

CSE881 HW3

Nan Cao, A52871775

Oct 4th, 2016

Problem 1

(a)

(i)

Please see the codes in the last part of this pdf.

(ii)

Please see the codes in the last part of this pdf.

(iii)

Use the codes in "Q1.m", we can mathematically find the following combination of non-linearly independent column pairs.

(4, 35); (4, 36); (4, 37); (22, 35); (22, 36); (22, 37); (30, 35); (30, 36); (30, 37); (31, 35); (31, 36); (31, 37); (32, 35); (32, 36); (32, 37); (33, 35); (33, 36); (33, 37); (34, 35); (34, 36); (34, 37); (35, 37); (35, 38); (35, 39); (35, 40); (35, 41); (35, 42); (35, 47); (35, 48); (36, 37); (36, 38); (36, 39); (36, 40); (36, 41); (36, 42); (36, 47); (36, 48); (37, 38); (37, 39); (37, 40); (37, 41); (37, 42); (37, 47); (37, 48).

Only group 4 that contains both of the columns in a pair.

Following are the names of columns:

30. weekday is monday: Was the article published on a Monday?

31. weekday is tuesday: Was the article published on a Tuesday?

32. weekday is wednesday: Was the article published on a Wednesday?

33. weekday is thursday: Was the article published on a Thursday?

34. weekday is friday: Was the article published on a Friday?

35. weekday is saturday: Was the article published on a Saturday?

36. weekday is sunday: Was the article published on a Sunday?

37. is weekend: Was the article published on the weekend?

Mathematically, all of following are correct answers:

(30, 35); (30, 36); (30, 37); (31, 35); (31, 36); (31, 37); (32, 35); (32, 36); (32, 37); (33, 35); (33, 36); (33, 37); (34, 35); (34, 36); (34, 37); (35, 37); (36, 37)

And I picked (36,37) as my answer. Firstly, it's mathematically correct. Secondly, the remaining variables show us an aesthetic feeling of mathematics.

iv)

Please see the codes in the last part of this pdf.

(b)

The correlation of MLR is -0.0339

The correlation of lasso is -0.229

The Top 10

Top 10 Para of MLR $1.0e+06$ 4.5335, 4.1681, 2.6387, 0.2139, 0.2008, 0.1887, 0.1764, 0.1423, 0.000, 0.0000

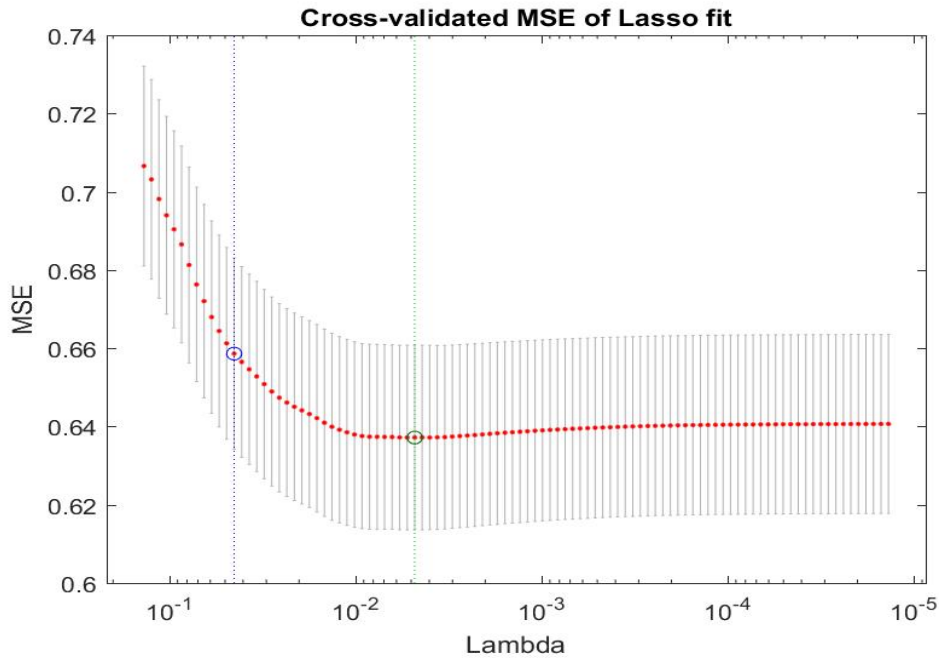


Figure 1: Problem 2b-lasso

Col number=45, 46, 4, 40, 39, 36, 38, 37, 57, 16

Top 10 Para of Lass=0.0814, 0.0736, 0.0433, 0.0424, 0.0327, 0.0235, 0.0192, 0.01, 0, 0.0159, 0.0039

Col number=35, 15, 27, 16, 26, 56, 38, 14, 13, 30

(c) The correlation of kernel is -0.082

(d) all 3 correlations (MLR lasso kernel) remain the same.

Problem 2

(a)

An outlier may have an unusual X or Y. An outlier is an influential point that have great effects on the slope of a regression model. If there are more than 1 outlier, their effect may be quite small, if the effect can be canceled by other's. For example:

x1=1, 2, 3, 4, 5, 6 y1=2.1, 3.9, 6.1, 8.1, 10.2, 12.3

x2=1, 2, 3, 4, 5, 6, 7 y2=2.1, 3.9, 6.1, 8.1, 10.2, 12.3, 30

x3=1, 2, 3, 4, 5, 6, 7 y3=2.1, 3.9, 6.1, 8.1, 10.2, 12.3, 3

(b)

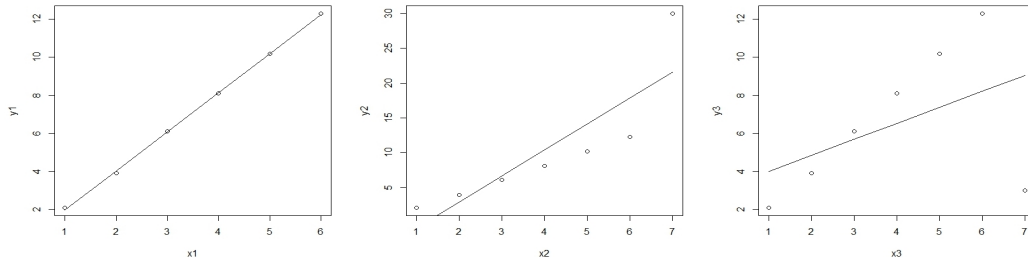


Figure 2: Problem 3c-2

$$\begin{aligned}
 & \min \sum a_i (y_i - w^T x_i)^2 + \lambda \|w\| \\
 & = \min (\sqrt{a_i} y_i - w^T \sqrt{a_i} x_i) + \lambda \|w\| \\
 & \text{so let } X^* \text{ denote } (\sqrt{a_1} x_1, \sqrt{a_2} x_2, \dots, \sqrt{a_n} x_n) \\
 & \text{let } Y^* \text{ denote } (\sqrt{a_1} y_1, \sqrt{a_2} y_2, \dots, \sqrt{a_n} y_n) \\
 & w = [X^{*T} X^* + \lambda I]^{-1} X^{*T} y^*
 \end{aligned}$$

Problem 3

$$\begin{aligned}
 E(\text{Sample}) &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \\
 &= 1
 \end{aligned}$$

$$\begin{aligned}
 E(A) &= \frac{1}{2} \left(-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) \\
 &= 0.9183
 \end{aligned}$$

$$\begin{aligned}
 E(B) &= \frac{1}{3} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + \frac{1}{12} \left(-\frac{5}{5} \log_5 \frac{5}{5} - \frac{0}{5} \log_2 \frac{0}{0} \right) + \frac{7}{12} \left(-\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} \right) \\
 &= 0.9081
 \end{aligned}$$

B is better because 1-E(B) is bigger.

Codes

```

1 % NAN CAO CSE881 HW4
2 % set dir in nan's win lap
3 cd C:\Users\nan66\Dropbox\CSE881\HW4\;
4 %set dir in nan's linux lap
5 % cd /home/nan/Dropbox/CSE881/HW4;
6 %%%%
7 %(a)
8 %%%%

```

```

9 ONP=csvread('OnlineNewsPopularity.csv',1,0);
10 Pre=ONP(:,1:58);%col 1~58 as Predict Variable
11 Tar=ONP(:,59);%col 59 as the Response/Target Variable
12 Tar=log(Tar);
13 Tar=Tar(:);
14 Tar(Tar==Inf)=0;
15 rPre=rank(Pre);%rank of Predict Matrix
16 a=0;%set 3 initial variables
17 b=0;
18 c=0;
19 for i=1:57
20 for j=(i+1):58 % try all the possible combination of 2 columns
21 A=Pre;
22 A(:,j)=[];%remove the one with greater column number first
23 A(:,i)=[];
24 rA=rank(A);
25 if rA==56;
26 a=a+1;%use a to calculate the number of the possible ij
27 b(a)=i;%save possible i
28 c(a)=j;%save possible j
29 end
30 end
31 end
32 d=[b;c]';%show all the possible combination of removing
33 d
34 %Remove column 36(is it Su) and column 37(is it Weekend);
35 Pre1=Pre;
36 Rm=[36,37];
37 Pre1=Pre;
38 Pre1(:,Rm)=[];
39 rank(Pre1)
40 %%%
41 %(b)
42 %%%
43 Len=length(Tar);
44 TrainPre=Pre1(1:2000,:);
45 TrainTar=Tar(1:2000);
46 TestPre=Pre1(2001:Len,:);
47 TestTar=Tar(2001:Len);
48 sTrainPre=zscore(TrainPre);%standardize
49 sTestPre=zscore(TestPre);
50 % TrainTar=zscore(TrainTar);%standardize
51 % TestTar=zscore(TestTar);
52 TrainPre1 = [TrainPre ones(2000,1)]; % add a column of 1s
53 sTrainPre1 = [sTrainPre ones(2000,1)];
54 TestPre1 = [TestPre ones(Len-2000,1)];
55 sTestPre1 = [sTestPre ones(Len-2000,1)];
56 %MLR
57 wMLR=regress(TrainTar,sTrainPre1);
58 %lasso
59 [w,stats]=lasso(sTrainPre1,TrainTar,'Alpha',1,'CV',10);
60 Figure1=lassoPlot(w,stats,'PlotType','CV');
61 saveas(Figure1,'Q1b','jpeg');
62 wbest=w(:,stats.Index1SE');

```

```

63
64 %cal top 10
65 [a1,b1]=sort(abs(wMLR),'descend');
66 a1(1:10)
67 b1(1:10)
68 [a2,b2]=sort(abs(wbest),'descend');
69 a2(1:10)
70 b2(1:10)
71 %cal predicted values
72 MLRPreVal=sTestPre1*wMLR;
73 CorMLR=corr(MLRPreVal,TestTar)
74 LasPreVal=sTestPre1*wbest;
75 CorLas=corr(LasPreVal,TestTar)
76 %%%
77 %(c)
78 %%%
79 Len=length(Tar);
80 TrainPre=Pre1(1:2000,:);
81 TrainTar=Tar(1:2000);
82 TestPre=Pre1(2001:Len,:);
83 TestTar=Tar(2001:Len);
84 lambda=0.001
85 %calculate alpha for function 'applyKernel'
86 dist1 = pdist(sTrainPre); % calculate distance between every pair of points
87 dist1 = squareform(dist1); % convert to square matrix
88 K = exp(-(1e-7) * dist1.^2); % mu: kernel parameter (specified by user)
89 alpha = inv(K + lambda*eye(2000)) * TrainTar;
90 pred = applyKernel(sTrainPre,TestPre, alpha,1e-7);
91 CorKer=corr(pred,TestTar)
92
93
94
95
96
97 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
98 %(d)
99 %%%
100 Pre2=Pre1;
101 Pre2(:,18:29)=log(Pre2(:,18:29));
102 Pre2(Pre2==Inf)=0;
103 Len=length(Tar);
104 TrainPre=Pre2(1:2000,:);
105 TrainTar=Tar(1:2000);
106 TestPre=Pre2(2001:Len,:);
107 TestTar=Tar(2001:Len);
108 sTrainPre=zscore(TrainPre);%standardize
109 sTestPre=zscore(TestPre);
110 TrainPre1 = [TrainPre ones(2000,1)]; % add a column of 1s
111 sTrainPre1 = [sTrainPre ones(2000,1)];
112 TestPre1 = [sTestPre ones(Len-2000,1)];
113 sTestPre1 = [sTestPre ones(Len-2000,1)];
114 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
115 %MLR
116 wMLR=regress(TrainTar,sTrainPre1);

```

```

117 %lasso
118 [w,stats]=lasso(sTrainPre1,TrainTar,'Alpha',1,'CV',10);
119 Figure1=lassoPlot(w,stats,'PlotType','CV');
120 saveas(Figure1,'Q1d','jpeg');
121 wbest=w(:,stats.Index1SE');
122
123 % %cal top 10
124 % [a1,b1]=sort(abs(wMLR),'descend');
125 % a1(1:10)
126 % b1(1:10)
127 % [a2,b2]=sort(abs(wbest),'descend');
128 % a2(1:10)
129 % b2(1:10)
130 %cal predicted values
131 MLRPreVal=sTestPre1*wMLR;
132 CorMLR2=corr(MLRPreVal,TestTar)
133 LasPreVal=sTestPre1*wbest;
134 CorLas2=corr(LasPreVal,TestTar)
135 %%%
136 %(c)
137 %%%
138 Len=length(Tar);
139 TrainPre=Pre1(1:2000,:);
140 TrainTar=Tar(1:2000);
141 TestPre=Pre1(2001:Len,:);
142 TestTar=Tar(2001:Len);
143 lambda=0.1;
144 %calculate alpha for function 'applyKernel'
145 dist1 = pdist(sTrainPre); % calculate distance between every pair of points
146 dist1 = squareform(dist1); % convert to square matrix
147 K = exp(-(1e-7) * dist1.^2); % mu: kernel parameter (specified by user)
148 alpha = inv(K + lambda*eye(2000)) * TrainTar;
149 pred = applyKernel(sTrainPre,TestPre,alpha,1e-7);
150 CorKer2=corr(pred,TestTar)

```