TECHNICAL UNIVERSITY OF DENMARK

Robert Mundziel (s222885)
Alberto Vendramini (s232103)
Emilie Grønborg Kristensen (s224923)

Project 2

02450 - Introduction to Machine Learning and Data Mining Fall 23

January 18, 2024

Name	Regression	Classification	Explanations	Exam Questions
Robert	20%	40%	33%	33%
Alberto	60%	20%	33%	33%
Emilie	20%	40%	33%	33%

1 Introduction

This project is a natural follow-up on project one where we concluded that the data was of good quality, but that there was little structure and patterns in the data. In this project we will train regression and classification models on the data using different techniques to do so. In the end of each section we will evaluate the performance of each model.

2 Regression

We are doing regression on the attribute Age. By doing this we are aware that in the data set age is given in integers, but to do the regression we are treating it as a continuous attribute.

1.1 Part A

In this section we will begin with examining the most elementary model; Linear Regression. As said before, we would like to predict the variable 'Age' based on all the other attributes, which are:

- SBP (Systolic Blood Pressure)
- Tobacco
- LDL (Low-Density Lipoprotein Cholesterol)
- Adiposity (Body fat percentage)
- Family History (of CHD)
- Type A (Personality type)
- Obesity (BMI)
- Alcohol Consumption
- CHD (Coronary Heart Disease)

We chose to keep all the elements in the Regression because there's nothing that we can be sure doesn't affect the attribute 'Age' directly or indirectly , for example Family History isn't really linked with Age, but it can explain why other attributes have that value - for example SBP (systolic blood pressure).

1.1.1 Data

The first thing that we have done was to exclude the Age column from our X (the attributes used for prediction) and save it as our y (the attribute to be predicted). We standardized the data such that each column has mean 0 and standard deviation of 1. After that we front pushed a column full of one, its purpose is to add a bias term that doesn't depend on a feature to our Linear Regression.

1.1.2 Regularization

In this section we perform Linear Regression with different Regularization Parameters, called λ, and estimate the Generalization Error.
After studying the generalization error in function of λ, we decided to analyze every power of 10, starting from 10^{-5} and ending with 10^{8}.
We chose this range because the error first drops and then increases, this enables us to verify also graphically that the results that we found are correct.

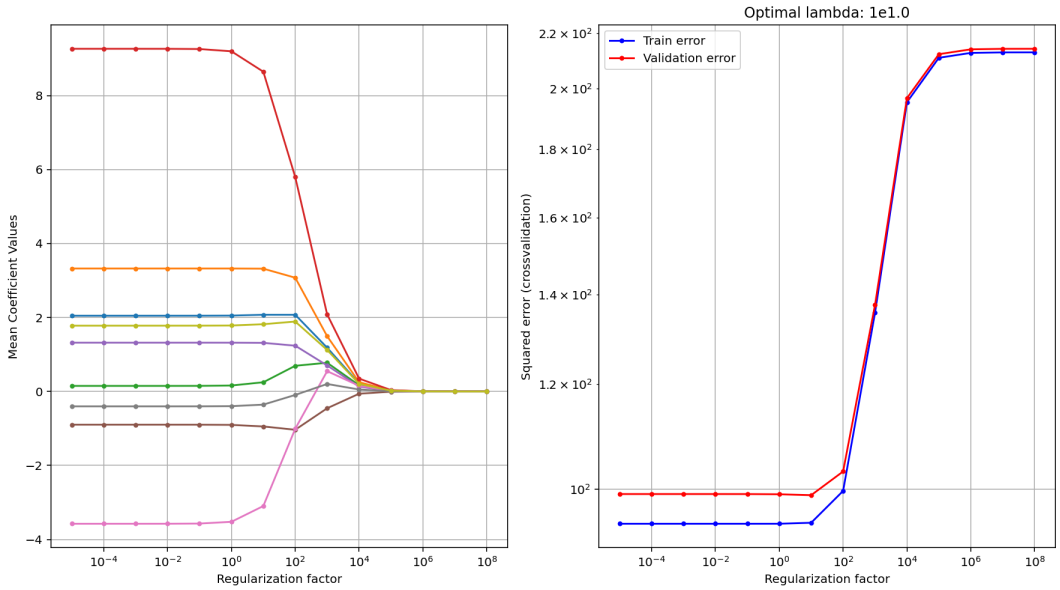

Figure 1: Plots found analyzing lambdas and weights

In the right plot of Figure 1 you can see how the train and test squared error behaves with different λ.
At first sight it is not very clear where it goes to the minimum, but looking closely it can be seen that there's a small dip on $\lambda = 10$ and then it increases drastically. The optimal regularization parameter is therefore $\lambda = 10$.

1.1.3 Output

After evaluating the optimal weights, in order to find a y given a vector X with every attribute (including 1 as first component for the bias), we need to do matrix multiplication.
Given $\tilde{x}$ as the vertical vector containing the attribute values and w^* the vertical vector containing the optimal weights, were $\hat{y}$ is the predicted y, we have:

$$\hat{y} = \tilde{x}^T w^*$$

Since the vector of the weights is a direct index of the importance of an attribute in prediction of Age, we'll go more in the detail.

$$w^* = [42.82 \quad 2.07 \quad 3.32 \quad 0.24 \quad 8.5 \quad 1.31 \quad -0.94 \quad -3.10 \quad -0.35 \quad 1.81]$$

It can be seen that pretty much every attribute has positive weight, besides typea, obesity, alcohol.

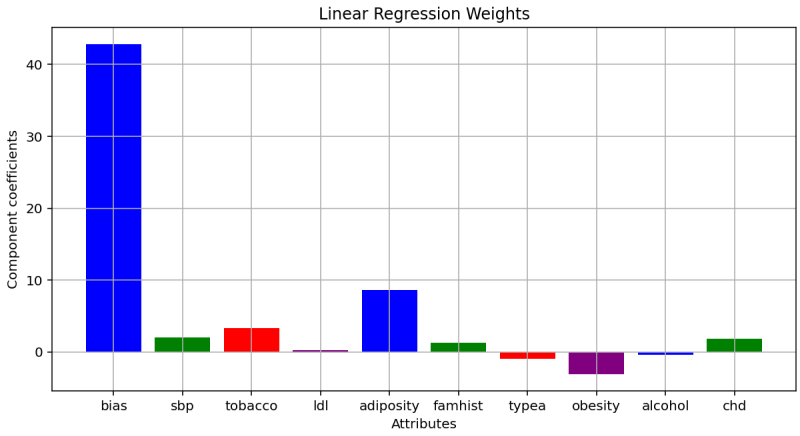

Figure 2: Optimal weights vector found after regularization

Since most of the weights are relatively small, we will now only describe the ones with the highest qbsolute value, since they have the highest impact:
The weight of Adiposity is the highest (excluding the bias), that means that a person with high body fat is more likely to be old, which is what we expected. At the same time decreasing Obesity leads to an increase in age, which at first sight seems contradicting when looking at Adiposity.
Obesity is measured by BMI, so we can expect that BMI decreases for older people, especially older than 42-43 years, while the body fat percentage increases because of a loss in muscle mass, partially replaced with fat due to aging.
Another interesting observation is that there is a positive weight on tobacco. This indicates a correlation between high tobacco consumption and age. This could maybe be avoided by using a larger data set with a more diverse population since tobacco consumption partly depends on cultural factors.[1]

A 1.2 Part B

In this section we will continue to predict 'Age' as before, but with more models. We will compare Regularized Linear Regression, Artificial Neural Network, and Baseline.
As Baseline model we will just compute the age mean of the training data and use this value to predict our y.

[1] https://tobaccofreelife.org/resources/culture-smoking/

1.2.1 Parameters

Concerning Linear Regression we'll stick with the Regularization parameters found in the previous section (Part A), so we'll analyze λ in power of 10 starting with $\lambda = 10^{-5}$ and ending with $\lambda = 10^{8}$. As complexity controlling parameter for Artificial Neural Networks we will use the number of Hidden Units.
After a few test runs we established that increasing the number of hidden units does not have a great positive impact on the generalization error after the value $h \approx 50$ and it's likely to be a waste of time. In figure 3 can be seen that the error after $h \approx 20$ stops decreasing rapidly, therefore we will consider $hs = [1, 5, 10, 15, 20, 25, 30, 35, 40, 50]$ We also tried considering hs in smaller ranges, such as every number between 1 and 30, but didn't find any better result.

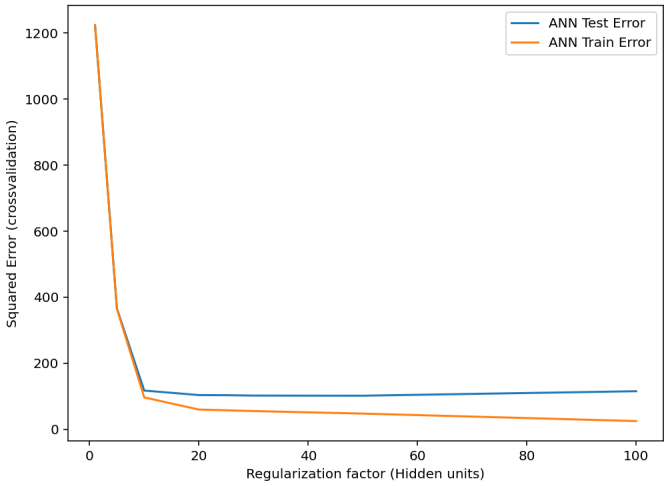

Figure 3: Example of a result of cross validation

1.2.2 Results

In this section we'll see how the error behaves and which parameters have been chosen as optimal.

Lambda

From Table 1 it can be seen that, regarding Linear Regression, the Regularization parameter λ is pretty much stable and almost always equal to $\lambda = 10$. This corresponds to what we have found in 1.1 Part A.

Hidden units

From Table 1 it can be seen that, regarding Artificial Neural Network, the Complexity controlling parameter h ranges between $h = 20$ to $h = 30$. with some exceptions of $h = 50$.

Error

The attribute that we want to predict ranges between 0 and 125.[2] The regression error that we have found is still too high to consider the goal achieved. Rather than precisely predict the age, the model can be used as an index of what the age range of that person can be, keeping in mind that it's not precise.

Outer fold	ANN		Linear Regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	25.0	118.58	10.0	117.66	215.57
2	20.0	62.59	10.0	84.22	218.29
3	30.0	90.19	10.0	113.11	224.21
4	20.0	75.45	10.0	68.83	209.95
5	50.0	146.03	10.0	128.37	208.55
6	50.0	73.28	10.0	61.77	167.45
7	30.0	117.04	10.0	121.56	253.4
8	25.0	94.74	10.0	92.33	223.05
9	20.0	57.2	10.0	73.66	196.76
10	30.0	138.46	1.0	133.23	218.04

Table 1: Cross validation results

1.2.3 Model comparison

In this section we will compare every model with each other in order to state which one is better use.

Baseline vs Linear Regression

Comparing the Baseline model and the Linear Regression model, we find that the confidence interval doesn't contain 0 - it is not even close, and is positive. The p-value is below the significance level of 5%:

$$\text{Confidence interval} = (94.96; 133.16)$$
$$p = 5.30 \cdot 10^{-28}$$

Combining the two results we see that there is significant statistical difference between them and conclude that the Linear Regression model is better that the Baseline.

Baseline vs Artificial Neural Network

Comparing the Baseline model and the Artificial Neural Network we find that the confidence interval doesn't contain 0, it is not even close, and is positive. The p-value is below the significance level of 5% :

$$\text{Confidence interval} = (93.75, 138.59)$$
$$p = 4.94 \cdot 10^{-22}$$

[2] See project 1 for context.

Combining the two results we see that there is significant statistical difference between the model and conclude that the Artificial Neural Network model is better that the Baseline.

Artificial Neural Network vs Linear Regression

Comparing the Artificial Neural Network model and the Linear Regression one we find that the confidence interval contains 0, the center is negative but very close to 0, as can be seen from Figure 5
The p-value is above high the significance level of 5%.

$$\text{Confidence interval} = (-12.25, 7.94)$$
$$p = 0.66$$

Since we cannot conclude that these models are significantly statistically different which leads us to thinking that the two models behave similarly.

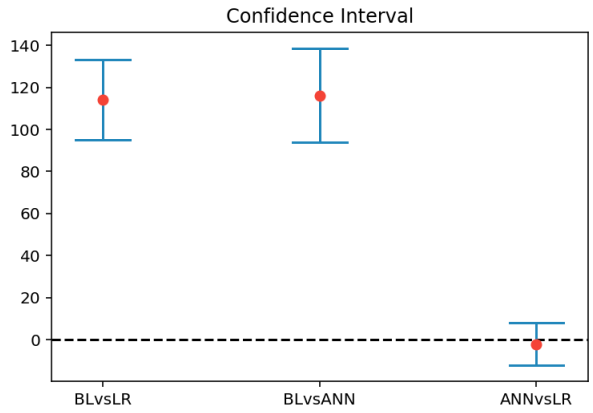

Figure 4: Confidence intervals

3 Classification

For the classification part we have chosen a binary classification problem, where the aim is to predict whether or not a person suffers from heart disease (CHD). This means that it is a binary classification problem.

The models we are training in this section are a baseline model, a logistic regression model and an artificial neural network (ANN). For each model we will do a two fold cross validation where $K_2 = K_1 = 10$.
The baseline model computes the largest class on the training data, and predicts everything in the test-data as belonging to that class.
Before training our logistic regression model and ANN we want to make sure that we do not over fit the train-data. To do this we have made a few test runs of λ-values for the logistic regression model and numbers of hidden layers (h) for the ANN-model to find a range where we expect the

models to perform their best without over fitting the train-data.
The range for λ is power of 10 between -5 and 8
Hidden layers (h) values are h = [1,5,10,15,20,25]

Based off the ranges for parameters we have trained our models and obtained the following results:

Outer fold	ANN		Logistic Regression		Baseline
	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	5	25.50%	10	25.53%	38.30%
2	1	22.17%	10	35.24%	29.79%
3	1	21.05%	10	21.74%	23.91%
4	1	26.96%	10	32.61%	41.83%
5	1	30.43%	10	26.09%	33.12%
6	1	22.44%	10	28.63%	30.68%
7	1	24.76%	10	28.26%	36.90%
8	1	20.53%	10	21.74%	43.48%
9	5	24.67%	10	32.61%	26.09%
10	1	23.55%	10	23.91%	39.13%

Table 2: Cross validation results

The error is calculated by using the formula $E = \frac{\text{Number of missclassified observations}}{N^{test}}$, meaning that the lower the E, the better is the model at classifying the data points in the test set.

Model comparison

We want to evaluate the models statistically, since any difference between the results potentially could be due to statistical uncertainty. We want to be sure that there is actual statistic significance between the model, before making any conclusions.
To do so we are using McNemar's test (Setup I) for comparing classifiers. We are choosing this method because we have a relatively small data set.

Baseline vs Logistic Regression

The confidence level is close to 0, which means that there is very little confidence in the estimate. This suggests a high level of uncertainty about whether the calculated interval contains the true parameter value. The p-value is below the significance level of 5%:

$$\text{Confidence interval} = (-0.33; -0.21)$$
$$p = 1.11 \cdot 10^{-21}$$

Upon comparing both outcomes, it's evident that a notable statistical disparity exists between them, leading us to conclude that the Linear Regression model outperforms the Baseline model significantly.

Baseline vs Artificial Neural Network

Confidence interval shows that the Baseline model and the Artificial Neural Network are not very similar in terms of statistical difference. The p-value is below the significance level of 5% :

$$\text{Confidence interval} = (-0.30, -0.11)$$
$$p = 7.01 \cdot 10^{-17}$$

Artificial Neural Network model outperforms the Baseline upon comparing. There is existence of statistical difference.

Artificial Neural Network vs Logistic Regression

Comparing the Artificial Neural Network model and the Logistic regression model we can find that confidence interval contains 0, but the center is not close to 0. The p-value is above high the significance level of 5%.

$$\text{Confidence interval} = (-0.01, 0.12)$$
$$p = 0.07$$

Those models have the highest value p so we can conclude that they are the most similar models of all of the models.

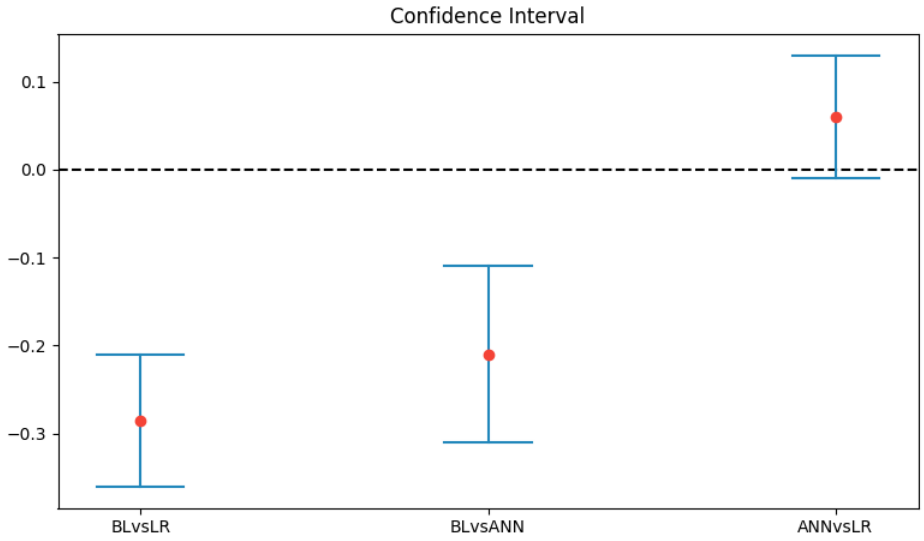

Figure 5: Confidence intervals for classification

In model comparison, the Artificial Neural Network (ANN) demonstrates superior predictive power owing to its adaptability and capacity to capture complex relationships within the data. Logistic Regression, while interpretable, lags behind ANN due to its limitations in handling intricate data structures. The Baseline model, though a starting point, showcases the weakest predictive performance among the three models, lacking the sophistication to capture meaningful patterns.

Logistic regression model

The last thing we want to do in the classification chapter is to take a deeper look at our logistic regression model.
The logistic regression model is given by the formula

$$\theta_i = \sigma(\tilde{x_i}^T w) = \frac{1}{1 + e^{-\tilde{x}_i^T w}}$$

Where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid function. By using the logstic sigmoid function we get a binary output, which is what we need to predict if a data point has heart disease (CHD) or not. This is also why the sigmoid function was not used for regression since the output should be continuous.
To explain logistic regression model we have to inspect the coefficients or weights assigned to each feature in the logistic regression model. Features with non-zero weights in logistic regression are considered relevant for predicting the outcome.

$$w^* = [0.15 \quad 0.33 \quad 0.31 \quad 0.14 \quad 0.41 \quad 0.30 \quad -0.20 \quad 0.01 \quad 0.54]$$

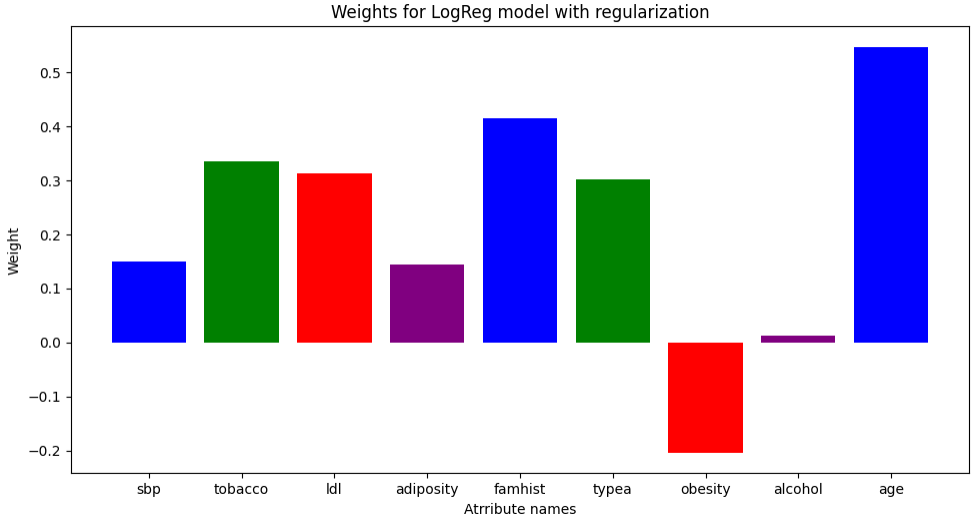

Figure 6: Weights to explain logistic regression model

When data is normalized and standardized before fitting the model, the resulting weights will share a similar scale. Consequently, it's appropriate to interpret that, for instance, the 'age' variable holds the most substantial positive influence in the equation. This interpretation aligns with the expectation that age likely correlates positively with an increased likelihood of heart disease.
Upon comparing these newly obtained weights to those previously identified in our selected regression problem, it becomes evident that both the significance and the extent of impact (including the direction of impact) on the output have changed for most of our attributes.

4 Discussion

Our analysis suggests that effectively predicting whether patients suffer from angiographic heart disease based on dataset parameters can be achieved with high accuracy using machine learning techniques such as logistic regression or classification ANNs. We've gained insight into weighing the trade-offs between resource-intensive training for complex ANNs and opting for simpler yet equally accurate models like logistic regression or possibly other models.

For the regression part of the report we have learned that it is definitely better to use either a linear regression model or an artificial neural network. However, there was no statistically significant difference between the two. But even by using either linear regression or ANN the squared test errors ranged between 57.2-138.46 indicating that the models on average perform with an estimation error of 7.6-11.8 years, which is rather high.

In our examination of the regression problem, we contend that enhancing our model's performance hinges on gathering new data. The current dataset exhibits minimal correlations among its attributes, indicating that improving predictive outcomes relies heavily on augmenting the dataset with additional, more informative features. The error of ANN, logistic regression and baseline was 24.21%, 26.67% and 33.89% respectively. It shows that ANN classification for this dataset was the most accurate and get the best scores. On the other hand the worst is baseline classification where one-third was classified wrong.

Since the data set was meant for prediction of whether or not a person suffered from CHD we were not able to find other studies where they looked at regression with the purpose of predicting the people's age. We were however able to find a machine learning study using the exact same data set as us, to perform a logistic regression on the data set. The study[3] did however exclude the attribute Obesity, so their result is not 100% comparable to ours, but it gives some insights.

For challenging datasets, logistic regression can handle noise and outliers better due to its simplicity and resistance to overfitting. ANNs, while powerful in capturing complex patterns, might struggle with noisy or insufficient data, potentially leading to overfitting and requiring more intricate regularization techniques. For the heart disease dataset we can see that ANN model works better. We consider our goals achieved, even if the errors are large, the results can give us a great index.

[3] https://philarchive.org/archive/KHDEML

5 Solutions for exam problems

1. Answer C;
   We can start by setting the threshold to 0.8 and consider Prediction B, TPR=0.5 FPR=0. Since it is not a point in the ROC plot we can exclude it. Now we set the trheshold to 0.6 and consider Prediction A, TPR=0.5 FPR=0.75. Since it is not a point in the ROC plot we can exlude it. The last 2 options are C and D, we can set the threshold after the 5th point. Considering option D we find TPR=0.5 FRP=0.25 which is not a point on the ROC curve, so the only possible answer is C.

2. Answer C;
   We can start by evaluating the entropies: $I(r) = 1 - \frac{37}{134}$
   $I(v1) = 1 - \frac{1}{1} = 0$
   $I(v2) = 1 - \frac{37}{133}$
   The impurity grain is then: $I(r) - \frac{133}{134} I(v2) = 0.0074$

3. Answer C;
   The first layer has 7 inputs, the second layer has 10 hidden units and the last layer consists of 4 outputs. The edges (and also weights) are then 7*10+10*4=110

4. Answer D;
   Following the branches of the tree and matching the colors in the plot with the visible cuts it is not hard to convince that the right answer is D.

5. Answer C;
   The time for ANN is 5*(4*5+1) * (Ttrain+ Ttest)=2625ms.
   The time for LR is 5*(4*5+1) * (Ttrain +Ttest)=945 ms
   The sum is equal to 3570.0 ms.

6. Answer B;
   The answer can be achieved by calculating manually (or via code) the $\hat{y}_k$ and then evaluating the probabilities. The only answer having k=4 with higher probability is the B.