CS534 MACHINE LEARNING

Da Lin 933334812
lind2@oregonstate.edu

Part0:

a.

The ID feature was removed, please check it in the following code named “part0.py”. ID feature make no sense for the learning model, it only presents the number of the house, remove it will increase the accuracy of the result.

b.

The date feature was split, please check it in the following code named “part0.py”. After we split date, we can use “year” or “month” to set a range for price prediction, it will help our decision making.

c.

	mean	std	range		Categorical	Percentage
dummy	1	0	0	Waterfront	0	99.3099
month	6.5924	3.1113	11		1	0.6901
day	15.8021	8.6213	30	Condition	1	0.12
year	2014.319	0.4659	1		2	0.7504
bedrooms	3.3752	0.9432	32		3	65.3227
bathrooms	2.1189	0.7651	7.25		4	25.6928
sqft_living	2080.223	911.2888	9520		5	8.1141
sqft_lot	15089.2	41201.83	1650787	Grade	4	0.1001
floors	1.5037	0.5426	2.5		5	1.041
waterfront	0.007	0.0834	1		6	9.3293
view	0.2294	0.7559	4		7	41.3313
condition	3.4091	0.6536	4		8	28.3984
grade	7.6732	1.1800	9		9	11.8218
sqft_above	1793.099	830.8239	8490		10	5.4655
sqft_basement	287.1239	434.9835	2720		11	2.0921
yr_built	1971.125	29.4791	115		12	0.3804
yr_renovated	81.2267	394.3601	2015		13	0.04
zipcode	98078.29	53.5157	198			
lat	47.5598	0.1386	0.6217			
long	-122.213	0.1414	1.195			
sqft_living15	1994.326	691.8657	5650			
sqft_lot15	12746.32	28239.83	870540			
price	5.3853	3.5737	68.08			

d.

Most of the features are useful for the task, such as bedrooms, bathrooms, sqft_basement, sqft_living, and floors etc. These kinds of features will directly affect the price of houses. Others like waterfront will also affect the price but not directly.

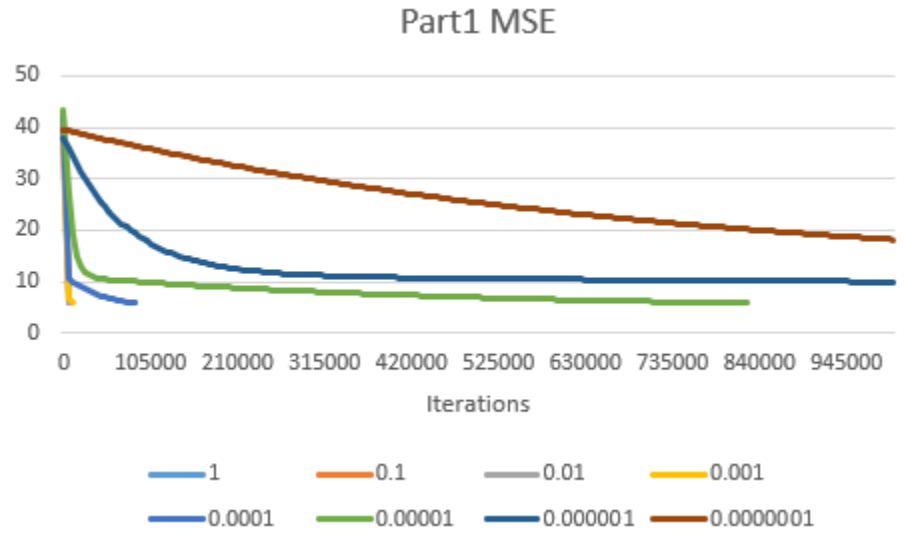
e.

All features were normalized, please check it in the following code named “part0.py”

Part1:

a.

Curve shows below. The best learning rate is 10^{-5}. Larger learning rate will cause the task end too early and lesser learning rate will cause the task too slow. Both of them can not obtain the optimal solution in a limited iteration. (In this case, iteration = 1000000)



b.

Learning rate	Train	Dev	Iteration
10^{0}	nan	nan	0(nan)
10^{-1}	5.796977864	6.082538	82
10^{-2}	5.80575021	6.077157	827
10^{-3}	5.800954682	6.088239	8232
10^{-4}	5.811817478	6.140678	84206
10^{-5}	5.825826676	6.119119	823146
10^{-6}	9.988410734	10.071861	1000000(limit)
10^{-7}	18.21746003	18.101539	1000000(limit)

Although the MSE for different learning rate seems similar (from 10^{-1} to 10^{-5}), the weight calculation accuracy will be affect if the task end too fast, so we need to choose the learning rate that cause the task end close the limited iteration, which means 10^{-5}. Others like 10^{-6} and 10^{-7} can not converge in a limited iteration.

c.

Features	Weight	Features	Weight
dummy	0.233110425884702	condition	0.705433401449338
month	-0.101833065885195	grade	3.09173717185788
day	-0.345700086403732	sqft_above	2.29036657131181
year	0.0827327076203772	sqft_basement	1.79137684237252
bedrooms	0.244109984814431	yr_built	-0.493923344079206
bathrooms	1.81844581609526	yr_renovated	0.723429665441942
sqft_living	2.24255828614358	zipcode	-0.323703419546284
sqft_lot	0.0246897462561008	lat	2.43835495342319
floors	1.33888089994827	long	-0.067712839198063
waterfront	0.671718935825266	sqft_living15	2.69599865163367
view	2.63699156651442	sqft_lot15	-0.033849811807896

The highest weight is “grade”, it is different to my assumption, that prove I am wrong in the expecting. Moreover, “lat”, “view”, “sqft_living15” also have a high weight, there are also strongly affect the price as “grade”.

Part2:

a.

Learning rate	Train MSE	Dev MSE
100	nan	nan
10	nan	nan
10^{-1}	nan	nan
10^{-2}	nan	nan
10^{-3}	nan	nan
10^{-4}	nan	nan
10^{-10}	869.360986	43.234230

Iteration	Train MSE
0	11283286.425909
2000	32596.009475
4000	9351.614042
6000	3167.007090
8000	1429.781704
10000	869.360986

If the data was non-normalized, in large learning rate, the result can not converge. When I try up to 10^{-10}, the task begins to converge. Normalized data set will be helpful for learning model. Because the number of features will be balanced and increase the accuracy of the learning result.