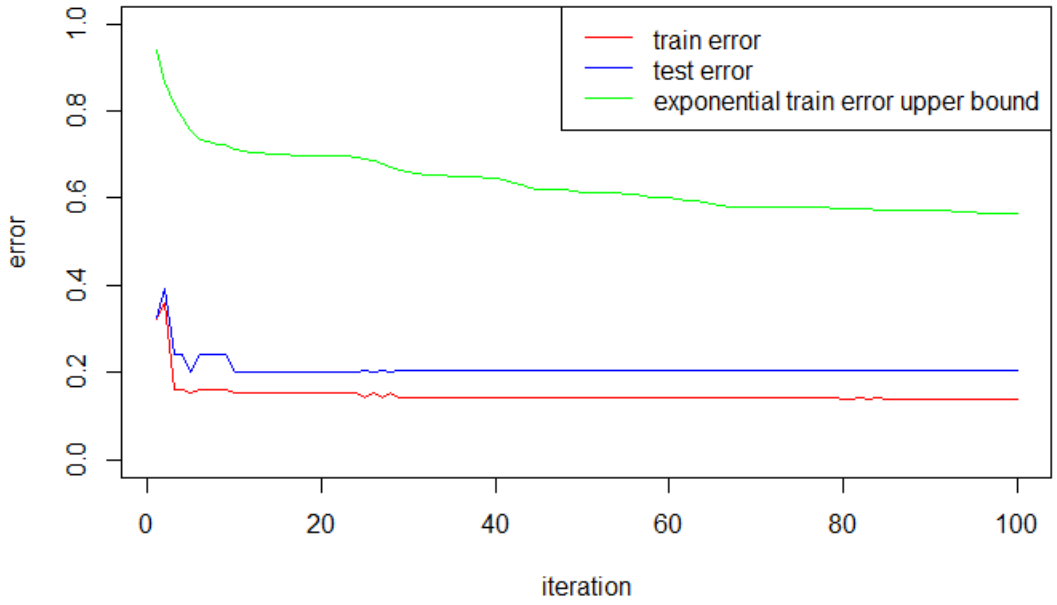
STAT542HW5

Xiruo Li (xiruoli2)

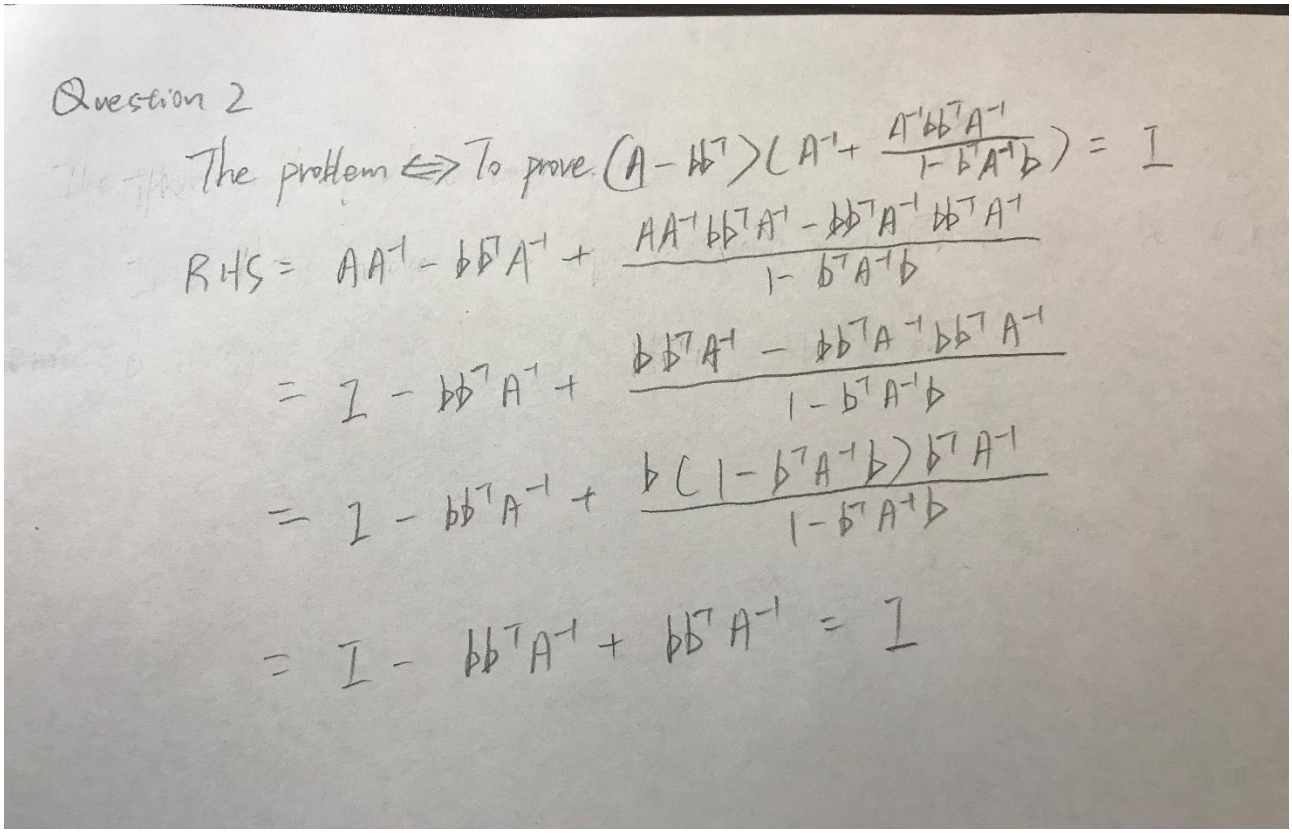
Question 1



The train error and test error decrease at the first several iterations, then the error become stable (around 0.2). And test error is higher than train error. Thus, there is overfitting in the model.

Besides, exponential error upper bound decrease exponentially and it's always higher than train error.

Question2



Question3

(1)

These arts are from 4 museums, so I want to divide them into 4 clusters that have distinct features of art in each museum.

(2)

Firstly, I remove 7 variables:

Id: all of id are unique;

artist: same function as artist id;

dateText, thumbnailCopyright, thumbnailUrl, Url: these 4 variables contain too much noisy and useless information for cluster;

depth: too many NA;

Then, I convert some character variable into categorical variable, and make them as factor variables:

artistRole: artist=1, others=0;

artistID: the most frequent case: 558=1, others=0;

title: first 5 frequent cases have the same meaning: "no title", which I make as 0. others=1;

medium: the most 5 frequent cases named as 1-5, others =6;

creditLine: the most frequent case=1, others=0;

dimensions: support=1, others=0;

Next, I normalize the numeric variables and substitute NA as 0 :

year, acquistionYear, width, height.

Finally, I also do other cleaning:

accession_number: extract the prefix (only letter left), then factorize;

units: mm=1, NA=0, then factorize;

inscription: date inscribed=1,NA=0, then factorize.

The final data contain 4 numeric variables and 9 factor variables.

Numeric: year, acquistionYear, width, height

Factor: artisRole, artisID, title, medium, creditLine, dimensions, accession_number, units, inscription.

(3)

I use k-prototypes algorithm.

Advantage: it combines k-means and k-modes algorithm. It's suitable for mix data in this problem (numerical variable+ categorical variables).

Disadvantage: we need to find the weights of categorical and numerical variables, to make a tradeoff for these two types of variables.

(4)

Summary table of the result (number of cluster=4):

--------Summary descriptives table by 'cluster'---------

	1 N=13204	2 N=27786	3 N=2986	4 N=25225
accession_number:				
A	38 (0.29%)	1621 (5.83%)	1 (0.03%)	78 (0.31%)
AR	0 (0.00%)	1078 (3.88%)	99 (3.32%)	0 (0.00%)
D	12561 (95.1%)	0 (0.00%)	1 (0.03%)	25082 (99.4%)
N	590 (4.47%)	2491 (8.96%)	653 (21.9%)	63 (0.25%)
P	0 (0.00%)	10869 (39.1%)	367 (12.3%)	0 (0.00%)
T	15 (0.11%)	11727 (42.2%)	1865 (62.5%)	2 (0.01%)
artistRole:				
0	32 (0.24%)	2240 (8.06%)	11 (0.37%)	11 (0.04%)
1	13172 (99.8%)	25546 (91.9%)	2975 (99.6%)	25214 (100.0%)
artistId:				
0	535 (4.05%)	26160 (94.1%)	2937 (98.4%)	180 (0.71%)
1	12669 (95.9%)	1626 (5.85%)	49 (1.64%)	25045 (99.3%)
title:				
0	5897 (44.7%)	5205 (18.7%)	93 (3.11%)	403 (1.60%)
1	7307 (55.3%)	22581 (81.3%)	2893 (96.9%)	24822 (98.4%)
medium:				
1	0 (0.00%)	1287 (4.63%)	7 (0.23%)	24873 (98.6%)
2	392 (2.97%)	1755 (6.32%)	1235 (41.4%)	1 (0.00%)
3	0 (0.00%)	2928 (10.5%)	56 (1.88%)	0 (0.00%)
4	0 (0.00%)	2695 (9.70%)	26 (0.87%)	0 (0.00%)
5	1004 (7.60%)	532 (1.91%)	3 (0.10%)	351 (1.39%)
6	11808 (89.4%)	18589 (66.9%)	1659 (55.6%)	0 (0.00%)
creditLine:				
0	438 (3.32%)	27786 (100%)	2941 (98.5%)	143 (0.57%)
1	12766 (96.7%)	0 (0.00%)	45 (1.51%)	25082 (99.4%)
year	-0.71 (0.27)	0.78 (0.91)	1.04 (0.97)	-0.61 (0.18)
acquisitionYear	-0.84 (0.12)	1.07 (0.47)	0.93 (0.63)	-0.85 (0.08)
dimensions:				
0	207 (1.57%)	18490 (66.5%)	1123 (37.6%)	62 (0.25%)
1	12997 (98.4%)	9296 (33.5%)	1863 (62.4%)	25163 (99.8%)
width	-0.30 (0.34)	0.17 (0.60)	3.44 (1.98)	-0.44 (0.13)
height	-0.21 (0.33)	0.11 (0.47)	3.02 (2.97)	-0.36 (0.12)
units:				
0	204 (1.54%)	3076 (11.1%)	0 (0.00%)	61 (0.24%)
1	13000 (98.5%)	24710 (88.9%)	2986 (100%)	25164 (99.8%)
inscription:				
0	13097 (99.2%)	22416 (80.7%)	2209 (74.0%)	25173 (99.8%)
1	107 (0.81%)	5370 (19.3%)	777 (26.0%)	52 (0.21%)

accession number: most of A, AR, N, P, T belongs to cluster 2, most of D belongs to cluster 1 and 4.

artistRole: most of non-artist belongs to cluster 2.

artistID: most of non-558 number ID belongs to cluster 2.

title: most of no-title belongs to cluster 1 and 2.

medium: most of medium 1 belong to cluster 4, most of medium 2 belong to cluster 2 and 3, most of medium 3, 4 belong to cluster 2, most of medium 5 belong to cluster 1 and 2.

year, acquisitionYear, width, height : they are low in cluster 1 and 4, they are high in cluster 2 and 3.

creditLine: the most frequent credit line belongs to cluster 1 and 4 mostly.

inscription: most of “data inscribed” is in cluster 2.