Danmarks Tekniske Universitet

DTU

02450 Introduction to Machine Learning & Data Mining

Report #2

Supervised learning: Classification and Regression

Authors

Dmitrij Mordasov - s203350
Mario Cesar Rodríguez - s203359

September 28, 2022

1 Regression

1.1 Regression: Part A

The authors of this report were inspired by the dataset to launch an investigation into how they, as (future) parents, could structure their families and their relationships with their children in a way that lead to (or is at least correlated with) their better academic performance. Thus, a collection of socio-economic factors that the authors deemed that parents may have a degree of control over was selected (13) from the larger attribute pool (33) of the dataset.

Whilst previous student's performance (G1 and G2), an attribute that improves the model's ability to predict the final grade in a subject could be used, in this paper the focus is only on attributes which parents have an effect on, namely:

Investigated attributes	**Discrete / Continuous**	**Types of attributes**
address - Student's home address	Discrete	Nominal
Pstatus - Parent's cohabitation status	Discrete	Nominal
Medu - Mother's education	Discrete	Ordinal
Fedu - Father's education	Discrete	Ordinal
Mjob - Mother's job	Discrete	Nominal
Fjob - Father's job	Discrete	Nominal
famsup - Family educational support	Discrete	Nominal
paid - Extra classes within course subject	Discrete	Nominal
internet - Internet access at home	Discrete	Nominal
nursery - attended nursery school	Discrete	Nominal
famrel - Quality of family relationships	Discrete	Ordinal
Dalc - Workday alcohol consumption	Discrete	Ordinal
Walc - Weekend alcohol consumption	Discrete	Ordinal
Grade - Final student grade (G3)	Discrete	Ratio

Table 1: Attributes (reasonably) within parental control

The final grade (G3) is used as the final outcome (or the result y) in this report. Whilst in essence a discrete value (ranging from 1 to 20 in steps of 1), it was considered to be continuous enough in order to be used for a regression task.

Categorical variables were one-hot encoded (equivalent to 1-of-C encoding) into nominal boolean variables. Based on a method in [1], the exam scores were transformed into a boolean pass/fail label, based on the criterion of achieving at least 10 out of 20 points, used for the classification part of this assignment. Data was also transformed to have a mean of 0 and a standard deviation of 1.

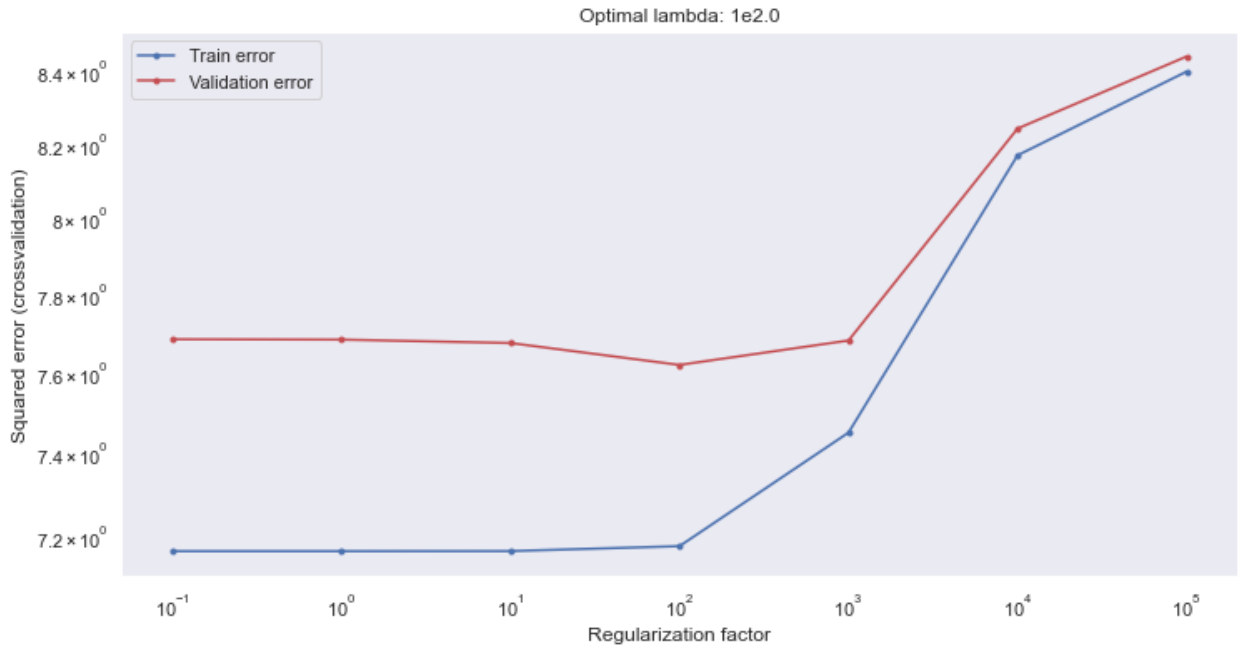

Figure 1: Estimated generalization error as a function of λ for linear regression model

In order to make predictions about data, one must make use of a model or function to represent the so-called "true function" of the data. Using linear regression, it is possible to create functions with different levels of complexity, which is fitted to training data. The complexity of the model is increased until the generalization (validation) error begins to increase, which occurs due to overfitting. Figure 1 shows how the training error increases as a function of the regularization factor. The validation error first decreases, then increases. The optimal regularization parameter corresponds to the first local minimum of the validation (test) error curve.

1.2 Regression: Part B

Three models are fitted to the training data, and then their performance is compared in a regression task of trying to predict a student's performance, given the information in Table 1. The three models compared consist of a regularised linear regression model (as introduced in the previous section), an artificial neural network, and a baseline model.

1.2.1 Method

A two-level K-fold cross-validation (with $K_1 = K_2 = 5$ folds) is implemented with the three used models.

The first model is a regularised linear regression model. In the inner cross-validation fold, model complexity is controlled by checking for an optimum regularisation constant λ, where tested λ values ranges from 50 to 500 in steps of 50, the range of which was determined by

trial runs, and helped by the investigation in the previous section.

The second model is an artificial neural network with one hidden layer, in which the number of hidden nodes is varied in the inner cross-validation fold in order to control the model complexity. Based on trial runs, the range of possible hidden node units ranges from 1 to 5.

For the third model, a featureless linear regression model that predicts the mean of the training set data is implemented as a baseline.

1.2.2 Results

The generalisation error in this case is to be assumed to be equal to the test error, using a mean-squared-error loss function. Table 2 shows us that the artificial neural network model is most accurate with only 1 hidden unit. The optimal regularization factor value for the linear regression model is somewhere between 50 to 500. This agrees with the optimal value found in the previous section. Lastly, the lowest generalization error occurs with the regularised linear regression model.

Outer fold	ANN		Linear regression		Baseline
i	h_i^*	E_i	λ_i^*	E_i	E_i
1	1	5.560	500	3.642	4.737
2	1	9.694	50	7.406	7.792
3	1	9.842	50	8.458	9.630
4	1	11.227	50	9.017	10.124
5	1	22.851	50	19.842	22.155

Table 2: Two-level cross-validation table used to compare the three models for regression

1.2.3 Statistical Test

Setup 1 (t-test) was chosen in order to statistically evaluate if there is a significant performance difference between the fitted ANN, linear regression model and baseline. The authors let the null hypothesis be that both models have the same performance. Table 3 shows how the p-value for ANN vs. Linear regression is the smallest, which is less than the significant p-value level of 0.05. This is evidence that ANN does not have a similar performance as Linear Regression. Furthermore, the Linear Regression model appears to have a smaller number than the significant p-value level, indicating that the performance is different than the baseline model. However, the ANN vs. Baseline p-value is higher than 0.05 so the models are considered to have the similar performance. While p-values are useful to get an indication if models perform similarly, they are less useful for determining a plausible interval of their performance difference. For this, the confidence interval is investigated.

The smallest confidence interval occurs when comparing ANN vs. Linear Baseline. This indicates that the distribution of the models' predictions are similar and within the same

range. Due to the smaller generalisation error for linear regression, the authors recommend the use of the linear regression model.

	ANN vs. Linear Regression	ANN vs. Baseline	Linear Regression vs. Baseline
P-value	0.0042597	0.09539696	0.0199845
Confidence interval	0.77783259, 2.18319917	-0.08336326, 0.68484182	-2.05376922, -0.30578397

Table 3: Performance evaluation for regression models

2 Classification

Once again, three models are fitted to the training data, now trying to predict, given data of Table 1, whether a student has either passed or failed his final Portuguese exam. The three models compared consist of a regularised logistic regression model, a classifier artificial neural network, and a dummy classifier baseline model.

2.1 Method

The two-fold cross-validation setup from the regression task is re-implemented for classification purposes now, using the same model complexity control parameters. However, the regularisation constant range for the logistic regression model now ranges from 10^{-12} to 10^{3} in steps of three orders of magnitude, and for the artificial neural network, the number of hidden nodes range from 1 to 3 - both ranges determined by trial runs.

2.2 Results

The generalisation error in this case is to be assumed to be equal to the test error, using the misclassification rate as the loss function. Table 4 shows how the regularization factors and hidden units stays constant in each fold. It is worth mentioning that the generalization factor is exact the same for the ANN and baseline. This is thought to occur due to the ANN predictions converging to the mean grade.

Outer fold	ANN		Logistic regression		Baseline
i	h_i^*	E_i	λ_i^*	E_i	E_i
1	1	3.846	1.e-12	3.846	3.846
2	1	14.615	1.e-12	14.615	14.615
3	1	6.154	1.e-12	6.923	6.154
4	1	20.769	1.e-12	21.539	20.769
5	1	31.783	1.e-12	32.558	31.783

Table 4: Two-level cross-validation table used to compare the three models for classification

2.3 Statistical Test

For the purposes of conducting the following statistical performance evaluation, the correlated t-test for cross-validation method described in box 11.4.1 was used. Again, for the p-values, the null hypothesis is that the two models have the same performance. Table 5 shows how in all cases, the P-value is greater than the significant p-value level of 0.05. ANN vs. Baseline has the highest p-value, which indicates that the ANN model's performance is most similar to the baseline model. The extremely small confidence interval in the ANN vs. baseline case indicates that both models practically predict the same values. ANN vs. Logistic Regression and Logistic Regression vs. Baseline have the same exact p-value and confidence interval. Since ANN and Baseline have the same generalisation error, the authors recommend the use of either ANN or the baseline for classification purposes.

	ANN vs. Logistic Regression	ANN vs. Baseline	Logistic Regression vs. Baseline
P-value	0.07049	0.62130	0.07049
Confidence interval	-0.00987236, 0.00061774	-3.43850124e-17, 2.32827822e-17	-0.00061774, 0.00987236

Table 5

2.4 Logistic vs linear regression model analysis

Using a regularisation constant λ as obtained in the first section, the weights of the trained model were analysed in order to see their importance for the classification task. It was found through multiple iterations that the authors of this report have found relevant insights during the data exploration in the first report, such as the student's father being a teacher and the highest-attained parental education level having high weights and correlation with good student performance. The model also gave a high weight to the student's access to internet, whether either of the parents stayed at home and if their parents were together or apart - and these three weights have also been often found to be the main features based on which the linear regression model made its decisions.

It was a a bit of a surprise to see the weights being different, as on a first assumption, as the logistic regression model predicts a certain student to be more likely to pass, one could guess that that student is also more likely to achieve a higher grade, hence the reasoning behind why the weights should be similar for the models. However, from the slight difference in weight magnitudes, it is seen that the question of 'what grade is a student most likely to get' is not the same as 'how likely is a student to pass' in this scenario, at least as seen by them models trained in this report.

3 Discussion

The regression problem, at first, showed the authors of this report that the optimum number of hidden layer nodes lays somewhere between 3 and 5 - however, that was for a number of maximum iterations of 10000. Once the maximum iterations for each training of the ANN was increased to 50000, it was found that the ANN with one hidden node consistently outperformed ANNs with multiple hidden nodes. Taking a closer look at the way the model behaved, it seems to mostly guess a value close to the mean of the grade distribution, with a few predictions being very far from the mean based on some learned rule - which does not seem to work too well, seeing its performance with regards to the baseline model that just guesses the mean.

The fact that the regular linear regression model has outperformed the artificial neural network helped hammer in the often-stated statement in machine learning education that not always is a neural network necessary for solving a problem, and that simpler, more conventional machine learning algorithms may be often sufficient. The fact that the authors of this report lack experience training and optimising neural networks may have helped this conclusion as well.

For the classification problem, it has been found that the ANN has basically trained itself to guess the most frequent class in the training data, much like the dummy classifier used as baseline. Any number of hidden units above one increased the misclassification rate of the network, where the same phenomena with the number of training iterations was discovered as in the regression problem. Furthermore, it has been found that for the classification problem, the regularisation term was either minimal, with no change in model error for most cases, yet sometimes a large change in error rate would happen the moment the regularisation term passed a certain (large) threshold - this is most likely due to the discrete 'lumpy' nature of the problem, and would likely change with a larger dataset.

Previous investigation on the dataset [1] has led to regression and classification results superior to those produced in this report. A quantitative comparison is deemed inappropriate due to the large discrepancy in method - whilst in this report, only 13 attributes are used in order to predict the student performance, all of which are socio-economic, in [1], all the attributes are used - including the previous grades of the student, which the researchers have found to be the strongest indicator of future performance and which have significantly improved their model results.

4 Problems

1. D) after evaluating the TPR and FPR for thresholds of $t = 0.5$ and $t = 0.55$, D was found to be the only option with corresponding points on the ROC curve.

2. C) $\Delta = I(r) - \Sigma_{k=1}^{K} \frac{N(v_k)}{N_r} I(v_k)$, where $I(r) = \frac{98}{135}$, $N(r) = 135$, $N(v1) = 1$, $N(v_2) = 134$, $I(v_1) = 0$, $I(v_2) = \frac{97}{134}$
3. A), as we have $7 \cdot 10 + 10 \cdot 4 = 110$ weights (10 hidden nodes, 4 output neurons), and $10 + 4 = 14$ bias terms.
4. D), obtained through following and visualising the decision tree rules.
5. C), $t = 5 \cdot [(25 + 9) + 4 \cdot (5 \cdot 9 + 5 \cdot 25)] = 3570ms$.
6. B), as $p(y = 4|\hat{y}) = 0.73$ whereas $p =$ almost 0 for other observations (and classes for this particular observation).

Work distribution

	Regression	Classification	Discussion	Questions
Dmitrij Mordasov	x	x	x	x
Mario Cesar Rodriguez	x	x	x	x

Table 6

References

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," 2008. Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal.