Danmarks Tekniske Universitet

DTU

Machine learning & Data mining Report 2

Authors

James Alexander Cowie - s192911
Ghassen Lassoued - s196609
Chandykunju Alex - s200113

April 17, 2020

Contents

1 Introduction **1**

2 Regression **2**
2.1 Regression Models-Comparison 4

3 Classification **6**
3.1 Two level cross validation for classification 6
3.2 Logistic Regression 7
3.3 K-Nearest Neighbors 7
3.4 Decision Tree 8

4 Discussion **9**

5 Conclusion **10**

6 Work Distributions **I**

List of Figures **I**

1 Introduction

This paper builds on the work done in the previous report on the data set HPRICE2. The attributes features, distributions and suitability for further machine learning processes were explored and discussed.
Figure 1 gives an overview of the data attributes.

	price	crime	nox	rooms	dist	radial	proptax	stratio	lowstat	lprice	lnox	lproptax
Description	Median housing price, $	Crimes committed per capita	Nitrous oxide, parts per 100 mill.	Avg. number of rooms per house	Weighted dist. to 5 employment centers	Accessibiliy index to radial highways	Property tax per $1000	Average student-teacher ratio	% of people 'lower status'	log(price)	log(nox)	log(proptax)
Type	Discrete, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Interval	Continuous, Ratio	Discrete, Ordinal	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio
count	506	506	506	506	506	506	506	506	506	506	506	506
mean	22511.5	3.61154	5.54978	6.28405	3.79575	9.54941	40.8237	18.4593	12.7015	9.94106	1.69309	5.9314
std	9208.86	8.59025	1.1584	0.702594	2.10614	8.70726	16.8537	2.16582	7.23807	0.409255	0.20141	0.396367
min	5000	0.006	3.85	3.56	1.13	1	18.7	12.6	1.73	8.51719	1.34807	5.23111
25%	16850	0.082	4.49	5.8825	2.1	4	27.9	17.4	6.9225	9.73209	1.50185	5.63121
50%	21200	0.2565	5.38	6.21	3.21	5	33	19.1	11.36	9.96176	1.68269	5.79909
75%	24999	3.677	6.24	6.62	5.1875	24	66.6	20.2	17.0575	10.1266	1.83098	6.50129
max	50001	88.976	8.71	8.78	12.13	24	71.1	22	39.07	10.8198	2.16447	6.56667

Figure 1: Attributes description

As the conclusion drawn was that the data set is suitable for machine learning purposes the task is now implementing some processes. Specifically solving a regression problem. The solution of which is compared with an artificial neural network (ANN) solution and a baseline.

We also attempt to define classes and classify the features radial and crime. We evaluate the performance of logistic regression, K-nearest neighbours and decision tree approaches relative to a baseline and each other.

2 Regression

We wish to investigate to what degree housing prices are predictable using the attributes listed in Figure 1
We define this problem as:

$$price = a_1 + a_2 rooms2 + a_3 dist + a_4 radial$$
$$+a_5 proptax + a_6 stratio + a_7 lowstat + a_8 crime + a_9 nox^2 + \varepsilon$$

In order to facilitate regularization the data is standardized. Ie feature transforming the input matrix $\mathbf{X}$ to have zero mean and a standard deviation of 1.

Equation 1 is the least square regularization cost function (L_2).

$$E_\lambda(\mathbf{w}, w_0) = ||\mathbf{y} - w_0 - \hat{\mathbf{X}}\mathbf{w}||^2 + \lambda||\mathbf{w}||^2, \lambda \geq 0 \tag{1}$$

$\lambda||\mathbf{w}||^2$ is known as the regularization term and λ specifically as the *regularization constant*. Correctly tuning this parameter allows the user to find the model with the lowest generalization error. Minimizing Equation 1 in the machine learning framework is named *ridge regression*.

The expectation for increasing values of λ is first a reduction in the generalization error until the regularization becomes too severe which results in an eventual increase in the error. Small λ values allow for higher variance while increasing λ reduces variance but increases bias. Another formulation of the trade off is between overfitting and underfitting.

Figure 13 show the effects of increasing λ and how it eventually constrains the coefficient values to 0 mean.

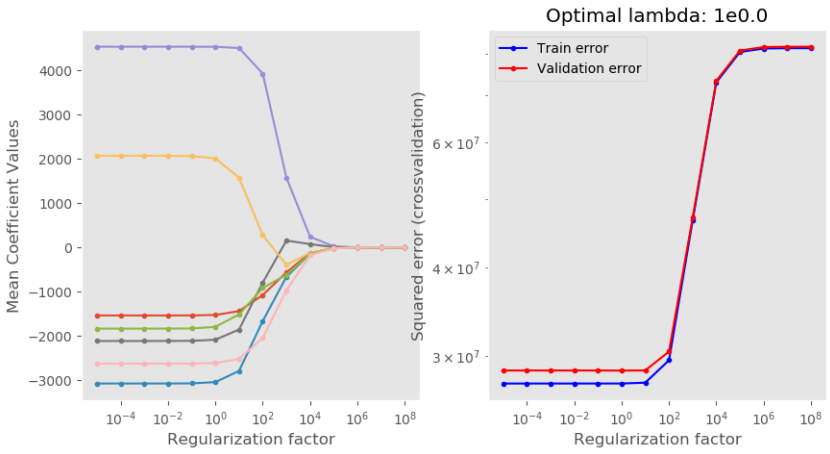


Figure 2: Regularization constant estimation

The resulting weights using $\lambda = 1$ are shown in Figure 3

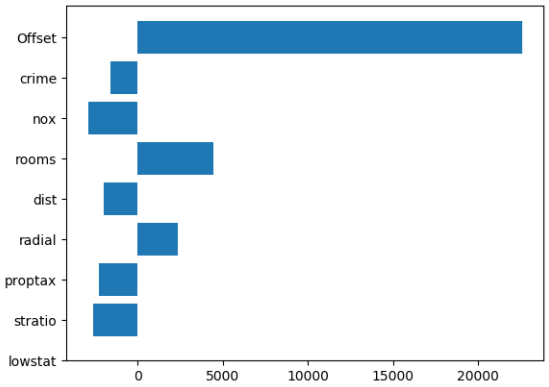


Figure 3: Parameter weights

The values of which are:

Offset	22567.16
Crime	-1605.31
NOX	-2921.9
Rooms	4461.28
Dist	-2021.3
Radial	2382.34
Proptax	-2253.86
stratio	-2598.38

Table 1: Parameter weights

The values in Table 1 are not surprising. The number of rooms for instance has a high influence on the price. More rooms indicate a larger house, which leads to higher prices. While a higher student teacher ratio could indicate less developed areas. It is however conceivable that city centres, where prices tend to be high, also have high student teacher ratios.

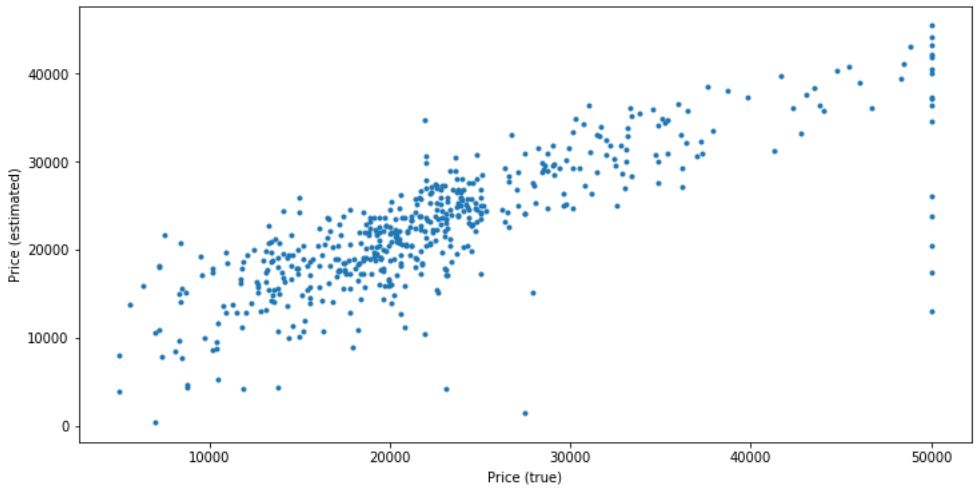


Figure 4: Predictions vs. true values

Figure 4 show the relative success of the regression models predictions. Where we experience trouble is particularly evident in higher price ranges. This may be improvable by implementing some feature transformations on the parameters with uneven distributions.

2.1 Regression Models-Comparison

In order to predict the price, 3 models were implemented: Ridge regression, Artificial neural network(ANN) and a Baseline. The comparison between them is done by applying 2 level cross-validation.
Baseline computes the mean of y on the training data, and use this value to predict y on the test data. As a result this model is only computed in the outer loop of the cross-validation. In ridge regression, the regularization parameter λ is tuned in the inner loop: the optimal λ is chosen from a range of λ values. For the ANN the number of hidden units were chosen similarly in the inner loop. The outer loop is used to evaluate the performance of the chosen models i.e calculate the generalization error for the optimal model found.
Since this is a regression, the output activation function is linear. For the hidden layer, the ReLU activation function was used.

The results are summarized in the following table:

outer fold	ANN		Ridge regression		Baseline
i	opt_h_i	E_test_i	opt_labmda_i	E_test_i	E_test_i
1	6	5.15634703e+08	3	34353176.2078	1.23073159e+08
2	7	3.92675133e+08	3	13592937.2763	6.05527271e+07
3	7	4.25514925e+08	3	41583695.6212	1.18528662e+08
4	7	3.71178499e+08	3	26749669.6409	7.09895856e+07
5	7	4.10565767e+08	3	20189287.6295	8.46899291e+07
6	7	4.22138058e+08	3	29883422.9337	7.96182089e+07
7	7	4.03478881e+08	3	18866718.5895	7.60693603e+07
8	7	2.99379545e+08	3	13662097.6914	7.75399521e+07
9	7	3.28489336e+08	3	17896484.4886	7.09896251e+07
10	7	4.11351333e+08	3	35226209.0786	8.84981734e+07

The statistical evaluation where SETUP I was used is summarized in the table below:

Pairwise Test	p_value	confidence interval_lower	confidence interval_upper	null H0
Ridge vs ANN	6.47499543e-85	-4.03881539e+08	-3.42329841e+08	rejected
Ridge vs Baseline	1.93181063e-22	-71472865.65578204	-48283593.76378753	rejected
Baseline vs ANN	4.34412191e-90	-3.37961524e+08	-2.88493396e+08	rejected

It is seen that the confidence intervals do not contain 0. The lower p_value is, the more evidence there is that A is better than B.
Applying the mean, the following estimation of generalization errors can be found:
-Ridge : 25200369.915739376
-ANN: 398040618.0286451
-Baseline: 85054938.21814734

According to statistical results and generalization errors, it can be said that Ridge Regression is better than ANN , Ridge regression is better than Baseline, Baseline is better than ANN.

In the previous section, optimal λ was 1. It is different in this section which is 3. This is due, probably, to the different methods used when implementing ridge regression or some inconsistencies in the code.

The reduced model given from the original literature is

$$log(price) = a_1 + a_2 rooms2 + a_3 log(dist) + a_4 log(radial)$$
$$+a_5 proptax + a_6 stratio + a_7 log(lowstat) + a_8 crime + a_9 nox^2 + \varepsilon$$

The model found by ridge regression is :

$$price = 22553.2346 - 1091.9152 crime - 1857.8889 nox + 3206.3679 rooms - 2338.7062 dist$$

$$+2414.3138 radial - 2069.5727 proptax - 2224.0594 strratio - 3426.5614 lowstat + \varepsilon$$

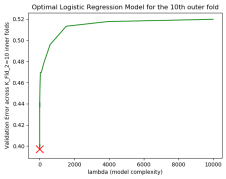
ε is expected to be normally distributed. This was not analysed in this report.

3 Classification

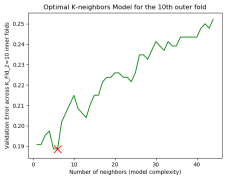
We attempt to classify the radial and crime attributes. The radial data showcase categorical behaviour with a rapid change in values at radial=24. Indicating a potentially classifiable behaviour. Moreover as we discussed in the previous report crime rate is best suited for classification so that we observed log(crime) feature and found that there is a cluster of different values between log(crime)=0. So we apply a simple threshold:
log(crime)>0 is high crime rate and log(crime)<=0 low crime rate.
Before performing the classification algorithms it is important to find the optimum parameters suited for classification.

3.1 Two level cross validation for classification

Searching for optimum parameters the following range of parameters in two level cross validation are set. For logistic regression classification, c=1/lambda is chosen by setting lambda value into logspace of -16 to 4 with 50 points:logspace(-16, 4, 50). For KNN classification the neighbouring distance parameters are set in the range of 1 to 45. Finally for the decision tree the the range of values for depth is set to interval 2 to 51. The cross validation script will search on these values and find the minimum error with optimum parameters.The output graph of the two level cross validation of data can be seen below.

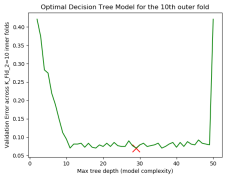


(a) logistic regression cross validation graph:radial

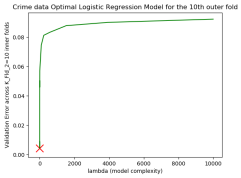


(b) KNN cross validation: radial

Figure 5: Cross validation results

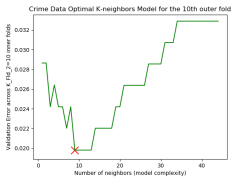


(a) Decision Tree cross validation: radial

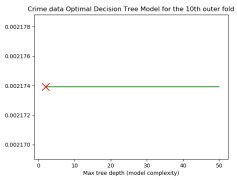


(b) logistic regression cross validation graph:crime

Figure 6: Cross validation results



(a) KNN cross validation: crime



(b) Decision tree cross validation: crime

Figure 7: Cross validation results

3.2 Logistic Regression

We have optimized the regularization value for logistic regression algorithm in classification. It seems 1e-8 is the best optimum value for radial classification

Fold	1	2	3	4	5	6	7	8	9	10
lambda	1e-08	1e-08	1e-08	1e-08	1e-08	1e-08	1e-08	1e-08	1e-08	1e-08
Error	0.41176471	0.37254902	0.39215686	0.35294118	0.33333333	0.33333333	0.48	0.4	0.38	0.36

Figure 8: Generalization error and λ for the logistic lasso regression algorithm on each fold on radial classification.

Since the log(crime) classification has only two classes it seems that the error became zero for most of the values of lambda

Fold	1	2	3	4	5	6	7	8	9	10
lambda	6.86e-05	6.86e-05	6.86e-05	6.86e-05	0.00017	6.86e-05	6.86e-05	0.00017	6.86e-05	2.68e-05
Error	0	0	0	0	0.01960784	0	0	0	0	0

Figure 9: Generalization error and λ for the logistic lasso regression algorithm on each fold on crime classification.

3.3 K-Nearest Neighbors

Since the data points are not regular in terms of distance we use cosine transform as metrics instead of euclidean distance. This provides better clustering of data compared to euclidean distance based classification.

Fold	1	2	3	4	5	6	7	8	9	10
Neighbours	3	5	1	5	3	5	3	3	5	3
Error	0	0	0.022	0	0	0	0	0	0	0

Figure 10: Generalization error and K:Neighbours for the K-Nearest Neighbor Classification algorithm on each fold radial data.

From the analysis it is evident that 3 and 5 neighbours give the least error in each fold for radial classification. We proceed with 3 folds.

Fold	1	2	3	4	5	6	7	8	9	10
Neighbours	11	7	7	5	5	5	6	12	7	9
Error	0	0	0.0	0	0	0	0	0	0	0

Figure 11: Generalization error and K:Neighbours for the K-Nearest Neighbor Classification algorithm on each fold crime data.

For the crime classification values 5,7,9,12 are optimum values. We will be using 5.

3.4 Decision Tree

Compared to other approaches the decision tree gives best result for radial classification. It seems depth value= 48 is best suited for the decision tree based classification on radial data.

Fold	1	2	3	4	5	6	7	8	9	10
Depth	33	48	26	34	41	26	41	28	40	23
Error	0.25	0.15	0.21	0.17	0.17	0.27	0.24	0.2	0.24	0.28

Figure 12: Generalization error and corresponding depth for the Decision Trees algorithm on each fold radial classification.

In addition to that when it comes to crime classification values like 5,7,9,11 giving the zero error most of the time. Since the crime classification is almost a binary classification.

Fold	1	2	3	4	5	6	7	8	9	10
Depth	11	7	7	5	5	5	6	12	7	9
Error	0	0	0	0	0	0	0	0	0	0

Figure 13: Generalization error and corresponding depth for the Decision Trees algorithm on each fold crime classification.

As we discussed earlier the crime data is easy to classify, hence it gives all zero error to every fold.

4 Discussion

The dataset is as mentioned a subset of the complete set originally used in [*Hedonic housing prices and the demand for clean air, Journal of Environmental Economics and Management, volume 5, number 5, pages 81-102, year 1976*]

The papers purpose was to attempt to model the willingness to pay for air quality improvements. Since we have two different aims we cannot directly compare results. However the attributes are shown to have correlations to one another that match our calculations.

There are improvements to be made. We suspect transforming the data into more suitable distributions will decrease the generalization error. There are also inconsistencies in the estimation λ. The baseline model does not use any of the features of $\mathbf{X}$. Yet, ANN is unable to outperform the baseline, meaning it does not make meaningful use of the features. This can either be because ANN model is implemented wrong, it is unsuited for the task, or because there is a serious data quality issue. The probable reason is that the implementation of ANN is wrong. Two methods were used to implement ANN: using the toolbox provided, and MLPRegressor from sklearn.neural_network. In both cases, results are unconvincing. It can be seen that ANN is not working properly also from the fact that the optimal number of hidden layers continue to increase.

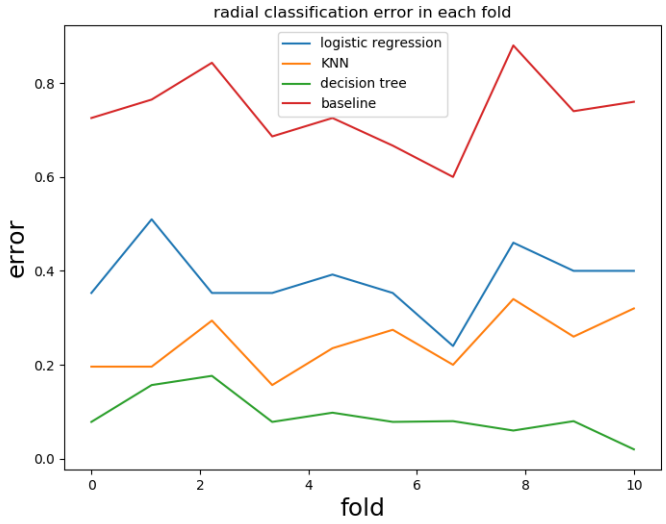


Figure 14: Radial classification error for different methods vs number of folds

From the figure it is evident that the base line error is very high compared to rest of the tree model. Which means the models we used have performed well and returned good results. The Logistic regression solutions have errors between 0.4 to 0.29. From Figure 14 it is evident that the seventh fold provides the model that performs best. The KNN also outperform both baseline and logistic regression model and returns an error between 0.25 to 0.1 the third fold has least error in that case. The method which performs best is decision tree which gives error between 0.14 to 0.06 ,the 10th fold has least error.

5 Conclusion

In this project, regression was used in order to predict median housing price. Different models were implemented: ridge regression, linear regression with feature selection, artificial neural network. Ridge regression outperformed the others.
Classification proved successful to a certain extent with the decision tree implementation outperforming the others.

6 Work Distributions

Each group member has contributed to every section in the report. That being said the primary contributions to the sections are as follows:

Introduction :
James 80% Ghassen 10% Alex 10%
regression part a :
James 85 % Ghassen 5% Alex 10%
regression part b :
James 10 % Ghassen 85% Alex 5%
Classification :
James 5 % Ghassen 5 % Alex 90 %
Discussion :
James 90% Ghassen 5% Alex 5%
conclusion:
James 25% Ghassen 50% Alex 25%

List of Figures

1 Attributes description . . . 1
2 Regularization constant estimation . . . 2
3 Parameter weights . . . 3
4 Predictions vs. true values . . . 4
5 Cross validation results . . . 6
6 Cross validation results . . . 6
7 Cross validation results . . . 7
8 Generalization error and λ for the logistic lasso regression algorithm on each fold on radial classification. . . . 7
9 Generalization error and λ for the logistic lasso regression algorithm on each fold on crime classification. . . . 7
10 Generalization error and K:Neighbours for the K-Nearest Neighbor Classification algorithm on each fold radial data. . . . 8
11 Generalization error and K:Neighbours for the K-Nearest Neighbor Classification algorithm on each fold crime data. . . . 8
12 Generalization error and corresponding depth for the Decision Trees algorithm on each fold radial classification. . . . 8
13 Generalization error and corresponding depth for the Decision Trees algorithm on each fold crime classification. . . . 9
14 Radial classification error for different methods vs number of folds . . . 10