitting. As naïve splitting is computationally expensive. No, it is not optimal as we have not searched all possible trees.

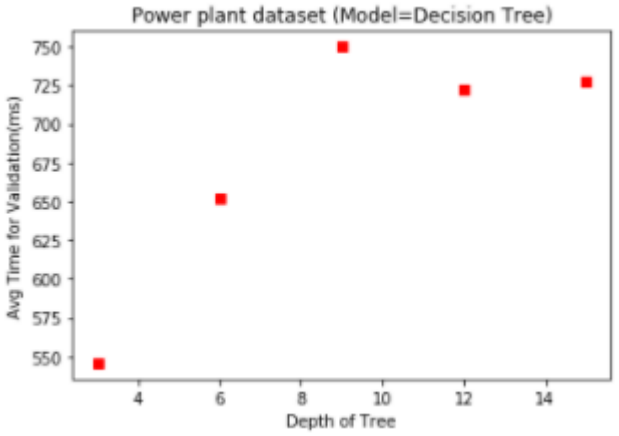
b) **Power Plant**

model chosen: max_depth=9, criterion='mae'

out of sample error=0.160256966417

full training error=0.118336262681

Out of Sample error is close to Full training error.



Model (depth)	Error1	Error2	Error3	Error4	Error5	Avg. Error
3	0.21363307694986075	0.21213404111498263	0.21483569790940768	0.21206802787456452	0.2074905393728223	0.21203227664432758
6	0.17165978655988864	0.17703962125435538	0.1741445038327526	0.16561086898954705	0.1742725013937816	0.17254545640605437
9	0.15516345369080786	0.16125527909407672	0.16353107804878048	0.16053877177700346	0.16079624947735191	**0.16025696641760409**
12	0.16517684087743736	0.15946469616724745	0.16460551567944243	0.16203833275261331	0.1554986247386759 3	0.16135680204308328
15	0.16459260759052918	0.17032658989547039	0.16705502473867595	0.16769339268292685	0.1700053177703486	0.16793458653552742

c) **Indoor Localization**

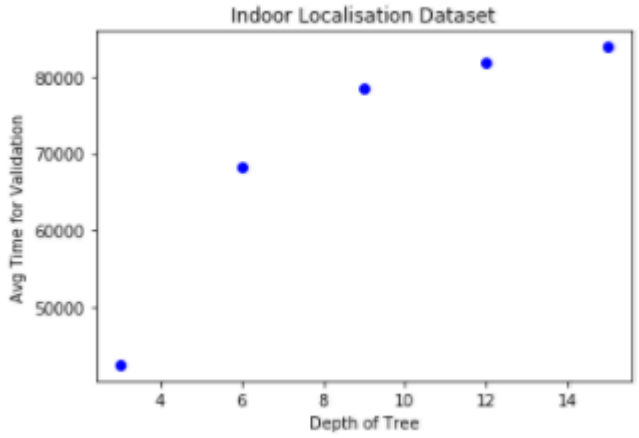
model chosen: max_depth=15

out of sample error=9.0142990917955039

full training error=7.48820330784

Out of Sample error is not very close to Full training error.

Model (depth)	Error1	Error2	Error3	Error4	Error5	Avg. Error
3	40.316933000376132	40.593297627382215	40.015660284926007	40.679415554176074	41.302495941183849	40.581560481608 84
6	27.481255895687 06	26.509645056920 76	28.268997951718084	27.535988266804615	26.338258458490092	27.226829125924 12
9	17.423034897881145	18.091903612274322	17.485970172686233	17.789486359794331	18.237538454602461	17.805586699447694
12	12.233991034435306	11.777981722291875	12.882466373651869	12.241183331640332	12.937726133182846	12.414669719023996
15	8.8434292245486468	9.0379293320586758	9.0100572337597189	9.0826730540506642	9.0974066145598194	**9.0142990917955039**



Q2)

a) N samples. Each chunk of size =M.
k=N/M is the number of subsets of training dataset.
We will train (N-M) samples for k times.
Total samples trained = (N-M)X(N/M)
Time taken ~ $N^2/M - N$
Complexity ~ $O(N^2/M)$

If M=5 , complexity is $O(N^2/5)$.

If M=N/2, then running time is linear. O(2N)

b) When we make M small, we actually increase k in k-fold cross validation. With this, the bias of estimated error will be small. Or we can say estimator can be fairly accurate.

Q3.KNN

a) Power Plant

model chosen n=3

out of sample error=0.22130739350784454

full training error= 0.361530806

Out of Sample error is close to Full training error.

Model (n_neighbors)	Error1	Error2	Error3	Error4	Error5	Avg. Error
3	0.5437207207520891	0.3709510025542784 3	0.1918652442328534 2	8.9421203028509396e-16	8.8916858516244428e-16	**0.2213073939350784454**
5	0.5396401577994429 7	0.4293734108672935 8	0.3303407982350208 8	0.2317099345758578 4	0.1150972827201783 6	0.3292323168395586 7
10	0.5813528027855153 8	0.4855198153256704 6	0.4240396581049698	0.3735414182720780 4	0.3188391844202899 1	0.4366585757817046 9
20	0.6084864923398328 7	0.5528452249912921 8	0.5021451544356712	0.4541370968385298 2	0.4223578424470457 7	0.5079943622104743 4
25	0.5888856829247911 5	0.5564655495228143 5	0.5153393037900603 2	0.4754626101724437 9	0.4463651015273132 9	0.5165036495874846 3

b) Indoor Localization

model chosen n=3

out of sample error=1.60727811

full training error=2.2239872

Out of Sample error is not very close to Full training error.

Model (n_neighbors)	Error1	Error2	Error3	Error4	Error5	Avg. Error
3	3.4594528604563703	2.0966570466399213	1.3172279068795469	0.5647910600522492 2	0.5982617030144983 3	**1.6072781154085174**
5	3.8368866274322975	2.6133712489343037	2.0162536138092459	1.4211278897680260	1.0011745483121846	2.1777627856512116
10	4.7261048981569713	3.3950221586822993	2.7334935904413640	2.2961816310783743	1.9353148300747409	3.0172234216 8675
20	5.79761992	4.402627948	3.656813968	3.156505595	2.814353202	3.965584128

	70310937	8622118	1183669	3714785	7662208	4298745
25	6.06134445 60481442	4.814073318 698596	3.982288090 134583	3.468855283 3918527	3.090503718 6056132	4.283412973 375758

Q4. Linear Model

a) The purpose of penalties in Ridge and Lasso regression is to control capacity. In several cases, all the attributes (p) of input data are not useful. The penalties makes coefficient β_i small for **i**th attribute of **X.** The larger the penalty (α), the smaller the value of β.
In ridge regression, β_s are made small but do not converge to zero. That is, no input attribute is ignored.
In lasso regression, β can be zero. This a selection of attributes occur.

b) Power Plant

Model chosen: alpha=0.01 (Lasso)

Out of sample error=0.18776822

Full training error=0.190764

Out of Sample error is close to Full training error.

Lasso (alpha)	**Error1**	**Error2**	**Error3**	**Error4**	**Error5**	**Avg. Error**
10^-6	0.189783 5245755 4071	0.193985380 69880425	0.191860269 21584834	0.190800045 36720838	0.190678764 92417468	0.19142159695 631528
10^-4	0.191755 0793772 2855	0.188871314 55626152	0.190059895 67954601	0.191086838 19736783	0.190679416 28063609	0.19049050881 8208
10^-2	0.184438 3544036 9816	0.186309279 54994064	0.187578892 74669534	0.189749688 57416435	0.190764887 68205507	**0.18776822059 131071**
1	0.266660 3571793 485	0.264340012 11203639	0.264054867 66376685	0.262189306 16776909	0.262453287 77656507	0.26393956617 989722
10	0.623296 2057636 6584	0.621489476 33336626	0.623656002 7055742	0.626516512 92788568	0.627689496 95425558	0.62452953893 694951

Ridge(alpha)	**Error1**	**Error2**	**Error3**	**Error4**	**Error5**	**Avg. Error**
10^-6	0.191957169 5636817	0.193288345 00600034	0.193177406 71953548	0.191956737 70276014	0.190678758 33731485	0.192211683 46585851
10^-4	0.187924906 46272467	0.191037361 19286774	0.190946324 74772946	0.189898398 29986058	0.190678758 34024595	0.190097149 80868567

10^{-2}	0.19414361825633891	0.19060084353701598	0.19089328768394551	0.19040849808191673	0.19067875863337849	0.19134500123851911
1	0.18650679327417996	0.18803605998555781	0.18933126988930352	0.19125137117950047	0.19067878794657855	0.18916085645502406
10	0.18752966366219898	0.18996838054764123	0.19006532963670314	0.19153979273873339	0.1906790544274605	0.18995644420254745

c) Indoor Localization

Model chosen: alpha=10 (Ridge)

Out of sample error= 18.970032206799651

Full training error=18.6044577636

Out of Sample error is very close to Full training error.

Lasso (alpha)	**Error1**	**Error2**	**Error3**	**Error4**	**Error5**	**Avg. Error**
10^{-6}	20.47708049290879	20.634091295324243	20.058269918698816	19.878756182890246	19.931362865996984	20.195912151163817
10^{-4}	20.151478357417044	19.455128957219078	20.460186593805226	20.777356272540423	20.170559352385183	20.202941906673392
10^{-2}	19.94083987632904	20.438487299950857	20.041168766019013	20.191730583255904	20.167033793553983	20.155852063821762
1	21.44192350345271	21.341933750531748	20.772501390215557	20.991333011626896	21.237534691471264	21.157045269459637
10	35.476969491427617	35.613805855557366	35.329753256371198	34.982710551830351	35.522383882598582	35.385124607557024

Ridge (alpha)	**Error1**	**Error2**	**Error3**	**Error4**	**Error5**	**Avg. Error**
10^{-6}	19.204266483693473	19.31476104522444	18.488004720819927	19.133594596324411	18.753334895918702	18.97879234839619
10^{-4}	19.006484640804544	18.824139184243098	19.038591711638318	18.980103339348965	19.029139946428252	18.975691764492643
10^{-2}	19.039378468528934	18.920775806994008	19.192728377638382	18.986285827973809	18.764728417721177	18.980779379771216
1	18.941915583117915	18.815454068655015	19.136983595819675	19.079018849427069	19.039240045015809	19.002522428407097
10	18.874853302804265	18.814607148918189	19.084361717063626	18.752438823573648	19.323900041638517	**18.970032206799651**
