02450- INTRODUCTION TO MACHINE LEARNING AND DATA MINING

Report 2

Authors:

Andrew Ardueser (s182332)

Alexandros Dorodoulis (s182148)

April 9, 2019

Contents

1 Regression - Part A

1.1 Introduction

The goal of our linear regression approach is to predict a patient's cholesterol level based on a number of other metrics that are typically found in the patient's medical history or could be observed and recorded through a non-invasive procedure. As measurement of cholesterol level is often gathered through the use of a blood test, the utility of a model to accurately predict cholesterol levels could reduce the number of needless invasive procedures. Our target variable ″chol″ is a continuous and ratio variable. Initially, all other variables are going to be used but some of them might be dropped in order to improve the overall estimation score.

In order to avoid bias when analyzing variables of massively varying scales, we normalized the data so the mean and variance for each quantitative variable would be 0 and 1, respectively, by subtracting by the variable's mean and dividing by the standard deviation. Additionally, we used 1-out-of-k coding to transform each categorical variable into k distinct variables with k being the number of classes within the variable.

The quantitative variables that were transformed for the analysis are shown below with a brief description:

> ″ ″:resting blood pressure (in mm Hg on admission to the hospital), continuous and ratio variable″thalach″: maximum heartrate achieved, continuous and ratio variable ″oldpeak″: ST depression induced by exercise relative to rest, continuous and ratio ″slope″: the slope of the peak exercise ST segment (Value 1: slope>0, Value 2: slope=0, Value 3: slope<0), discrete and ordinal variable ″ca″: number of major vessels (0-3) colored by fluoroscopy, discrete and ordinal variable

The 1-out-of-k transformation was performed on the categorical variables of the dataset, shown below with a brief description:

- ″sex″: sex of subject (1 = male; 0 = female), discrete and nominal variable
- ″cp″: chest pain type (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic), discrete and nominal variable
- ″fbs″: fasting blood sugar >120 mg/dl (1 = True, 0 = False), discrete and nominal variable
- ″restecg″: resting electrocardiogram (ECG) results (Value 0: normal, Value 1: having ST-T wave abnormality, Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria), discrete and nominal variable
- ″exang″: exercise-induced angina (1 = yes, 0 = no), discrete and nominal variable
- ″thal″: 3 = normal; 6 = fixed defect; 7 = reversable defect, discrete and nominal variable
- ″target″: presence of heart disease (1 = True, 0 = False), discrete and nominal variable

1.2 Generalization error

In order to find the most effective coefficients for our linear regression model, we introduced a set of regularization parameters λ as a way to minimize the complexity of our model while still aiming for the highest accuracy it can attain. This parameter λ will reduce the level of overfitting in the model.

Including all of our parameters, we see the generalization error for the test and train data sets in Figure 1.

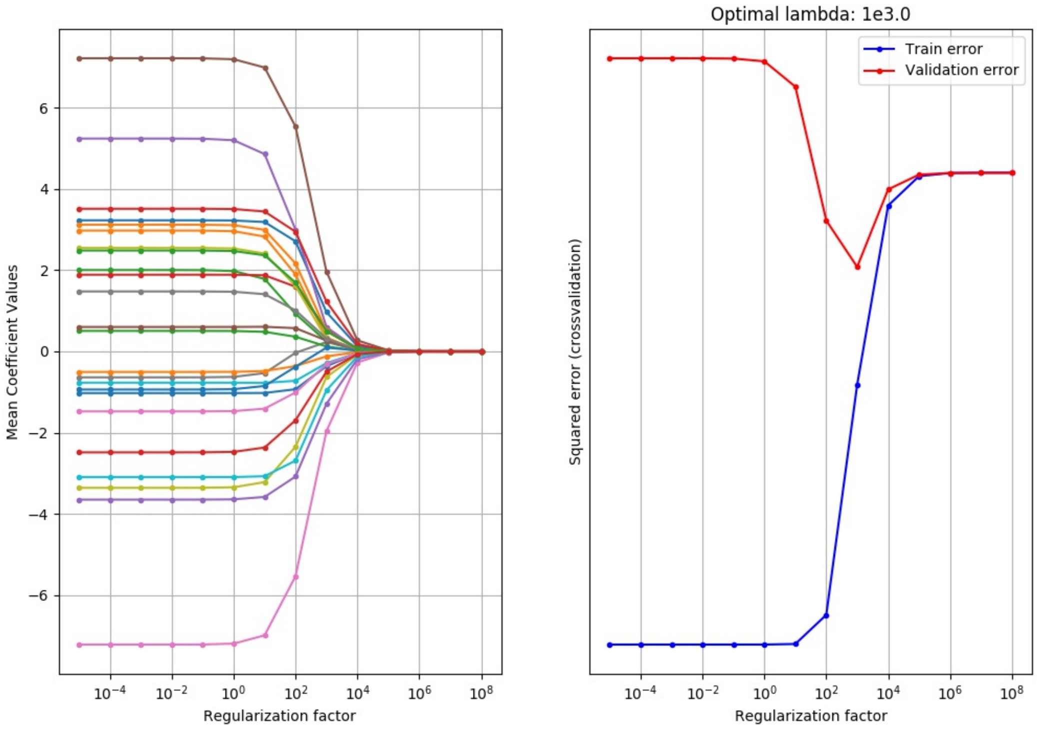

Figure 1: Generalization Error for Test and Train Data

From the left panel, we see that the mean coefficient values converge quickly towards 0 between the λ values of 10^2 and 10^5. We can interpret this as the regularization parameter's effort to reduce the impact of non-significant variables has the most impact between those values.

From the right panel, we see that the generalization error has a minimum that occurs at $\lambda = 10^3$. From this, we are able to conclude that utilizing the variable coefficients calculated from that optimal regularization value will give the most accurate predictions. These coefficients are displayed in Figure 2.

To predict the "chol" value of a patient in the test set, we multiply the each value of the patient's attributes by the corresponding coefficient. The sum of these products and the bias weight value will yield the patient's predicted cholesterol value. To interpret how these variable coefficients actually effect a prediction, let us take a small subset of the features: age, sex, and target. The starting point for every prediction is the bias; the bias is the value on which every other variable and its corresponding coefficient will be added.

Considering a 30 year-old, male patient with a diagnosed heart disease, can use these attributes to get a rough estimate of a cholesterol prediction. Given the following equation:

$$\hat{y} = w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 \tag{1}$$

By substituting in our bias coefficient for w_0, age coefficient for w_1, etc. and age=30 for x_1, sex_0=1 for x_2, the prediction becomes:

$$\hat{y} = 246.767 + 2.001(30) + 1.210(1) + 0.458(1) = 308.465 \tag{2}$$

Feature Name	Coefficient
bias	246.767
age	2.001
trestbps	1.311
thalach	0.310
oldpeak	0.468
ca	0.670
slope	0.206
sex_0	1.210
sex_1	-1.210
cp_0	0.377
cp_1	0.010
cp_2	-0.178
cp_3	-0.209
fbs_0	-0.049
fbs_1	0.049
$restecg_0$	0.920
$restecg_1$	-0.966
$restecg_2$	0.046
$exang_0$	-0.245
$exang_1$	0.245
$thal_0$	-0.070
$thal_1$	-0.333
$thal_2$	0.028
$thal_3$	0.375
$target_0$	0.458
$target_1$	-0.458

Figure 2: Variable Coefficients

To further explain the function of the coefficients by taking age as an example, the coefficient is positive. For every unit of age (in this case, 1 year), the predicted value for cholesterol increases by 2.001. Conversely, the sex_1 variable has a negative value. If a patient is female, the predicted cholesterol value is lowered by 1.210. One is able to draw the inferences that as patients age, their cholesterol level is likely to rise. Additionally, women are likely to have lower cholesterol levels compared to men. Further it follows that, because the parameters are normalized, the features with smaller coefficients will have a smaller impact on the predicted value.

The highest level of serum cholesterol to be considered healthy is 200 mg/dl. As the age range for our data set has an average of 54.4, ranging from 29 to 77 and over half of the patients have a diagnosed heart disease, the bias having a value above the acceptable range is somewhat understandable.

One note-able observation is that $target_1$ (denoting the patient has a heart disease) has a negative associated coefficient. While one would assume a patient with a heart disease would have a higher cholesterol level, this may be counteracted by the fact that a heart disease patient is more likely to be on medication or a diet to control their cholesterol.

2 Regression - Part B

2.1 Two-level corss-validation

This part of the assignment is going to implement a two-level cross-validation between 3 models with K_1=K_2=10. The first model is our baseline which is a linear regression without any features. The second one is going to be a linear regression with λ as a control variable. Finally, the third one is an Artificial Neural Network with a control variable the number of hidden layers.

The range of λ's that is chosen is for exploration is from $10^{-}3$ until 10^7. This range should include the optimal range that was discovered in the previous analysis. The hidden units of an artificial neural network should be between the number of the input parameters and the number of the output parameters. So, the selected range is going to be between 1 and 19 hidden layers.

Outer Fold		ANN	Linear	Regression	Baseline
i	h_i^*	E^test_i	λ_i^*	E^test_i	E^test_i
1	2	0.136390	10^1	0.25320	0.1234
2	2	0.132332	10^1	0.07271	0.26305
3	1	0.1150	10^2	0.19052	0.38284
4	2	0.132	10^1	0.1454	0.3010273
5	2	0.148	10^1	0.23376	0.2154
6	1	0.14369	10^2	0.26408	0.2146
7	1	0.13004	10^2	0.2772	0.24566
8	2	0.17316	10^3	0.17762	0.10699
9	2	0.237212	10^0	0.18027	0.23696
10	2	0.0113123	10^0	0.2540	0.2055

Table 1: Two-level cross-validation table used to compare the three models

From the cross-validation table 1 we can infer that the optimal λ is between 1 and 10^3. This is a bit different from what was discovered when we where searching for the optimal without a two-level cross validation. But, as seen in Figure 1 the optimal range of λ from the cross-validation is inside the range of ’s that the validation error is decreasing.

In order to compare the statistical difference between the three models we are going to use the credibility interval method. The results of this method are the following:

Model	z_L	z_U	Result
Stats for ANN / Baseline	-0.16698	-0.02007	Classifiers are significantly different.
Stats for Linear Reg. / Baseline	-0.10812	0.0588	Classifiers are not significantly different
Stats for ANN / Linear Reg.	-0.13682	-0.000912	Classifiers are significantly different.

Table 2: Statistical comparison of each model pair

From the table 2 above it's clear that the Artificial Neural Network has a significant difference with both the Linear Regression and the Baseline. While the Linear Regression and the baseline don't have a significant difference between them.

From the table 1 it's clear that the ANN model outperforms both the Linear Regression and the Baseline. While the Linear Regression seems to be overall better than the baseline but not in a statistically important degree.

Finally in order to improve our models we could lower the folds of the cross validation from 10 to 5. This might help the train of our models because our dataset is relatively small.

3 Classification

3.1 Problem Explanation

For the classification, we have chosen to predict the ”target” variable, denoting if a patient has been diagnosed with a heart disease. As this is the intended use for the dataset from Kaggle, we believe there will be a good amount of insights to be made. To give further rationalize to the usefulness of solving this problem, the prediction of heart disease based on relatively normal exam results (the least common likely being the cholesterol test and exercise-induced angina), the ability to accurately predict the presence of heart disease would be extremely useful for doctors.

As the output variable is predicting the presence or absence of heart disease, it is a binary classification problem.

3.2 Error Comparison

To compare multiple methods of classification, a baseline model, a regularized logistic regression, and an artificial neural network will be used to predict the target variable in each of 10 training/test sets. The test error by which each model will be evaluated is as follows:

$$E = \frac{\{Number\ of\ misclassified\ observations\}}{N^{test}} \tag{3}$$

The baseline model will simply predict that each patient will have the same heart disease condition as the most populous class (e.g. if more patients in the sample have been diagnosed with heart disease, the model will assume all in the test sample have heart disease). This simplistic model will have no regularizing parameters, merely showing how ineffective the method is and thereby illustrate the difference of a logistic regression model without features to one with features.

The logistic regression model is similar to the linear regression model in that it uses the values of its features in order to train the weights of each variable (e.g. coefficients). However, it deviates in the representations of its output. By using a sigmoid function, the logistic regression model is able to transform its prediction value into a probability that a test observation will be in a given class. The sum of these probabilities always sums to one, by the nature of the sigmoid function. For the logistic regression model, we will have a calculated regularization parameter chosen from a range of acceptable values before ultimately calculating the variable coefficients with the hyper-parameter value with the lowest training error.

Finally, the artificial neural network will iteratively calculate the test error given multiple numbers of hidden neurons. From the resulting prediction, a value ¡=0.5 will be assumed to be class 0 while a value ¿0.5 will be interpreted as being class 1. The lowest test error in each fold will be tracked with the corresponding number of hidden neurons.

Outer Fold	ANN		Logistic	Regression	Baseline
i	h_i^*	E^test_i	λ_i^*	E^test_i	E^test_i
1	1	0.0667	10^0	0.0323	0.4839
2	1	0.1	10^{-2}	0.1290	0.4194
3	3	0.1333	10^{-2}	0.0645	0.3226
4	1	0.1333	10^1	0.2	0.2667
5	1	0.1667	10^{-2}	0.1667	0.4667
6	2	0.2667	10^0	0.2667	0.3667
7	1	0.2333	10^{-1}	0.2333	0.4667
8	1	0.1333	10^{-1}	0.1667	0.5
9	1	0.1333	10^0	0.1333	0.3667
10	1	0.1	10^{-2}	0.1	0.4

Table 3: Two-level cross-validation table used to compare the three models

From the results, we see that the worst performance (by far) was by that of the baseline model. While its accuracy reached 0.2667 at its best fold, this is still far worse than the other models. This is expected, as a model which disregards the features of the dataset is expected to be inaccurate.

From the results of the logistic regression model, we see that the regularization parameter λ is optimized in the range from 10^{-2} to 10^1. Given more time, we would likely explore this smaller range in order to find even more efficient values of λ.

From the ANN, we see that the number of neurons in the hidden layer tends to be one. This makes sense, as the dataset is relatively simple. Without a large number of parameters, we are unlikely to need a large number of hidden neurons.

Finally, we see that the ANN and logistic regression are contentious in terms of efficiency. Many of the folds see the two models equally effective at predicting the target variable. However, a more objective calculation is needed to sufficiently decide if one model is superior. To address this issue, we will use the equation:

$$E_A^{gen} - E_B^{gen} = 1/K \sum_{n=1}^{K} z_k, z_k = E_{A,k}^{test} - E_{B,k}^{test} \tag{4}$$

Comparing the ANN and logistic regression, we find $E_A^{gen} - E_B^{gen} = -.00204$. Given a certainty interval of 0.05 and using a non-standardized Student's t-distribution, we are able to find that there is a significant difference between models if either z_L is greater than 0 or z_U is less than 0.

Comparing the models pairwise, we find the following:

Model	z_L	z_U	Result
ANN/Logistic Reg	-0.0290	0.0238	Not Signif. Different
Baseline/ANN	0.1874	0.3311	Signif. Different
Baseline/Logistic Reg	0.1774	0.3359	Signif. Different

Table 4: Statistical comparison of each model pair

As was already apparent, both the ANN and logistic regression models performed significantly better than the baseline model. However, the ANN and logistic regression appeared to be extremely similar in their effectiveness. Some recommendations for further research into this dataset would be to decrease the range of the regularization parameter for the logistic regression to be from 10^{-2} to 10^{1} to narrow the search for the optimal value of λ. Furthermore, adding a second hidden layer may have a significant effect on the accuracy of the artificial neural network. While increasing the number of neurons in a single level may not have high impact, a second layer with one neuron may provide significant gains in prediction effectiveness. Diving further into these necessary hyper-parameters could yield useful insights and better predictions.

4 Discussion

4.1 Learning

Though we have had basic experience with basic linear regression in the past, this project helped to emphasize the importance of normalizing the data before the analysis. In the beginning, our results were less than acceptable. This was due to our use of completely un-normalized data. Additionally, the concept of the artificial neural network was both intriguing and challenging to implement properly. This project laid some of the foundation that will be inevitably be needed for more complicated machine learning courses.

While the regression form of the artificial neural network, applying it to a classification problem gave a different perspective on how the tool can be applied. The diversity of application of the ANN will be incredibly useful in classes like Deep Learning.

4.2 Previous Analysis

As our dataset was obtained through Kaggle, there are naturally many people that have analyzed it. One example is by Caner Dabakoglu and it can be accessed here. He performs multiple methods of classification to predict the target variable. His results were surprisingly similar to ours. The accuracy of logistic regression was 86.89% compared to our 85.07% . Our ANN had 85.33%, which is comparable to the rest of the methods he used. However, his models with the highest prediction accuracy were the random forest and K-nearest neighbors models. The structure of the data may make these more viable options, but exploring them would be targets for further work.

Another interesting example can be found here. This study managed to achieve a classification score of 93.44%. The main differences from the implementation in this paper are:

1. This research convert different categorical variable into indicator variables.
2. in order to achieve this result the classification was done using Gaussian Naive Bayes instead of Artificial Neural Network.

5 Contribution Table

Section	Contributor(s)
Report Writing	Andrew Ardueser, Alexandros Dorodoulis
Linear Regression Model	Andrew Ardueser
ANN - Regression Model	Alexandros Dorodoulis
ANN - Classification Model	Andrew Ardueser
Logistic Regression/Baseline Model	Andrew Ardueser
Statistical Analysis	Andrew Ardueser, Alexandros Dorodoulis
Discussion	Andrew Ardueser, Alexandros Dorodoulis

Table 5: Contribution Table