Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Project 2

s191985 Julian Böhm (Regression, part a)
s196119 Emil Chrisander (Classification)
s192184 Jorge Montalvo Arvizu (Regression, part b)

November 2019

1 Regression, part a (s191985)

This section discusses the linear regression analysis of the South African heart disease data set. At first, it has to be defined which of the variables is the predicted one. In case of a logistic regression we would choose the binary variable *chd* (heart disease or no heart disease) as a dependent variable because in the overall problem we are interested in how the other variables affects this state. However, the task requires to solve the problem with a linear regression model. Therefore, we have to pick a continuous variable and decided on *ldl* which is low density lipoprotein cholesterol (often referred to as the 'bad' cholesterol). We decided on *ldl* because it showed the highest correlation rate with *chd* in the last report (together with age, adiposity and tobacco consumption). Moreover, we estimate high *ldl* levels to be connected to *obesity* and *adiposity*. All of the nine variables are included in the model as independent ones.
To execute the regression analyses, we had to apply some feature transformations to our data set. The variables *chd* and *famhist* (given in categories) had to be binarized. A one-out-of-coding was not necessary in our data frame. Furthermore, the data was standardized by subtracting the mean and dividing by the standard deviation. This results in a mean of 0 and a standard deviation of 1 for each column.
Linear regression model [1]

$$\begin{aligned} y_i &= f(x_i, w) = \tilde{x}_i^T w \\ E(w) &= ||y - \tilde{X}^T w||^2 \end{aligned} \tag{1}$$

Based on this formulas one can estimate the optimal weights w^* by minimizing the error. This is computationally achieved by applying the least squares regression model (*lm.LinearRegression()*).
Figure 1 shows the true *ldl* and the estimated *ldl*. From a visual observation, it can be seen that the model does not take into account the extreme values of the true *ldl* and 'compresses' the scale of the estimated *ldl*. Since these extreme values are just a few cases, we can concluded on a rather good fit of the model. Additionally, the underlying graph describes the residuals which are in a rather good accuracy of approximately -4 to 3 percent.

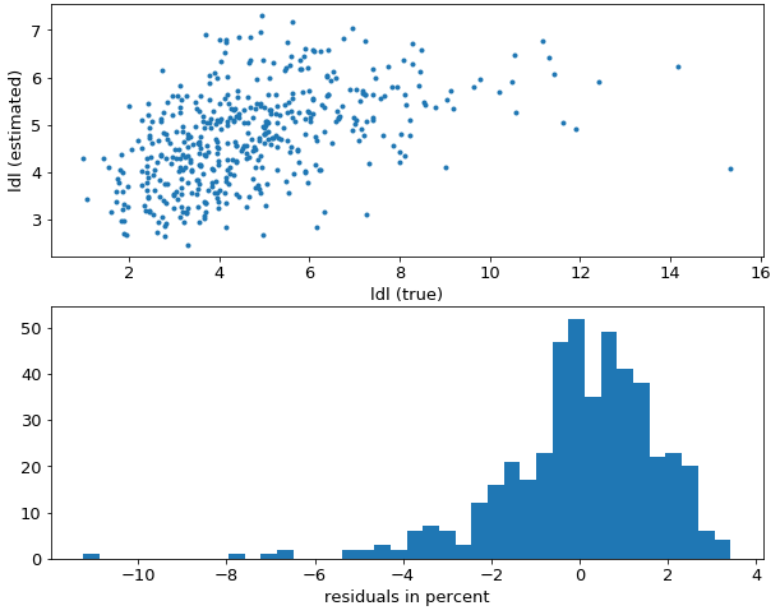
Following, we introduce a regularization parameter λ (equation 14.3) to the model:

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

For the model, we set the range of λ from 10^{-5} to 10^{11} and observed an ideal λ of $10^2 = 100$. In figure 2, it can be noted that the generalization error drops at $\lambda = 100$ and increases with higher λ values.
Table 1 lists the weights in the last fold. The highest values are obtained for variables *adiposity* (0.46), *chd* (0.26) and *obesity* (0.23). This supports our

[1] Equation 8.3 and 14.1; Introduction to Machine Learning and Data Mining, 2019; Tue Herlau, Mikkel N. Schmidt, Morten Mørup.

Figure 1: Residuals

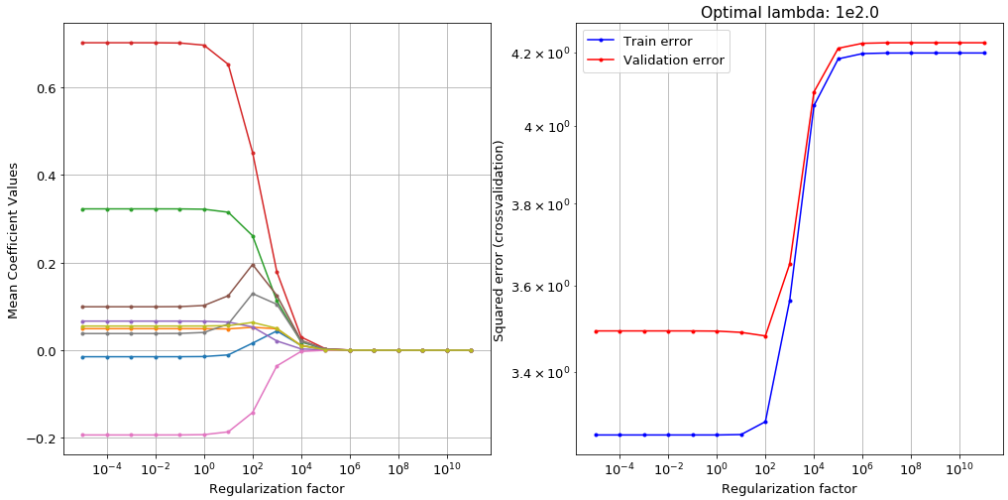


assumption that *ldl* is somehow connected to and *adiposity* and *obesity*. In addition, the *chd* seems to have the highest effect on the level of *ldl*.

Table 1: Overview of weights w^* in the last fold

Weights in last fold (K = 10)	
Offset	4.75
sbp	0.0
tobacco	0.04
chd	0.26
adiposity	0.46
typea	0.08
obesity	0.23
alcohol	-0.13
age	0.13
famhist_present	0.12

Figure 2: Mean coefficient values and squared errors over different regularization factors



2 Regression, part b (s192184)

2.1 Parameters and Models

In this section, we selected the same variable to perform the regression as in Section 2, i.e. our objective was the continuous variable **ldl**. Also, we used the normalized attributes by subtracting the mean and dividing by the standard deviation. Then, we focused on comparing the following three models: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline.

To compute these models, we implemented two-level cross-validation with $K_1 = K_2 = 10$; given the excessively long computing times to tune the ANN model with the number of hidden units h we first tried with $K_1 = K_2 = 5$. However, we played with two parameters to maintain computing times low and to obtain a feasible ANN model, i.e. maximum iterations and hidden units. Initially the maximum iteration parameter was set between 10,000 and 50,000, that way we could quickly compute the optimal number of hidden units (or layers), check the necessary number of iterations for convergence, and then twitch the model for additional robustness. During this initial test phase, we saw that 1) the number of iterations to converge were between 30,000 and 50,000, and 2) by increasing the complexity of the model (by increasing the number of hidden layers h), the error increased.

Therefore, after the initial test runs, we used the following range of h with a maximum iteration value of 50,000.

$$h \in [1:5] \tag{3}$$

For λ we used an interval close to 10^2, as we were expecting the optimal value to be around 10^2 as analyzed in the previous section:

$$\lambda \in [10:200] \tag{4}$$

Furthermore, the error measure we used for the regression is the squared loss per observation:

$$E = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} (y_i - \hat{y}_i)^2 \tag{5}$$

Finally, as the project description required, the baseline model was a simple linear regression with no features, i.e. the mean of y on the training data was used to predict the y on the test data.

2.2 Results and Comparison

After running the models with the parameters from the previous subsection, we obtained the results shown in Table 2. We also added a final row with the estimated generalization error of each model, as explained in Section 3.3 and calculated with formula (7).

Table 2: Summary of 2-level 10-fold CV for regression

Outer Fold		ANN		Linear regression		Baseline
i	*data size*	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	47	2	403.2	80	383.1	450.4
2	47	2	532	80	435.6	527.1
3	46	1	449.3	100	373.3	385.7
4	46	1	351.9	100	311.6	385.9
5	46	1	345.8	100	290.3	399.1
6	46	1	295.5	100	267	357.8
7	46	1	280.3	100	263.4	353.2
8	46	1	576.3	100	499.7	572.4
9	46	1	375	100	387.5	498.1
10	46	1	271.4	100	206.7	355.7
$\hat{E}^{test}$			388.4		342.1	428.8

The results show, given the generalization error, that the "best-performing model" was the normal regression model with an estimated generalization error of 342.1, the "second best-performing model" was the ANN with an estimated generalization error of 388.4, and the "worst-performing model" was the baseline with an estimated generalization error of 428.8. This last standing was expected, since the baseline model is really simple as it only computes the average of the data. What is interesting, is that the specific errors of each model vary between each fold and the optimal regularization parameters h and λ in outer folds 1 and 2 are different, with λ being less strict in those folds compared to the results of the previous section. The difference in the regularization parameter for the linear regression model may be that the observations in outer fold 1 and 2 may have high variance and slow bias, contrary to the slightly-lower variance and slightly-higher bias of the data in observations in the other outer folds. Therefore, a more flexible regularization parameter works better for the first two outer folds of the model. However, as we can see in the results, the optimal regularization parameter of the whole model is argued to be 10^2, as in the previous section, given the results of the other eight folds.

2.3 Statistical test

Given the "close" results of the three models, we're interested in answering the question: which of the three models is the best one and how do they compare between each other? We attempted to answer this question by performing a statistical set (setup II) on the results of the previous subsection. We used the more robust setup II to test our models against variability from different training sets in our statistical estimates, i.e. we used the method layed out in section

11.4.[2]

Table 3: Summary of Setup II statistical test for regression

H_0	p value	Lower CI	Upper CI	Conclusion
$E^{test}_{baseline} - E^{test}_{linear} = 0$	0.001	0.491	1.241	H_0 rejected
$E^{test}_{ANN} - E^{test}_{linear} = 0$	0.689	-0.07	0.101	H_0 not rejected
$E^{test}_{baseline} - E^{test}_{ANN} = 0$	0.002	0.435	1.265	H_0 rejected

For the statistical test, we used $J = K$ to obtain J splits on the data set (train and test) and estimated the generalization error of the three models by taking into account the randomization we want. The results are shown in Table 3. We can see from the results that the p-value between the baseline model and both the ANN model and linear regression model is lower than 0.01-0.02, so it shows significance and we can conclude that the error between these models are not the same. However, between the ANN model and the linear regression model we can't say the same.
In conclusion, we can say that if we were to use our models against new data, we'd not use the baseline model because the linear regression model and the ANN model are better than it. Between these two, we can't say they're identical but their results perform better compared to the baseline model. As a recommendation, I'd look more into the tuning parameters of ANN, since the available GPU on our computers was a restriction and it could've performed better than the linear regression model if we could tweak other parameters and not only the number of hidden layers h.

3 Classification (s196119)

3.1 Our setting

In this section we will perform a classification of the variable *chd*. Recall, that chd a the binary variable which indicate whether a person is diagnosed with a heart disease. Hence, our classification task is to predict whether a person has a heart disease conditional on the nine attributes. To avoid issues from differences in scale and variation, we normalize our nine attributes prior to the classification analysis. Famhist, a binary categorial variable, is transformed to an indicator variable using the one hot encoding principle. Although our data set is imbalanced (we have more non-diagnosed than diagnosed persons), we do not address this problem. The reason being, that we are not told explicitly how to deal with this issue in the project 2 task description.

[2]Introduction to Machine Learning and Data Mining, 2019; Tue Herlau, Mikkel N. Schmidt, Morten Mørup

3.2 Choice of models

We decided to apply KNN as our second model. We did in fact implement ANN in the 2 level K-fold cross validation. Sadly, our laptops do not have a GPU, which resulted in infeasible long computing times for the hyperparameter tuning task of ANN[3]. Hence, for pragmatic reason we decided to choose KNN as our second model. After a serial of trial runs we decided to let the trial space of K_{KNN}, the regularization parameter, be in the interval:

$$K_{\text{KNN}} \in [1 : 50] \tag{6}$$

Our optimal K_{KNN} was always in inside this interval during the trial runs and the test error grew large whenever the K_{KNN} exceeded 40. Hence, we do not expect that the global minimum of test error to be above 50. We tried every K within the interval.

For λ, the regularization parameter of the logistic regression model, we decided let the trial interval be:

$$\lambda \in [10^{-1} : 10^{3}] \tag{7}$$

We initially tried with a very large interval in our trial CV run. We realized that outside the interval stated above, there were never any candidates for the global minimum. Therefore, we reduced the interval to an interval that always included the candidate for the global minimum. Moreover, we decided to have 30 steps in the interval. Finally, we applied the *L2* penalty term and the *liblinear* solver option.

The baseline model does not have a regularization model, as it always pick the majority class. We now proceed to the results of the classification task.

3.3 Results of classification

Table 4 shows the results of our 2-level 10 cross validation classification. We have added a final row with the estimated generalization error of the KNN, logistic regression and baseline model. The estimated generalization error is calculated as:

$$\hat{E}^{gen} = \sum_{i=1}^{10} \frac{|D_i^{test}|}{N} E_i^{test} \tag{8}$$

We applied the most frequent selected model to estimate the generalization error. I.e. 31 nearest neighbours for the KNN method. Our best performing classification model is the logistic regression. We estimate the generalization error of the logistic regression to be 26.4. Our second best performer is the KNN model with a generalization error of 28.4. Finally and unsurprisingly the baseline model is worst performer with a generalization error of 34.6. Recall,

[3]Selecting the optimal level of hidden units in the first hidden layer

that 34.6 percent of the persons in the sample is diagnosed with a heart disease. Hence, it is not very surprising that our baseline model on average guess incorrect 34.6 percent of the time. Despite the fact that the logistic model has a lower estimated generalization error, we cannot not conclude that it is in fact a better model for our classification challenge. This is a consequence of the fundamental statistical uncertainty. In the next sub section we perform a setup II statistical test, to test the hypothesis of whether the logistic regression is a better classifier within a reasonable statistic uncertainty. However, before we do this we would like to add a few more words to table 4. As one can see from the table, the estimated test errors differ quite a bit within each outer fold. We expect this to be a direct consequence of the fact that the data size for each outer fold is relatively small. Hence, the estimated test error is more sensitive to incorrect predictions. That is, a few more incorrect prediction will have a great impact on the test error within each outer fold. Finally, we see that the KNN and Logistic regression are fairly consistent about their optimal choice of regularization parameter. This gives us confident that the optimal choice of regularization parameter is in fact the global optimal choice. We now proceed to the statistical test subsection.

Table 4: Summary of 2-level 10-fold CV for classification

Outer Fold		KNN		Logistic regression		Baseline
i	*data size*	K_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	47	23	23.4	22.1	23.4	36.2
2	47	23	25.5	22.1	21.3	27.7
3	46	31	32.6	11.7	34.8	41.3
4	46	31	19.6	11.7	19.6	30.4
5	46	31	23.9	11.7	26.1	28.3
6	46	31	39.1	11.7	37.0	32.6
7	46	31	26.1	11.7	26.1	32.6
8	46	31	30.4	11.7	23.9	34.8
9	46	31	30.4	11.7	26.1	47.8
10	46	31	32.6	11.7	26.1	34.8
$\hat{E}^{test}$			28.4		26.4	34.6

3.4 Statistical test (Setup II)

We decided to perform a setup II test, because we are interested in evaluating how well we can expect our model to perform on an unknown data set, generated from the same population. In other words: *Should we expect the logistic model to outperform the KNN model on a new data set on heart disease from South African villages?*

To compute the statistical test we follow the approach outlined in method box 11.4.1 (correlated t-test for cross validation) in the text book. We compute a separate CV from the CV that lead to our estimations in table 4. We do this, because we want to avoid performing a statistical test on the same data that we have used to perform model selection of. By computing a new random CV we ensure that CV split for the statistical setup II is independent from the CV used for model selection. We use the most frequent occurring regularization parameter as a model. We applied $J = 10$ outer folds for the statistical test.
The results of the statistical test is stated in table 5. As one can see from the table, we reject (on a five percent significance level) that the baseline model has the same generalization error as the KNN and logistic regression. However, we cannot reject that the generalization error for the KNN and logistic regression is the same in the population. Thus, our conclusion is that if we were given the task to predict heart diseases on a new data set from South African villages, it would be a better approach to use KNN or a logistic regression than a simple baseline model.

Table 5: Summary of Setup II statistical test for classification

H_0	p value	Lower CI	Upper CI	Conclusion
$E^{test}_{baseline} - E^{test}_{logistic} = 0$	0.004	0.033	0.123	H_0 rejected
$E^{test}_{KNN} - E^{test}_{logistic} = 0$	0.559	-0.041	0.071	H_0 not rejected
$E^{test}_{baseline} - E^{test}_{KNN} = 0$	0.031	0.007	0.118	H_0 rejected

4 Discussion

4.1 What did we learn?

From the regression part, we learned that the optimal regularization parameter λ was an important parameter in the behaviour of the model; given it's ability to keep the complexity of the model in an optimal point, the error we obtained was lower. This can be also seen in the ANN model, where the optimal number of hidden layers were always 1 or 2 only, thus by keeping the model simple we obtained good results. Also, from the statistical test, we learned that it is important to do these tests, since it may be easy to see the generalization error results and argue that one model is better than the other just with that parameter in mind. In reality, we'd like to further test our model with new

data never seen before by our model and keep testing and improving it. On a side note, an important constraint during this project was the lack of GPU and processing power to quickly calculate the ANN model; the time-constraint was always an issue and we had to be very selective on the tests we wanted to run when choosing the final parameters of our model.

4.2 How can we relate to research performed on same the data set?

Recall, that the original paper by Rousseauw et al (1984) makes use of a much more detailed data set than ours. It primarily investigates the relationship between chest pain, gender and chd. We do not have access to the chest pain or gender attribute, making it difficult for us to compare our results with their work. Nevertheless, we argue that our classification results show that it would be possible to perform a fairly reliable screening for heart disease based on our nine attributes. However, more work would have to be done, before it could be implemented as a health policy tool. Firstly, more data from new persons would have to collected to investigate whether the accuracy could be improved through leveraging on more data. Secondly, a wider range of models would have to be applied and tested to explore potential accuracy improvements. Thirdly, we would have to address the class imbalance problem. Finally, we would also have to look at other score metrics such as precision, recall, and AUC to better understand where our classification is good (and less good). I.e. for a screening device we would ideally want to eliminate false negatives.