Danmarks
Tekniske
Universitet

DTU

Project 2 - South African Heart Disease

02450 Introduction Machine Learning and Data Mining

GROUP 240

Anders Krenk - s204749
Bence Gattyan - s204773

Section	1	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	4.4	4.5	5
Anders Krenk	0	60	in2.1	in2.1	30	30	70	20	30	20	90	10	90
Bence Gattyan	10	40	in2.1	in2.1	70	70	30	80	70	80	10	90	10
Exam exercise	6.1	6.2	6.3	6.4	6.5								
Anders Krenk	50	50	50	50	50								
Bence Gattyan	50	50	50	50	50								

Table 1: Individual contributions for each section

November 15, 2022

1 Introduction

This is our project 2 that is the second part of our machine learning project. We will continue to use the same data set and now we will do classification and regression on it.

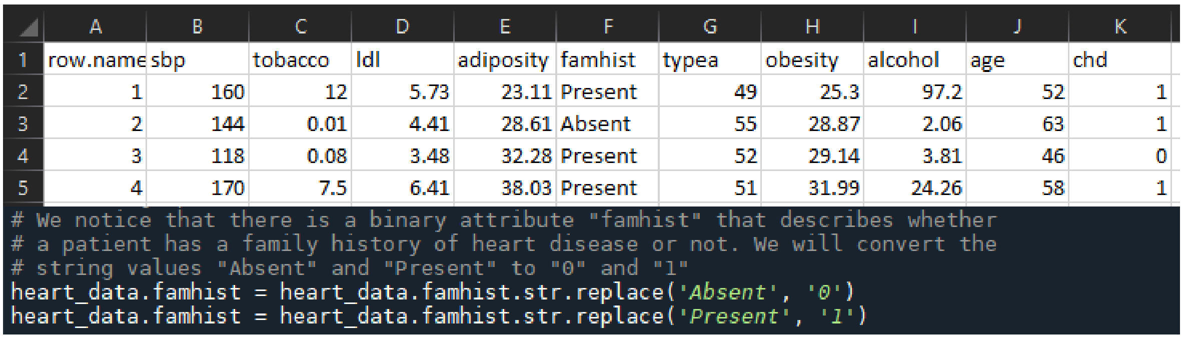
2 Regression - part a

2.1

For regression we have decided to try to predict the age attribute because its a continuous ratio type attribute instead of the binary discrete type CHD response attribute.

Feature Transformation:

We need to binarize famhist attribute so it appears as 1/0 instead of Present/Absent. Then we standardized the data subtracting the mean and dividing by the variance.



	A	B	C	D	E	F	G	H	I	J	K
1	row.name	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
2	1	160	12	5.73	23.11	Present	49	25.3	97.2	52	1
3	2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
4	3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
5	4	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	1

```
# We notice that there is a binary attribute "famhist" that describes whether
# a patient has a family history of heart disease or not. We will convert the
# string values "Absent" and "Present" to "0" and "1"
heart_data.famhist = heart_data.famhist.str.replace('Absent', '0')
heart_data.famhist = heart_data.famhist.str.replace('Present', '1')
```

Figure 1: Famhist is transformed from present/absent to 1/0)

For this part, we will work with regression problems on our data and statistically evaluate the results. We start by introducing a regularization parameter lambda, whose purpose is to control the tradeoff between fitting the data well and overfitting. The regularization parameter will scale other parameters accordingly to achieve the highest accuracy. We find the best regularization parameter with the help of the exercise script 811.*py*. We initialize a sequence of values ranging from 1 - $10 * *6$, which will be the different regularization parameters that we test. We chose this interval because the generalization error drops a bit and then increases. A 5-fold cross-validation is performed (K = 5 and not 10 to decrease runtime as suggested in the project description). On the plot below are the estimated test errors plotted against the different regularization factor values:

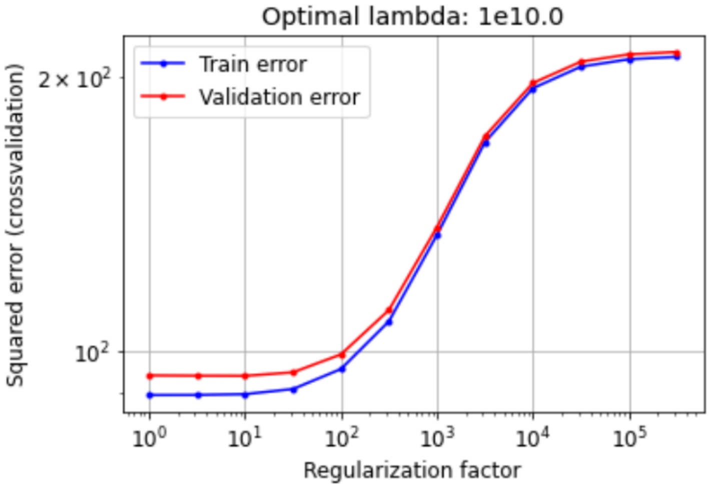


Figure 2: Regularized Linear Regression Generalization Error

The optimal value for the regularization parameter is seen to be lambda = 10, as it is the value that gives the smallest estimated test error. Below is a table showing the different weights for the variables (values adjust slightly with each run):

Attribute	Offset	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	chd
Weight	43.18	2.09	3.09	0.74	8.43	1.58	-0.88	-3.10	-0.22	1.79

These values are interesting because they tell us how much each attribute contributes to the prediction of the model. A larger weight, either negative or positive, will have a bigger influence on the model rather than the weights closer to zero. A positive weight indicates a more probable outcome as opposed to a negative one. The Offset attribute is significantly close to the mean age that the model makes sense, since no influence on the model would result in a prediction of a constant, the average. As for the other attributes, we can see that ldl has the lowest influence by far. Other attributes such as adiposity and tobacco usage are apparently better indicators of age for this model. One particular observation is the weights attributed to obesity and adiposity. As both attributes are synonyms of each other, adiposity meaning "the state of being fat" or "a tendency to become obese", one would assume that these attributes would have similar influence to the outcome, but they aren't at all. They have different signs and stray from 0 significantly, but apart from that the weight distribution seems to make sense across the other attributes. A larger weight will result in an older age prediction and a negative weight will result in younger age.

3 Regression - part b

3.1 Implementing two-level cross-validation

In this part we will compare Linear Regression to Artificial Neural Network and Baseline method to determine which is the best regression model. Baseline here means we don't use input parameters i.e. compute the mean of the training data. Including Baseline model helps put our results into perspective and is common practice in machine learning.

We used Exercise 8.2.6 as a basis of our code and ran K=5 Folds using hidden units h as complexity measure in range $[1, 2, 3, 4, 5, 6, 7, 8, 9]$. We came up with this range when researching ANN's and were guided to have the number of hidden nodes between the number of input parameters (9) and output parameters (1). As for the number of folds we use $K_1 = K_2 = 5$ folds because it was very time-consuming to compute 10 folds. We use random-state=240 to keep the test splits the same for each test. For measuring the loss we used squared loss formula:

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

3.2 Making the table

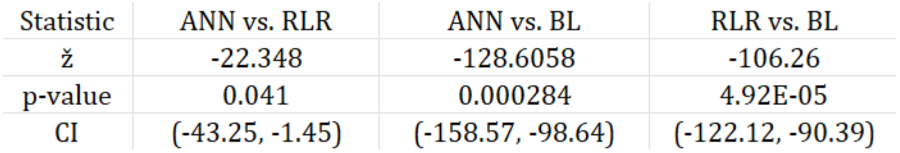
Table 2: ANN vs Linear Regression vs Baseline model results

Outer Fold	ANN		Linear Regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	4	76.10	10	98.8777	191.15
2	2	84.40	10	80.4291	190.632
3	3	77.65	10	116.808	239.244
4	2	83.99	10	103.226	197.44
5	1	83.75	10	118.28	230.453

Again, we found λ=10 as the ideal regularization factor, and ANN provided the best results on an average of 2.4 hidden nodes. ANN has lower errors than Linear Regression and both have lower errors than Baseline which can indicate that both are capable of providing a solution, ANN being the more capable one.

3.3 Statistical evaluation of selected models

We now want to investigate if there is a significant performance difference between the three models. We've compared the models using cross-validation with 5 folds. They will be compared pairwise with paired t-tests (setup 1), so: ANN (5 hidden units) vs. RLR (lambda = 10), ANN vs. BL and RLR vs. BL. We are testing the null hypothesis, that the mean difference between model A and model B will be approximately 0, meaning the performance is the same. If the mean differences aren't the same, the performances differ significantly enough to suggest that one model is better than the other, unless the p-value offers another perspective. Below are the mean differences, p-values and confidence intervals of each t-test (our alpha vale is set to 0.05):



Statistic	ANN vs. RLR	ANN vs. BL	RLR vs. BL
ž	-22.348	-128.6058	-106.26
p-value	0.041	0.000284	4.92E-05
CI	(-43.25, -1.45)	(-158.57, -98.64)	(-122.12, -90.39)

Figure 3: Paired t-tests

Judging by the p-values, ANN and RLR perform significantly better than BL, as the p-value is smaller than our alpha value of 0.05. This means that the null hypothesis can be rejected in both cases, meaning ANN and RLR perform significantly better than the BL. As for ANN vs. RLR, the p-value is just low enough to suggest that the models do perform differently, meaning that the null hypothesis is also rejected. The confidence intervals across all cases also suggests that they perform significantly more different, as 0 is not within the intervals. But as seen earlier with its lower error rates, ANN is the best regression model.

4 Classification

4.1 Our aim in classification

Now we shift our focus back on trying to predict CHD response using classification. We will compare three methods again, Baseline, Logistic Regression and our chosen method, Artificial Neural Network. As mentioned before, CHD response is binary discrete type attribute meaning we are solving a binary classification problem.

4.2 Setting up comparison between methods

For Logistic Regression we use λ as complexity measure between the values $[10^{-1}, 10^{6}]$ with 50 values, for ANN we use the number of hidden units h in range $[1, 2, 3, 4, 5, 6, 7, 8, 9]$ and Baseline, corresponding to Logistic Regression with a bias term and no features. We are using $K_1 = K_2 = 5$ folds as training the ANN is very time consuming. And we use the same seed 240 for making the folds. For error measure we use error rate:

$$E = \frac{\text{Number of misclassified observations}}{N^{test}}$$

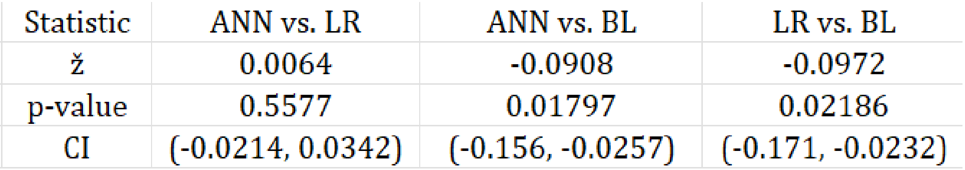
4.3 Making the table

Table 3: ANN vs Logistic Regression vs Baseline model results

Outer Fold	ANN		Logistic Regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	1	0.258	2.68	0.2473	0.4086
2	3	0.301	3.73	0.3118	0.4194
3	3	0.239	37.28	0.2391	0.337
4	1	0.271	3.73	0.2283	0.347
5	3	0.206	0.1	0.2174	0.2174

4.4 Statistical evaluation of selected models

We now want to investigate if there is a significant performance difference between the three models, like we did with the regression models. We've compared the models using cross-validation with 5 folds. They will be compared pairwise with paired t-tests (setup 1), so: ANN (now 4 hidden units) vs. LR (lambda = 1), ANN vs. BL and LR vs. BL. We are testing the null hypothesis, that the mean difference between model A and model B will be approximately 0, meaning the performance is the same. If the mean differences aren't the same, the performances differ significantly enough to suggest that one model is better than the other, unless the p-value offers another perspective. Below are the mean differences, p-values and confidence intervals of each t-test (our alpha vale is set to 0.05):



Statistic	ANN vs. LR	ANN vs. BL	LR vs. BL
ž	0.0064	-0.0908	-0.0972
p-value	0.5577	0.01797	0.02186
CI	(-0.0214, 0.0342)	(-0.156, -0.0257)	(-0.171, -0.0232)

Figure 4: Paired t-tests

As expected, the ANN and LR perform significantly better than the BL. Although the mean differences are almost 0, the low p-values that the tests possess (under 0.05) claim that ANN and LR perform much better than the BL. This means that the null hypothesis can be rejected. Onto ANN vs. LR, the p-value isn't low enough to suggest that the models perform differently, meaning there isn't enough evidence to reject the null hypothesis. They are, however, judging by the p-values and error rates shown earlier, much better models than the BL, and will therefore have similar performance as the best models.

4.5 Training Logistic Regression

Now we train Logistic Regression with the optimal λ value 0.1. We choose this because it gave the best result in our previous test. Using $K = 5$ fold cross-validation error rate was found to be 25.95%

Attribute	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
Weight	0.00901124	0.0727624	0.204316	0.0346567	0.839872	0.0471755	-0.0858392	-0.000747782	0.0384571

Table 4: Weights obtained

The logistic regression makes a prediction based on the weights and the inputs of the attributes. The most important features seem to be famhist, ldl meaning a family history or low density lipoprotein cholesterol make it more likely to have a positive CHD response. Beside these systolic blood pressure, tobacco usage, adiposity, typea behaviour and age also contribute positively to the likeliness of a positive response, while obesity and alcohol usage contribute negatively making CHD response less likely to be positive. In the regression part of the report we were predicting age attribute therefore these values are not immediately relevant to predicting age.

5 Discussion

5.1 Regression and Classification - what we learned

This project gave us an opportunity to work with regression and classification models for the same data set, and with our results we have compared which models (within regression and classification) perform better than others with our dataset.

For the regression models, we used a baseline model, an artificial neural network and a regularized linear regression model. We discovered through statistical evaluation and observation that the ANN performed the best to predict the age attribute of a patient. We think its because the more simple RLR model cant make up for the complexity of the dataset with 9 attributes that an ANN can. However, the RLR table of weights sheds light on which attributes have the biggest influence on age, where some of the values seem to follow a pattern, like high positive weights of tobacco use and adiposity suggesting an older age.

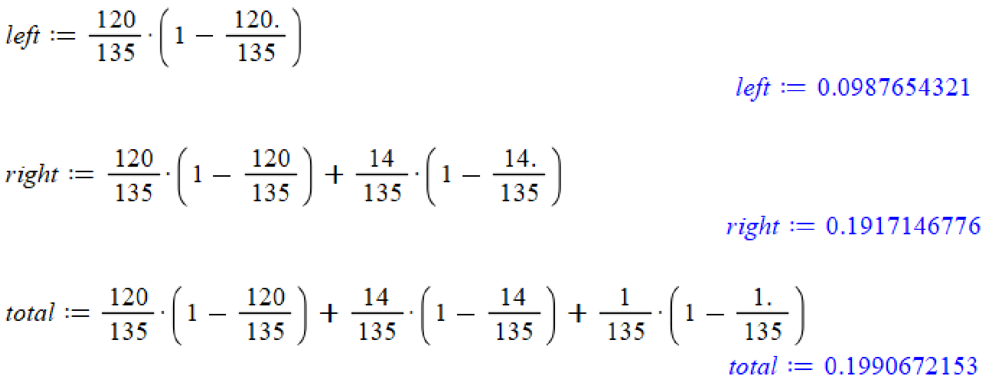
As for the classification models, we used a baseline model, an artificial neural network and a logistic regression model. Our goal was to classify if a patient has a CHD response of negative or positive, represented with the binary values 0 and 1. Again, as expected, the ANN and LR models were significantly better than the BL. Interestingly though, the ANN and LR didnt show enough evidence to perform differently, and so we conclude that they perform somewhat equally well.

5.2 Previous analyses

The ahd dataset was found on https://web.stanford.edu/ hastie/ElemStatLearn/, and is analyzed in the book "The Elements of Statistical Learning". Their analysis consists of predicting CHD response in patients, where they also fit a regularized linear regression. It is not as extensive as our analysis in some cases (cross-validation f.x), and therefore we cannot compare the performances of our analyses.

6 Our attempts at the exam questions

6.1 Question 2



$$left := \frac{120}{135} \cdot \left(1 - \frac{120.}{135}\right)$$

$$left := 0.0987654321$$

$$right := \frac{120}{135} \cdot \left(1 - \frac{120}{135}\right) + \frac{14}{135} \cdot \left(1 - \frac{14.}{135}\right)$$

$$right := 0.1917146776$$

$$total := \frac{120}{135} \cdot \left(1 - \frac{120}{135}\right) + \frac{14}{135} \cdot \left(1 - \frac{14}{135}\right) + \frac{1}{135} \cdot \left(1 - \frac{1.}{135}\right)$$

$$total := 0.1990672153$$

Figure 5

The correct answer is C.

6.2 Question 3

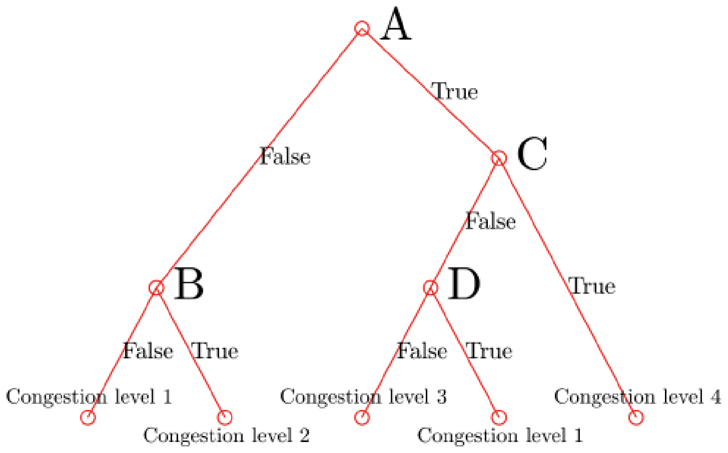


Figure 6

There are 7 x values which will feed into 10 hidden values, the sigmoid non-linearity only has one parameter per link, so this will give $7 * 10 = 70$ parameters. The 10 hidden values will then link to the 4 y-values and this will give $4 * 10 = 40$ parameters. The total amount of parameters is $40 + 70 = 110$ The correct answer is C.

6.3 Question 4

We look at the branches leading to congestion level 4. Node A separates the observations by b1 being larger than -0.76, and then node C separates the observations by b1 being larger than -0.16, and this should give congestion level 4, which matches the classification boundary. The correct answer is D.

6.4 Question 5

The number of models is found by $K1(K2L+1) = 5(4*5+1) = 105$
Time for linear regression: $8*105+1*105 = 945ms$
Time for ANN: $20*105+5*105 = 2625ms$ and t.
Total time: $2625+945 = 3570ms$
The correct answer is C.

References