Assignment 1&2 Take2 Report

Name: Abdelrahman Ashraf Adel

ID: 40-5927

This is a report on the first and the second assignments after applying techniques from lecture 5 and 6 to them, while changing the data set of the first one to the new house data set acquired.

The report is divided into 4 phases, starting with the data sets description, the techniques applied to the data, analysis of the results and finally the conclusion.

Data set description and observations:

We have two different data sets each used for different models. The first data set is the house price data-set which was utilized in the linear regression model in assignment one. The second one is the student exam scores data set used in the second assignment which applied logistic regression models.

- House price data-set is a large one composed of 18 thousand samples and 21 features. OfCourse not all the data contribute have the same contribution to the prediction of the label as each other, so to make it simpler some features were dropped after a correlation matrix was conducted to measure the correlation between all the features and the label "Price". Then any feature with a correlation less then 0.4 was dropped.

- Student exam scores data set composed of 100 samples and 3 features. At which the last column was taken as the label.

There are two approaches for dividing the data sets, the first one is to divide it to 70% training set and 30% testing set and the other approach which was applied is to divide the data to 3 sets, data for training which takes 60%, data for cross validation 20% and the last 20% is for testing.

Techniques applied to the data:

- **Data normalization**

The two data sets were normalized before applying any techniques to them to reduce their redundancy and improve the data integrity for better results.

- **Random sampling**

For the first data set using the linear regression, six models were created differing in their degrees. Which varied from one to six. The learning rate and the number of iterations parameters were fixed for all models as the learning rate was standardized at 0.1 and the iterations for 400 same as the old assignments. In the first-degree model all the features are first degree. The second all the features are first except the sqft living feature which is second degree. The third same as the second but with the sqft above feature third degree. The fourth goes with the same rhythm and have the sqft living 15 with the 4th degree. The fifth have the grade feature with 5th degree. And the last with the bathroom feature in 6th degree.

The models trained and tested with two different approaches. The first-time using shuffling only and the other using random sampling and shuffling.

For the second data set using logistic regression, three models were created also differing in their degrees from one to three. The models were trained and tested twice with the same the data house price data set. The learning rate and the number of iterations parameters were fixed for all models as the learning rate was standardized at 0.1 and the iterations for 400 same as the old assignments. In the first-degree model all the features were first degrees. In the second the exam 1 score feature was second degree and in the last the exam 2 score feature was 3rd degree where exam 1 score feature same as the second.

- **K-fold sampling**

The K-fold sampling was applied on the house price data set from the first exercise with K equal 5. The data was divided accordingly into 5 sets each containing 3600 sample and was used for the last model which have the 6th degree.

- **Regularization**

Regularization technique was applied on both data sets utilizing the linear and logistic regression. A set of lambdas were created containing 0 0.01 0.05 0.1 0.5 1 5 10. For the linear regression different degree models were trained using each lambda and tested. The logistic regression models with different degrees were trained and tested for each different lambda from the previous set.

Analysing the results:

1. First Assignment results for different techniques

This is the results concluded from applying random sampling and shuffling techniques for each degree model.

Degree	First	Second	Third	Fourth	Fifth	Sixth
J (CV)	34339153416.014427	30014636120.37325	30005388815.93281	29985026013.978413	27260867640.791466	26927211279.592083

J (test)	24097534927.29127	21049939081.942837	21045026694.735073	21036156176.456947	17942869579.033318	17777734998.318657

And this is the results from testing the models without random sampling.

Degree	First	Second	Third	Fourth	Fifth	Sixth
J (CV)	25244160206.976276	22034153443.112682	22008755640.055515	21980888955.498722	19154499854.237453	18937756243.515877
J (test)	24369572452.806427	21243835443.11221	21246399414.904533	21238495286.15241	18195676777.772175	18038803725.88177

Although the J(CV) of the above table is slightly better than those of the first table but the J(test) in both are nearly the same.

This are the results of the K-fold technique where the metric used is the cost function applied on the test set. The five sets were named S1,2,3,4 and 5. in each time one of the sets was taken as the test sample and the others were the train sets.

Degree	First	Second	Third	Fourth	Fifth
Test Sample	S5	S4	S3	S2	S1
J (CV)	17044704401.954456	18624164667.758224	22694434694.140636	22350126942.5694	21184492467.779766

And those are the results from the regularization technique applied on different lambdas and different degrees using CV set.

Degree/Lambdas	0	0.01	0.05	0.1	0.5	1	5	10
1st	25244160208.976276	5848471707166.148	29141161202905.555	58256526523380.75	291159588687438.1	582238772992865.5	2908888078008638.5	5812246865343023.0
2nd	22034153442.112682	6102080724821.959	30422030633811.633	60821436181703.68	303995408764866.1	607909703396084.8	3037098948453307.0	6068281103914253.0
3rd	22008755639.055515	6095689893653.899	30390176172410.902	60757747915152.914	303676879437048.7	607272196551472.8	3033892647925022.0	6061821734392028.0
4th	21980888956.498722	6101609367485.173	30419882668769.473	60817182926152.24	303973932106684.56	607865745546577.6	3036837219570326.5	6067653109437821.0

5th	19154499859.237453	6069642739411.186	30271366357579.156	60523004873509.945	302515474358475.3	604954472501224.8	3022404536061575.5	6039070183379429.0
6th	18937756247.515877	6077063518959.113	30309337291646.06	60599163639835.836	302897141323682.6	605718038452755.8	3026223819932054.0	6046710316890518.0

2. Second Assignment results for different techniques

The table below shows the results of applying random sampling and shuffling on different degrees.

Degree	First	Second	Third
J (CV)	0.2558754664228678	0.1411465031307492	0.3026520093719782
J (test)	0.2742096209034244	0.09288470067252748	0.35126795488388396

And these are the results of applying the same technique without the random sampling on the same degrees.

Degree	First	Second	Third
J (CV)	0.1670884574310742	0.1570884574310742	0.3509194210074731
J (test)	0.2524452938856521	0.2024452938856521	0.29889376500910814

The last results in this exercise are the ones conducted from the regularization with different lambda values for different degrees using CV set.

Degree /lambda	0	0.01	0.05	0.1	0.5	1	5	10
1st	0.28240653030885154	0.3095196628337426	0.41797218289175525	0.5535378242054041	1.6380624292690227	2.993716976145704	13.838904863962645	27.395267770469196
2nd	0.3143647249063731	0.3143839703964817	0.31445565072230036	0.3145495987929592	0.31110874090211554	0.31761497658318516	0.32550299804625465	0.34461227507681336
3rd	0.3097337573003858	0.3109055290958805	0.31341241778878737	0.31657023747496843	0.31329802257101497	0.32443726581422067	0.3283905505620721	0.335014127467089

Conclusion:

After applying the different techniques mentioned above to the two different sets of house prices and the student exam score sets using linear and logistic regression. The random sampling technique didn’t improve the prediction accuracy, but in the logistic regression the random sampling did contribute and improved the results a little. Although the CV test results changed by changing between the sampling techniques using random and without the random sampling but the J test results didn’t change.For the regularization technique the first lambda equal to zero resulted in the best accuracy for both regression models.