Danmarks Tekniske Universitet

DTU

Introduction to Machine Learning and Data Mining

Project 2

Authors

Dimitris Bokos - s213233
Melina Siskou - s213158

November, 2021

Contents

1 Regression - Part A

1.1 Our Regression problem

We will predict LDL choresterol, a continuous ratio variable based on three features: adiposity, obesity and age. The decision was made based on the high correlation findings on our previous Project 1.

We decided on LDL since it is the 3rd principal component of our primary objective which is the variable CHD (coronary heart disease), a binary variable which indicates if a person is diagnosed with heart disease. Therefore, it would be interesting to see if we could predict LDL from the other principal components and test each others correlation. It is also a continuous ratio variable which makes it suitable for a linear regression analysis.

1.2 Introducing Lambdas

We introduced the regularization parameter λ with a range of:

$$\lambda \in [10^{-5} : 10^{5}]$$

We believe that most lambdas will fall between 0.01 to 100 but we decided on a bigger spectrum in order to be more impartial.

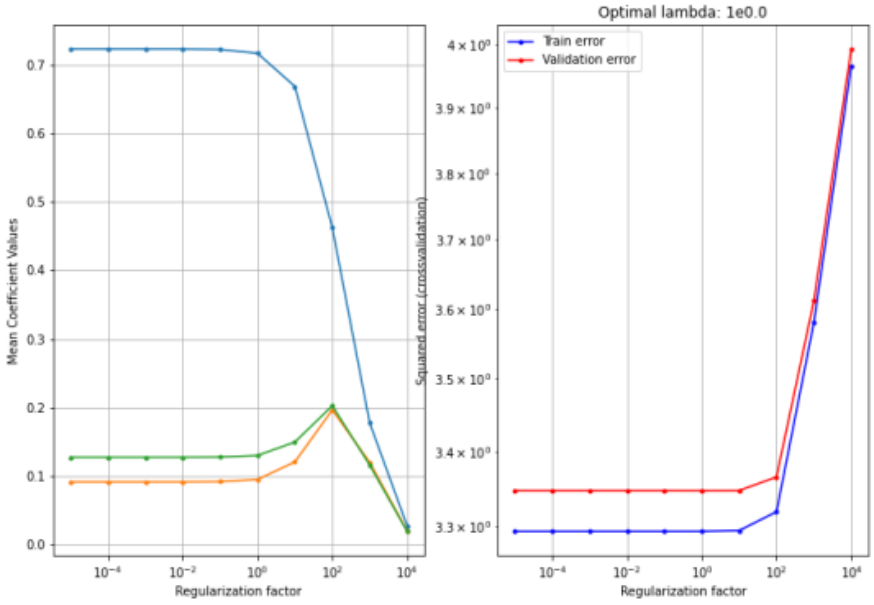

Figure 1: Estimated generalization error as a function of Îť

1.3 Predicting new data observations

A new data observation would be predicted by fitting it to our trained model with the lowest generalization error. Linear regression follows the formula: $y = X\beta + \epsilon$ where y is the

dependent variable, X are the regressors, β our weights and ϵ the loss function.

The selected attributes that we use affect the new prediction in terms of accuracy. The weights of the regressors determine the model, therefore determine the result directly. The results of these tests make sense to us, since they are closely predicting the true value of the observation.

2 Regression - Part B

2.1 Implementing two-level cross-validations

1. For the baseline model we got the mean of the Y value only from the train data for each loop in the inner cross-fold separately in order to keep the mean unbiased and with no knowledge of the values in the test. We also kept the indexes of the train and test splits in order to reuse them on the following two algorithms as well and allow statistical comparison.

2. For the ANN model we decided to let the hidden units be in the interval:

$$h \in [0 : 10]$$

since starting from 0, including 1 and going up to maximum 10 is an optimal number of hidden units.

2.2 Two-Level Cross-Validation table

The three models have been computed on a Two-Level Cross-Validation algorithm. The inner level of cross validation was used to find the optimal model parameters and the outer level estimated the generalization error. We used 10-fold cross validation for inner and outer loops.

Table 1: Two-level cross-validation table used to compare the three models in the linear regression problem

Outer Fold	ANN		Linear Regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	10	3.026	10	3.6720	4.2545
2	10	3.889	10	3.081	4.4289
3	10	3.1710	10	4.188	4.9815
4	8	3.6721	10	5.7001	6.2914
5	8	5.2425	10	3.6620	4.2833
6	10	4.7572	10	3.5995	4.6642
7	4	3.2678	10	2.9432	3.8428
8	10	2.0158	10	2.2875	3.2905
9	10	2.9318	9	2.5488	3.0692
10	10	4.02036	10	3.0799	3.7545
$\overline{E_i^{test}}$		3.1131		3.4742	4.2861

In a quick glance our best performing regression model is the ANN. We estimate the generalization error to be on average 31%. Our second best performer is the linear regression model with a generalization error of 34.7%. Finally, the baseline model is the worst performer with a generalization error of 42.8%, which was expected, since the baseline model didn't fit the model, rather it followed a straight line and performed marginally better that a random model. For example, R^2 gets value 0.1817 and sometimes goes below 0. We expected the ANN model to perform a lot better than the linear regression model, but in the end it was marginally better.

3 Classification

3.1 Our Classification problem

We will now perform a classification of the variable CHD (coronary heart disease), a binary variable which indicates if a person is diagnosed with heart disease. The first 9 attributes will be considered the independent variables of our classification model. CHD will be the labelled class of the model and will either have a positive (1) or a negative (0) predictive outcome. We address issues of scale and variation by standardizing our dataset first.

3.2 Classification Models

Three different classification methods will be applied to the dataset for prediction: logistic regression, k-nearest neighbours and the baseline model, with two levels of cross validation applied to each.

1. For the KNN model, after some trial runs, we decided to let the regularization parameter be in the interval:

$$K_{KNN} \in [1 : 20]$$

2. For the logistic regression model we decided to let the λ regularization parameter be in the interval:

$$\lambda \in [0.1 : 50]$$

since we noticed that were not any candidates for the global minimum outside this range.

3. The baseline model does not have a regularization model as it always picks the majority class.

3.3 Two-Level Cross-Validation

The three models have been computed on a Two-Level Cross-Validation algorithm. The inner level of cross validation was used to find the optimal model parameters and the outer level estimated the generalization error. We used 10-fold cross validation for inner and outer loops. The optimal optimization parameters have been recorded for each fold, and displayed in the following table, as well as the error E_i^{test} associated with each of the folds. This error is determined by dividing the number of misclassified observations by the number of observations of the test set.

Table 2: Two-level cross-validation table used to compare the three models in the classification problem

Outer Fold	KNN		Logistic Regression		Baseline
i	k_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	15	0.2978	22.4	0.2340	0.2765
2	18	0.3191	47.7	0.2765	0.4255
3	5	0.3260	6.6	0.3043	0.3260
4	12	0.3043	46.7	0.2391	0.2826
5	11	0.2391	0.1	0.4130	0.4130
6	19	0.3260	3.4	0.2608	0.3043
7	15	0.3695	6.7	0.2608	0.3043
8	11	0.3913	41.7	0.3260	0.2826
9	19	0.5	22.9	0.2608	0.5
10	17	0.3478	0.1	0.3260	0.3478
$\overline{E_i^{test}}$		0.2901		0.3421	0.3462

Our best performing classification model is the logistic regression. We estimate the generalization error to be on average 29%. Our second best performer is the KNN model

with a generalization error of 34.2%. Finally, the baseline model is the worst performer with a generalization error of 34.6%, which was expected, since 34.6% of our sample is diagnosed with CHD. For logistical regression, the lowest error rate is 23.4% for a λ around 22.4. For KNN, the lowest error rate is 29.7% for k=11.

3.4 Statistical Evaluation

For the classification, we decided to perform a setup I test, using the McNemar test. As you can see from the table below, we reject (on a five percent significance level since alpha = 0.05) that the logistic regression model has the same generalization error as the baseline or the KNN models. This means that there is a significant difference between those pairs and the classifiers have a different proportion of errors on the test set. However, we cannot reject that the generalization error for the KNN and baseline is the same. This means that KNN and baseline models have a similar proportion of errors on the test set. Thus, our conclusion is that if we were given a new dataset to predict CHD, it would be a better approach to use a Logistic Regression.

Table 3: Summary of Setup I statistical test for classification

H_0	p value	Lower CI	Upper CI	Conclusion
Logistic=KNN	0.0106	0.0138	0.0899	reject H_0
Logistic=Baseline	0.0291	0.0078	0.1046	reject H_0
KNN=Baseline	0.9196	-0.0376	0.0462	failed to reject H_0

4 Discussion

4.1 What have we learned

First of all, for the regression part we tried to calculate the levels of LDL cholesterol of our observations based on our other features. We created three models: a linear regression, an artificial neural network and a baseline model. We then trained them on the same splits and compared them on the number of misclassified observations. We have concluded that the best approach would be to choose the artificial neural network model, followed by linear regression as a close second.

We strongly believe that the reason the ANN model did not vastly surpass the other models and the reason why the baseline model was not exponentially worse that the other models, was the small amount of observations in our dataset (only ∼450) and the chosen number of features (3)

Secondly, for the classification part, we wished to predict whether or not a patient will develop a CHD (binary classification problem). We created three models: a logistic regression, a k-nearest-neighbor and a baseline model. We then trained them on the same splits and compared them on the number of misclassified observations. We've concluded that the best approach would be to choose the Logistic Regression model.

Overall, from both regression and classification methodologies we can conclude that finding the optimal regularization parameters is essential in the behaviour of a model and allows us to obtain a lower error.

What is more, the statistical test showed us the importance of checking our models using the optimal parameters found. It allowed us to compare the generalization errors of model pairs and decide which is the most preferable approach when dealing with a possible new dataset.

4.2 Comparison with previous research

Looking through some other research [1] [2] [3] there is generally a lot more pre-processing on the data, so it would make sense that the classifiers from these analysis perform better than ours. However, Neural Networks, Logistic Regression and KNN were found to be quite effective in other research as well. On a larger scale the process we followed is very similar compared to these analysis. What is more, a common conclusion we had with found research is that additional data with the same features would surely help improve our predictions and validate our findings.

5 Exam Problems

- **Question 1 : Option C**

Table 4: Calculations of TPR and FPR at different thresholds

Threshold	A/B/C/D	TPR	FPR	Valid
0.8	A	0.25	0.25	Yes
	B	0.5	0.	No
	C	0.25	0.25	Yes
	D	0.25	0.25	Yes
0.5	A	0.75	1.0	Yes
	C	0.75	1.0	Yes
	D	1.0	0.75	No
0.65	A	0.5	0.75	No
	C	0.75	0.5	Yes

- **Question 3 : Option B**

$7 * 10 * 4 = 280$

- **Question 4 : Option D**

- **Question 5 : Option C**

$NN = 5 * (5 * 4 * (20 + 5) + 20 + 5) = 2625ms$
$LR = 5 * (5 * 4 * (8 + 1) + 8 + 1) = 945ms$
$Total = NN + LR = 3570ms$

6 Collaboration

	s213233	**s213158**
Regression - Part A	✓	
Regression - Part B	✓	
Classification		✓
Discussion	✓	✓

References

[1] Amanda H. Gonsalves, Fadi Thabtah, Rami Mustafa A. Mohammad, and Gurpreet Singh, "Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis [In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies]," ICDLT 2019. doi:10.1145/3342999.3343015.

[2] F. Babic, J. Olejar, Z. Vantova, and J. Paralic, "Predictive and descriptive analysis for heart disease diagnosis," *in Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, pp. 155–163, September 2017.

[3] Hisham Khdair1, Naga M Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," *International Institute of Business and Information Technology, Federation University Associate Adelaide, Australia)*, vol. 12.