**Assignment 1&2 Take 2**

**Name: Eyad Hesham Dawood ID: 40-5465**

**Data-sets description and modifications:**

* The data-set used in the linear regression models, for training and testing is the house price data-set given for this assignment.

The data-set is composed of 18000 samples and 21 features.

Not all the features were used in either the training phase or the testing phase of the models, because not all the features have the same effect or relation with the label ‘price’. So a correlation matrix was constructed to measure the relation between the features and the label, so the standard will be, if the correlation coefficient value is less than 0.2 the feature will be dropped. So the features dropped are : ‘date', 'id', 'zipcode', 'price', 'sqft\_lot', 'condition' ,'yr\_built' ,'yr\_renovated' ,'long', 'sqft\_lot15'. These features have correlation coefficients with negative values and zero values.

All the remaining features which will be used are normalized using the mean and the standard deviation to make the convergence of the model easier and faster.

* The data-set used in the logistic regression models, for training and testing is the student exam score data-set.

The data-set is composed of 100 samples, 2 features and 1 label.

Both data-sets were splitted into 60% of the data are for training as a training set, 20% of the data are for cross validation as a cross validation set and 20% of the data are for testing as a testing set.

**Experiments performed**

1. **Degrees and random sampling**

* For linear regression, 6 models were created with different degrees equals to [1, 2, 3, 4, 5, 6]. Each model was trained on the training set for 400 epochs and with learning rate equals to 0.1. The models were trained and tested on two different versions of the data-set, one is with random sampling and shuffling applied and the other is without random sampling and shuffling applied.
* For **logistic regression**, 3 models were created with different degrees equals to [1, 2, 3]. The models were trained and tested on two different versions of the data-set, one is with random sampling and shuffling applied and the other is without random sampling and shuffling applied.

1. **K-fold sampling technique**

The K-fold sampling was applied on the dataset of house price using K = 5. So 5 sample sets were created each with 3600 samples and were used for training and testing for the model with degree = 6 using learning rate = 0.1 and trained for 400 epochs for each sample set.

1. **Regularization**

In this experiment, the regularization technique was applied on the linear regression and the logistic regression. A set of lambas was created equals to [0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10]

* For **linear regression**, different models with different degrees from 1 to 6, were trained and tested using each value of lambda in the set of lambdas using learning rate equals to 0.1 and trained for 400 epochs each.
* For **logistic regression**, different models with different degrees from 1 to 3 were trained and tested using each value of lambda in the set of lambdas.

**Results and analysis**

1. **Experiment 1:**

The results metric is the cost function applied on the cross validation set and the testing set for the experiment.

For **linear regression,** Table 1 and Table 2 shows the results of applying random sampling and shuffling techniques on different models with different degrees. As shown in the tables, the model with degree = 6 which was trained and tested without random sampling achieved the highest accuracy and lowest J\_test.

For **logistic regression**, Table 3 and Table 4 shows the results of applying random sampling on different models with different degrees. As shown in the tables, the model with degree = 2 which was trained and tested with random sampling achieved the highest accuracy and lowest J\_test.

Table 1 model different degrees with random sampling

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Which feature | J\_cross | J\_test |
| 1 | All features with degree = 1 | 34339153417.014427 | 24097534928.29127 |
| 2 | Feature sqft\_living is with degree = 2 | 30014636121.37325 | 21049939080.942837 |
| 3 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3 | 30005388816.93281 | 21045026693.735073 |
| 4 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4 | 29985026014.978413 | 21036156175.456947 |
| 5 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4  Feature grade is with degree = 5 | 27260867641.791466 | 17942869578.033318 |
| 6 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4  Feature grade is with degree = 5  Feature bathroom is with degree = 6 | 26927211278.592083 | 17777734990.318657 |

Table 2 different degrees without random sampling

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Which feature | J\_cross | J\_test |
| 1 | All features with degree = 1 | 25244160208.976276 | 24369572459.806427 |
| 2 | Feature sqft\_living is with degree = 2 | 22034153442.112682 | 21243835447.11221 |
| 3 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3 | 22008755639.055515 | 21246399412.904533 |
| 4 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4 | 21980888956.498722 | 21238495287.15241 |
| 5 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4  Feature grade is with degree = 5 | 19154499859.237453 | 18195676778.772175 |
| 6 | Feature sqft\_living is with degree = 2  Feature sqft\_above is with degree = 3  Feature sqft\_living15 is with degree = 4  Feature grade is with degree = 5  Feature bathroom is with degree = 6 | 18937756247.515877 | 18038803723.88177 |

Table 3 different models degrees with random sampling

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Which feature? | J\_cross | J\_test |
| 1 | All features with degree = 1 | 0.2558754664228678 | 0.2742096209034244 |
| 2 | Exam 1 score feature is with degree = 2 | 0.1411465031307492 | 0.09288470067252748 |
| 3 | Exam 1 score feature is with degree = 2  Exam 2 score feature is with degree = 3 | 0.3026520093719782 | 0.35126795488388396 |

Table 4 different models degrees without random sampling

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Which feature? | J\_cross | J\_test |
| 1 | All features with degree = 1 | 0.1670884574310742 | 0.2524452938856521 |
| 2 | Exam 1 score feature is with degree = 2 | 0.1570884574310742 | 0.2024452938856521 |
| 3 | Exam 1 score feature is with degree = 2  Exam 2 score feature is with degree = 3 | 0.3509194210074731 | 0.29889376500910814 |

1. **Experiment 2**

Table 3 shows the results of applying the K-fold technique with K=5 using a model with degree = 6. The metric used is the cost function applied on the test sample set.

Table 5 K-fold with K=5 using model with degree 6

|  |  |  |  |
| --- | --- | --- | --- |
| Iteration | Train Samples | Test Sample | J\_Test |
| 1 | S1,S2,S3,S4 | S5 | 17044704400.954456 |
| 2 | S1,S2,S3,S5 | S4 | 18624164668.758224 |
| 3 | S1,S2,S4,S5 | S3 | 22694434693.140636 |
| 4 | S1,S3,S4,S5 | S2 | 22350126940.5694 |
| 5 | S2,S3,S4,S5 | S1 | 21184492468.779766 |

1. **Regularization**

The result metric used is the cost function J applied on the cross validation set.

For **linear regression**, Table 6 shows the results of different models with different degrees from 1 to 6 trained and tested using lambda values in the lambdas set. As shown in the table, the lowest cost function and highest accuracy was achieved by the model of degree = 6 using lambda = 0

For **logistic regression**, Table 7 shows the results of different models with different degrees from 1 to 3 trained and tested using lambda values in the lambdas set. As shown in the table, the lowest cost function and highest accuracy was achieved by the model of degree = 1 using lambda = 0

Table 6 regularization with different lambdas and degrees using cross validation set

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Degree | ƛ=0 | ƛ=0.01 | ƛ=0.05 | ƛ=0.1 | ƛ= 0.5 | ƛ=1 | ƛ=5 | ƛ=10 |
| 1 | 25244160208.976276 | 5848471707166.148 | 29141161202905.555 | 58256526523380.75 | 291159588687438.1 | 582238772992865.5 | 2908888078008638.5 | 5812246865343023.0 |
| 2 | 22034153442.112682 | 6102080724821.959 | 30422030633811.633 | 60821436181703.68 | 303995408764866.1 | 607909703396084.8 | 3037098948453307.0 | 6068281103914253.0 |
| 3 | 22008755639.055515 | 6095689893653.899 | 30390176172410.902 | 60757747915152.914 | 303676879437048.7 | 607272196551472.8 | 3033892647925022.0 | 6061821734392028.0 |
| 4 | 21980888956.498722 | 6101609367485.173 | 30419882668769.473 | 60817182926152.24 | 303973932106684.56 | 607865745546577.6 | 3036837219570326.5 | 6067653109437821.0 |
| 5 | 19154499859.237453 | 6069642739411.186 | 30271366357579.156 | 60523004873509.945 | 302515474358475.3 | 604954472501224.8 | 3022404536061575.5 | 6039070183379429.0 |
| 6 | 18937756247.515877 | 6077063518959.113 | 30309337291646.06 | 60599163639835.836 | 302897141323682.6 | 605718038452755.8 | 3026223819932054.0 | 6046710316890518.0 |

Table 7 regularization with different lambdas and degrees using cross validation set for logistic regression

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Degree | ƛ=0 | ƛ=0.01 | ƛ=0.05 | ƛ=0.1 | ƛ= 0.5 | ƛ=1 | ƛ=5 | ƛ=10 |
| 1 | 0.28240653030885154 | 0.3095196628337426 | 0.41797218289175525 | 0.5535378242054041 | 1.6380624292690227 | 2.993716976145704 | 13.838904863962645 | 27.395267770469196 |
| 2 | 0.3143647249063731 | 0.3143839703964817 | 0.31445565072230036 | 0.3145495987929592 | 0.31110874090211554 | 0.31761497658318516 | 0.32550299804625465 | 0.34461227507681336 |
| 3 | 0.3097337573003858 | 0.3109055290958805 | 0.31341241778878737 | 0.31657023747496843 | 0.31329802257101497 | 0.32443726581422067 | 0.3283905505620721 | 0.335014127467089 |

**Conclusion**

To sum up, three experiments were performed with linear regression and logistic regression using two data-sets, the house price data-set and the student exams scores data-set. The random sampling technique applied on linear regression did not contribute in increasing the prediction accuracy while in logistic regression, random sampling increased the classification accuracy. Also applying regularization technique with lambda = 0 achieved the best accuracy in both linear regression and logistic regression, so one can deduce that regularization technique in this case did not contribute in improving the accuracy of the models in both cases.