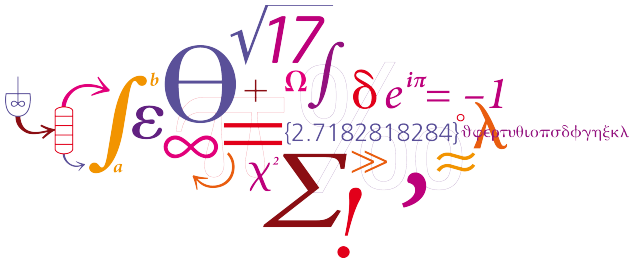
Technical University of Denmark

02450 - Introduction to Machine Learning and Data Mining

Project 2

Lucas Besnard - s201778
Jan Cedric Issel - s201232
Ayoub El Outati - s201602



November 16, 2020

DTU Compute
Department of Applied Mathematics and
Computer Science

1 Regression (a)

In this section, we explain which variable of our data set we aim to predict based on which other variables. For the prediction we test different linear regression models and estimate their generalization error for different regularization parameters λ. Finally, we use the model with the lowest generalization error to make a prediction and we evaluate how the selection of the attributes affects the quality of the respective prediction.

1.1

The attributes of our data set are focused on the characteristics of housing blocks in California. As stated in the previous report, for the regression task it is our goal to predict the median house value in the different housing blocks. Therefore, our choice for the predicted variable is the median house value. In order to use our data set for a regression model, we apply some feature transformations to it. First, the ocean proximity, a non-numerical nominal scaled feature, is transformed with one-out-of-K encoding. Thus, we can use it as an input feature for the regression. Next, we evaluate which input features should be taken into account for the prediction by applying sequential feature selection (forward selection) to all features of the data set. For this procedure, the data set is split into five equal sized subsets with 5-fold cross validation. Each subset is evaluated with a linear regression model once with all features selected and once with a specific feature combinations according to the forward selection algorithm. The selection results for the five folds are visualized in Figure 1. As can be seen, there is no clear tendency to drop a specific feature. A comparison of the prediction accuracy R^2 of the models with and without feature selection shows that both for train and test data set the accuracies don't differ at all. In both cases a prediction accuracy on the test set of $R^2 = 0.6859$ is achieved. A reason for this indifference can be that the main contribution to the model accuracy R^2 comes from four features which are visualized in Figure 2. The remaining features also contribute to the accuracy but affect it only marginally such that their selective combination with forward selection has no significant effect. Thus, it can be concluded that all features can be used for the following regression. Finally, we prepare the data for the following regularization attempt by performing standardization ($\mu = 0$, $\sigma = 1$) on the data set.

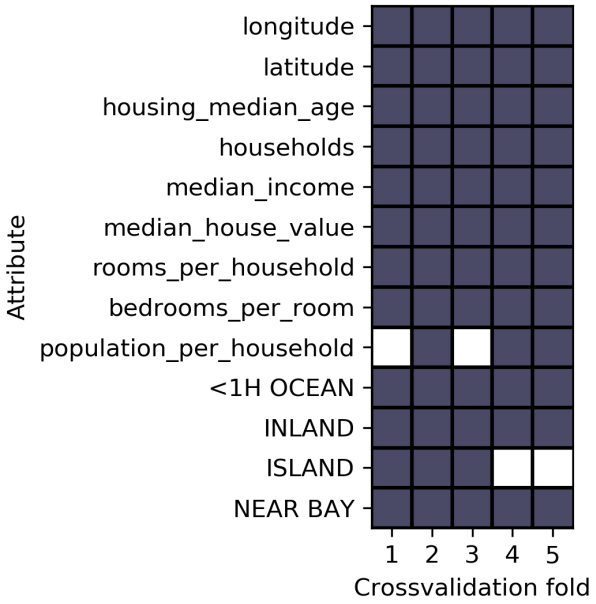


Figure 1: Forward selection with 5-fold cross-validation

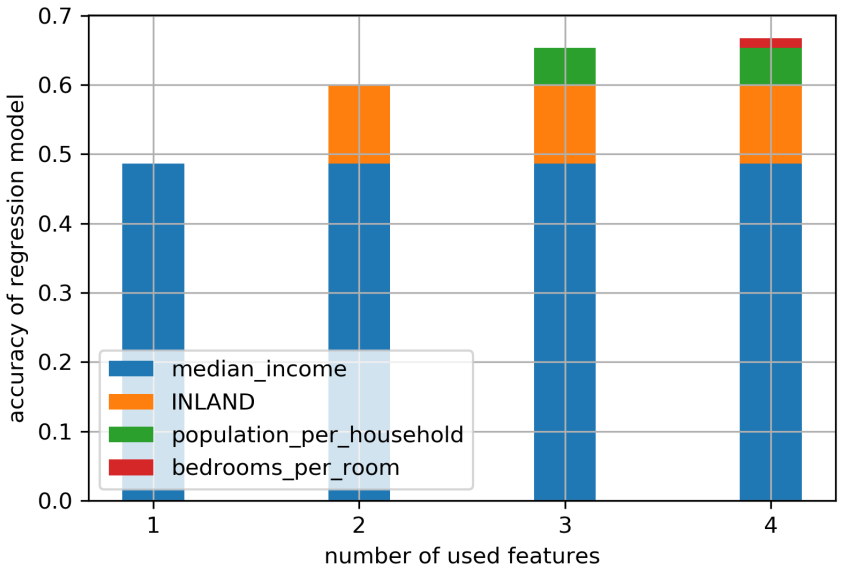
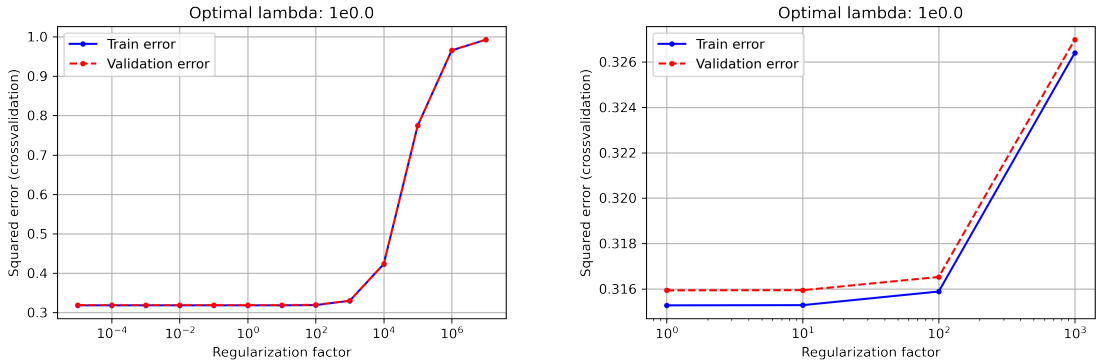


Figure 2: Contribution of features to model accuracy R^2

1.2

Regularization on a machine learning model is used in order to minimize the generalization error / test error of the model. This means that the ideal model should perform as good on unknown data as on the data it was trained on. A linear regression model can be regularized by penalizing large weights in the model with a parameter λ. To examine the generalization error of the regression goal described in Section 1.1, we set the range to

$\lambda = [10^{-5}; 10^{7}]$. The change of the squared error for the model trained on the full data set, can be seen in Figure 3a. The train error and validation error behave completely similar independent of the chosen λ. This behavior can be explained with the relatively high number of 20198 observations in the data set which helps the model to generalize and thus prevents overfitting on the training set. However, if we choose a small enough range of $\lambda = [10^{0}; 10^{3}]$ as visualized in 3b one can see that the validation error is actually slightly higher. In general, the error seems to increase with the increase of λ. The optimal value of λ corresponds to the lowest minimum squared error of the validation set. Therefore, in this case for the linear regression model a regularization value of $\lambda = 1$ should be used.



(a) **(b)**

Figure 3: Estimated generalization errors as a function of λ

1.3

We continue with the linear regression model that shows the lowest generalization error which is regularized with $\lambda = 1$. The trained regression model yields a weight w_i for each feature that influences the prediction. Thus, a prediction based on new observation data is done by weighting the new observations with their respective weight w_i. In Table 1, the weights of the final fold from the previously found optimal linear regression model are shown. Especially the median income as well as latitude and longitude stand out for their weight values. Since the latitude and longitude weights have relatively high negative values, they will reduce the median house value for large values of latitude and longitude. This result makes sense because larger latitude and longitude for the California area means moving away from the coast where housing tends to be more expensive. The median income affects the median housing value in a positive way. This makes sense as well since people with higher income tend to invest more into real estate.

Attribute name	Weight in last fold
longitude	-0.49
latitude	-0.51
housing median age	0.14
households	0.05
median income	0.7
rooms per household	0.08
bedrooms per room	0.2
population per household	-0.22
<1h ocean	0.05
inland	-0.08
island	0.02
near bay	0.01
near ocean	0.03

Table 1: Weights of the last fold

2 Regression (b)

In this section, we aim to find out which model predicts the median house value best by means of the generalization error E_i^{test}. Therefore, we compare the generalization error of the derived linear regression model from the previous section, an artificial neural network (ANN) and a baseline model. In order to reduce the bias of our results, we use two-level cross-validation. Finally, we evaluate the performance difference between the three models with hypotheses tests.

2.1

The two-level cross-validation consists of an outer fold with K_1 folds and respectively an inner fold with K_2 folds. In order to limit the time consumption for the training of the different models to a reasonable limit, we use $K_1 = K_2 = 5$. We want to find out, how complex each model should be to deliver the best, i.e. the lowest, generalization error E_i^{test}. Therefore, a complexity-controlling parameter has to be chosen. For the linear regression model, the regularization parameter λ is varied in a range of $\lambda = [10^{-5}; 10^8]$. For the ANN, the number of neurons in the hidden layer h is varied for the values $h = [1, 5, 10, 20, 40, 80]$. As a baseline model, the mean of y_{train} is used to predict y_{test}.

2.2

Based on the previously defined parameter ranges we test the three models with two-level cross-validation. From each set of inner folds of size $K = 5$, the complexity-controlling parameter that yields the lowest generalization error is chosen. The results of this procedure are shown in Table 2. It shows that the ANN compared to the other models yields the lowest generalization error and thus performs best on this data set. The optimal number of hidden neurons should be $h_i^* = 40$ according to the majority of the folds. For the linear regression model, without an exception the optimal $\lambda_i^* = 1.0$ is computed which is the same result as in the previous evaluation of the regularization parameter.

Unsurprisingly, the baseline model performs worst with an average generalization error of $E_i^{test} = 1$. However, it sets a benchmark that tells us that linear regression and ANN actually provide a significant improvement to the prediction accuracy of the median house value.

Outer fold	Linear regression		ANN		Baseline
i	λ_i^*	E_i^{test}	h_i^*	E_i^{test}	E_i^{test}
1	1.0	0.321	40	0.246	0.990
2	1.0	0.301	40	0.212	1.036
3	1.0	0.307	80	0.211	0.978
4	1.0	0.317	40	0.235	0.996
5	1.0	0.325	40	0.243	0.999

Table 2: Comparison of the generalization errors for linear regression, ANN and a baseline model

2.3

Afterwards, we must statistically evaluate if there is a significant difference of performance between the optimal models that we selected from the two-level cross-validation, by using one of the following methods : Setup I or Setup II. These comparisons will be made pairwise :

- ANN vs Linear regression
- ANN vs Baseline
- Linear regression vs Baseline

Having the choice between the setups mentioned above, we have decided to apply the Setup I method. This method is based on the fact that our conclusions will only be valid for our data set and not for other data sets generated from the same mechanism. From that point of view, Setup II seems to be a better alternative as it gives a more general conclusion. Nevertheless, in our problem, we have a clearly defined data set of the Californian house market which makes the Setup I applicable. In addition, it is too computationally expensive to train multiple models especially with the ANN model, which justifies our preference towards Setup I.

Based on the paired t-test for regression models, we once more use cross-validation to train our models on the complete data set. Indeed, as we can see in Table 2, the optimal parameters for the ANN model is not the same in each outer fold. The ANN model with 40 hidden units is the overall optimal model, but it was not trained on the complete data set. Therefore, we apply another K-fold cross-validation for our three models :

- Regularized linear regression with $\lambda = 1.0$
- ANN with $h_i = 40$
- Baseline model

We compute the predictions and then use the square loss to calculate the per-observation losses. However, one important assumption to validate is the normality assumption of the test error. In this situation, we know that this approach depends on the central limit theorem and should only be applied when the size of our test set is larger than 30 observations, which is the case. Therefore, we can finally compare the generalization error between a model A and a model B by using the estimated difference in test errors which is defined as

$$z = E_A^{test} - E_B^{test}. \tag{1}$$

Finally, we obtain the confidence interval and p-value of each pairwise comparison in Table 3.

Outer fold	Linear regression vs ANN		ANN vs Baseline		Linear regression vs Baseline	
i	CI	p-value	CI	p-value	CI	p-value
1	[0.096;0.131]	2.10e-38	[-0.799;-0.765]	0	[-0.709;-0.628]	5.31e-204
2	[0.104;0.132]	3.76e-59	[-0.868;-0.840]	0	[-0.778;-0.693]	1.61e-221
3	[0.107;0.137]	1.21e-58	[-0.809;-0.779]	0	[-0.713;-0.632]	5.54e-208
4	[0.095;0.132]	4.92e-33	[-0.811;-0.774]	0	[-0.720;-0.638]	2.35e-205
5	[0.093;0.131]	1.52e-30	[-0.806;-0.768]	0	[-0.716;-0.633]	1.74e-199

Table 3: Confidence interval (CI) at 95 % and p-value for each pair wise comparison

As we can see in Table 3, the confidence interval (CI) at 95% provides a range of plausible values for the the true difference of the test errors between the models that we compare. In the first pairwise comparison between the ANN (model B) and the linear regression (model A) we apply Equation 1. We can notice that on average this difference is located in a CI between 0.099 and 0.133 on a 95% confidence level. This confirms that the ANN model is better than the regularized linear regression. The p-value validates this result since we observe a low value close to zero which is inferior to the significance level $\alpha = 0.05$. Thus, the hypothesis of same performance between ANN and linear regression is rejected. About the comparison of the ANN (model A) and the baseline model (model B), we notice a negative range of values for the confidence interval, which highlights the performance of the ANN over the baseline model. We can also note a similar observation for the linear regression (model A) and baseline (model B) comparison. Therefore, we can conclude that both models are better than the baseline model. Concerning the p-values, all of them are around zero, which confirms that none of the models are identical.
Based on our findings for the different generalization errors of the three examined models and the validation of these errors with the hypothesis tests, we can recommend the use of the ANN for the prediction of the median house value since it generalizes best.

3 Classification

In this section, we test different models on a classification problem. Once again, we compare the model's generalization errors for different complexity comparing parameters and use two-level cross-validation. Next, we compare the performance difference between the three models with hypothesis tests.

3.1

As an appropriate classification problem, we have chosen to use the non-numerical nominal scaled feature 'ocean proximity' which has been transformed with one-out-of-K encoding in Section 1.1. The ocean proximity is measured in five different classes (<1H OCEAN, INLAND, ISLAND, NEAR BAY, NEAR OCEAN). Thus, it is a multiclass classification problem. The respective model will predict the ocean proximity using all the remaining features of the data set as input.

3.2

We evaluate the performance in terms of the generalization error E_i^{test} of a multinomial regression model since we have a multiclass problem, an ANN and and a baseline model. For the multinomial regression model, the regularization parameter λ is used as for complexity controlling. We vary on a range of $\lambda = [10^{-5}; 10^8]$. For the ANN, we use the number of hidden neurons h as parameter. As in the previous section, h is varied for the values $h = [1, 5, 10, 20, 40, 80]$. The baseline model will use the largest class of the data set according to the training data to predict everything in the test-data as belonging to that class.

3.3

Once again, we use two-level cross-validation with the number of inner and outer folds set to $K_1 = K_2 = 5$. Based on the previously defined parameter ranges we test the three chosen models and evaluate for which parameter the respective model yields the lowest generalization error E_i^{test}. The results of this procedure are shown in Table 4. It becomes apparent that the generalization errors of the ANN and the multinomial regression model are rather close to each other. However, the ANN performs slightly better in this case. This could be the case because the ANN is able to model more complex relations than the regression is, for example by using several hidden layers of neurons. However, since we only use one hidden layer for this evaluation, the potential of the ANN is still limited. The consensus for the optimal number of hidden neurons is $h_i^* = 80$. In the multinomial regression model, the optimal regularization value is not consistent among the outer folds but should be between λ_i^* = 1e-03 and λ_i^* = 1e-05. Since these are rather small values the regularization effect on the model is low.

Outer fold	ANN		Multin. regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	80	0.114	$1e{-}05$	0.190	0.683
2	80	0.108	$1e{-}03$	0.203	0.684
3	80	0.112	$1e{-}05$	0.193	0.682
4	80	0.110	$1e{-}04$	0.195	0.681
5	80	0.109	$1e{-}05$	0.193	0.680

Table 4: Comparison of the generalization errors for ANN, multinomial regression and a baseline model

3.4

We perform a statistical evaluation of our three models pairwise. We use the McNemera's test (setup I) with $\alpha = 0.05$. A summary of the results can be found in Table 5 where θ_L denotes the lower bound and θ_U the upper bound of the respective confidence interval. θ_E is the respective estimation for the difference between two models. When we compare the ANN with the multinomial regression, we show that the probability to obtain such a distribution, if the methods were of equal performance, is 1.60e-254. This enables us to reject the null hypothesis that the two methods are of equal performance. The neural network shows a better performance than the multinomial regression. A confidence interval for this difference of performance with a probability of 95% is $[0.079; 0.089]$ and the estimation of this difference is $\theta_E = 0.084$. Next, we show that the two methods are clearly superior to the baseline model. For both methods the probability of obtaining such a distribution if they where of the same performance is so small that Python prints 0. The interval of confidence for the difference of performance between ANN and the baseline model is $[0.564; 0.579]$ and the estimation for this difference is $\theta_E = 0.571$. The interval of confidence for the difference of performance between the multinomial regression and the baseline model is $[0.480; 0.495]$ and the estimation for this difference is $\theta_E = 0.488$. Since all the p-values are very close to zero, it can be confirmed that none of the examined models are identical.
Based on our findings for the different generalization errors of the three examined models and the validation of these errors with the hypothesis tests, we can recommend the use of the ANN for the classification of the ocean proximity of a housing block since the ANN generalizes best.

Model combination	p-value	θ_L	θ_U	θ_E
ANN - Multinomial regression	1.60e-254	0.079	0.089	0.084
ANN - Baseline	0	0.564	0.579	0.571
Multinomial regression - Baseline	0	0.480	0.495	0.488

Table 5: Pairwise statistical evaluation of ANN, multinomial regression and a baseline model

3.5

We continue with the multinomial regression that shows the lowest error rate which uses $\lambda = 1e{-}04$. For each classes the trained multinomial regression yields a weight for each feature. For each observation, the prediction is based on this weights. The class that obtains the highest score is selected. In Table 6, the weight of each feature for each class is represented for a trained multinomial regression. We can see that the main factors taken into account are the longitude and latitude. This make sense since the classes correspond to the proximity of the house block to the see. We can also note that the households have notable influence on the classification. Comparing this with the evaluation in Section 1.1, it becomes apparent that for this classification task different input features are relevant than for the regression task.

Attribute name	<1h ocean	inland	island	near bay	near ocean
longitude	13.0	27.6	-61.1	12.9	7.48
latitude	21.1	37.3	-97.5	24.2	14.9
housing median age	-0.405	-0.298	0.55	0.49	-0.34
households	2.41	2.40	-9.77	2.61	2.34
median income	0.398	0.067	-1.29	0.767	0.056
median house value	-0.426	-0.703	2.26	-0.576	-0.56
rooms per household	-0.963	-0.535	3.17	-1.18	-0.49
bedrooms per room	-0.860	-1.12	3.41	-0.69	-0.74
population per household	-0.752	-0.765	-2.27	0.4.34	0.317

Table 6: Weights per feature and predicted class

4 Discussion

4.1

During our analysis of the California housing prices data set, we gained considerable insight about the applicability of regression and classification models. In the first regression part we evaluated if and which value of regularization benefits the generalization error of the regression model the most. We learned for this data set in particular that a value of $\lambda = 1$ is optimal for the prediction of the median house value. We learned in general that the size of the data set affects the generalization error strongly. For a smaller data set it is crucial to use regularization because otherwise the model tends to overfit on the small training data set. In the subsequent regression part we aimed to compare the performance of different models on our data set and validated the results with the use of the t-test. We learned that the ANN performs best for our regression task. However, considering the time consumption of the training and considering that the linear regression model doesn't perform much worse we can conclude that the use of the ANN is not simply always justified. One has to weigh up time consumption and performance increase. From our classification task we learned that the ANN is superior in terms of the generalization error as well. However, here the performance increase compared to the multinomial regression is rather small again. All in all, we learned that we set a realistic goal regarding our machine learning tasks in the beginning and learned that a sufficient amount of data is very beneficial for the training of regression as well as classification models.

4.2

The California housing data set is very popular among introductory machine learning tutorials and has thus been subject to many blog posts as well as Jupyter notebooks on Kaggle. Here and here, one can see that our prediction results for the regression part are close to the approaches of others. Especially in the book 'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow' from Aurélien Géron, a similar approach for the comparison of different generalization errors has be taken. For the classification

task it was hard to find comparable literature since the feature we chose to use for the classification is not used by other people working on this dataset.

5 Responsibilities

Section	Responsible for more than 40%	Contributor
Regression A	s201232	
Regression B	s201602	s201232
Classification	s201778	s201232
Discussion	s201232	