Project 2

Glass Identification Data Set

Albert Frisch Møller s214610 & Mark Andrawes s214654

	Section 1	Section 2	Section 3	Section 4	Section 5	Exam questions
s214610	40%	40%	60%	60%	40%	60%
s214654	60%	60%	40%	40%	60%	40%

1 Introduction

In Project 1, the Glass Identification Dataset was analyzed and investigated in order to determine if the the main machine learning aim - to perform regression and classification - was feasible. On the basis of our findings from Project 1, it was concluded that this aim is indeed feasible. The purpose of this project is to now apply machine learning techniques in order to successfully perform regression and classification on the dataset. More specifically, the aim of the project is the following: To predict the refractive index of a glass sample based on the rest of its attributes, and to predict the class label of a glass sample.

2 Regression, Part a

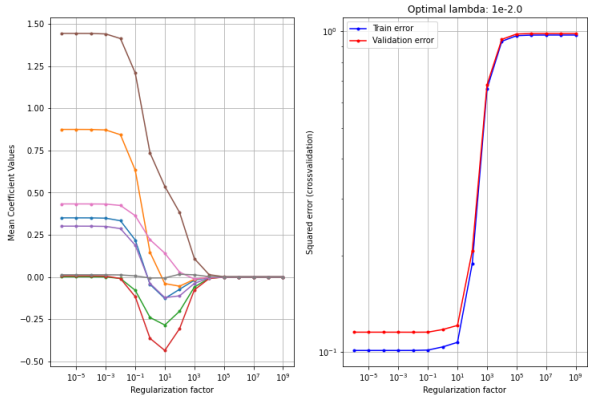
In this section, we wish to perform linear regression on the Glass Identification dataset in order to solve the aforementioned regression problem - namely predicting the refractive index of a sample based on the rest of the attributes. For our dataset, there was no need to apply feature transformations such as one-out-of-K coding. However, we standardized the data matrix by subtracting the mean and dividing by the standard deviation in order to obtain a mean of 0 and a standard deviation of 1 for each column.

2.1 Regularization Parameter & Generalization Error

The regularization parameter controls the complexity of a given model and is used to penalize large weights which are caused by the different scales of the attributes in the dataset. This regularization term is shown in the cost function below (equation 14.3 from the book):

$$E_\lambda(w, w_0) = \|y - w_0 1 - \hat{X}w\|^2 + \lambda\|w\|^2, \lambda \geq 0$$

Where w represents a vector containing the weights of the model, $\lambda\|w\|^2$ is the regularization term, and λ is the regularization constant. A range of values of λ was chosen to be from 10^{-5} to 10^9 (in powers of 10), and the generalization error was calculated for each value using $K = 10$ fold cross validation. It was observed that the optimal λ value that returned the smallest generalization error was $\lambda_{opt} = 10^{-2}$. Given below is a plot of the squared error as a function of the regularization factor:



Figur 1: A plot of the squared error as a function of the regularization factor

Although it is difficult to see from the plot, it can be noted that there is a very slight drop in the error as the regularization factor approaches 10^{-2}, and it then begins to increase significantly afterwards. As the generalization error reaches a minimum at $\lambda = 10^{-2}$, then the bias and variance trade-off is at its optimal point.

2.2 Results

The optimal regularized linear model was found to be,

$$y = -0.01 + 0.34x_1 + 0.82x_2 - 0.02x_3 + 0.01x_4 + 0.28x_5 + 1.42x_6 + 0.49x_7 + 0.01x_8$$

where the x parameters denote respectively the Na, Mg, Al, Si, K, Ca, Ba, and Fe attributes. From the chosen optimal regularized linear model we see that the weights of Ca, Mg and Ba have a strong positive weight and therefore influence on predicting the refractive index. The result makes sense as we that almost all the weights are positive, which will result in a positive predicted refractive index value. In project 1, we also saw that there was a strong positive correlation between refractive index and Ca content. It therefore makes sense that the weight of Ca is the most positive. The refractive index of a new data observation is predicted by replacing the x parameters with their respective value, and then computing a y-value which denotes the predicted refractive index value.

3 Regression, Part b

In this section, we wish to compare three different regression models, namely the regularized linear regression model, an artificial neural network and a baseline model. The aim of this investigation is to first determine if their is a convergence in terms of the model's complexity controlling parameters, i.e. if the $K = 10$ folds will select the same optimal values for the number of hidden units and the regularization parameter. We also wish to determine if the three models

have significantly different performance.

3.1 Description of the Three Models

The number of hidden units was used as the complexity-controlling parameter for the ANN. The regularization parameter, λ, was used as the complexity-controlling parameter for the regularized linear model. The squared loss per observation was used as an error measure for each of the models. As our data-set is relatively small, we decided to use hidden units, h, in the range from 1 to 10. This would also allow us to see if their was a convergence in terms of the optimal number of hidden units being selected within each inner fold. Note that a minimum squared error loss function was used for training the ANN model. Similarly, the chosen range for λ was in the range of 10^{-5} to 10^{9}. This range would ensure that all regularization parameters were covered. It is to be expected that the optimal λ does not lie in either of the extreme endpoints, as a model that balances bias and variance is the most optimal.

3.2 Comparison of their Performance

Two level cross-validation was performed to ensure that optimal values for λ and h, which minimized the squared loss per observation, could be determined and used in the outer fold. Table 1. displays the obtained optimal complexity parameters and the squared loss per observation in the outer fold i:

Outer fold, i	ANN, h_i	ANN, E_i	LR, λ_i	LR, E_i	Baseline, E_i
1	2	1.13	0.1	0.15	1.13
2	1	1.41	0.1	0.049	1.41
3	10	0.84	0.00001	0.14	0.84
4	4	0.21	0.01	0.034	0.21
5	2	0.18	0.1	0.27	1.18
6	1	2.05	0.1	0.091	2.05
7	4	0.79	0.01	0.065	0.79
8	1	0.55	0.01	0.037	0.55
9	2	1.24	0.01	0.076	1.24
10	2	0.67	0.00001	0.24	0.68

Table 1: A Table displaying the regression results for the 3 different models. Note that i denotes the fold i in the $K_1 = 10$ folds.

From the above table, we can see that the optimal λ value which appears the most frequently in the 10 folds, is $\lambda = 0.01$. We notice that this is the same optimal value found in *Regression, Part a.* We also notice that $h = 2$ appears the most frequent.

3.3 Statistical Evaluation

In order to determine if the three models have significant performance differences, a paired t-test is to be computed. Our null-hypotheses for all 3 paired t-tests are: $H_0 : E^{gen}_{ANN} - E^{gen}_{Baseline} = 0$ vs. $H_1 : E^{gen}_{ANN} - E^{gen}_{Baseline} \neq 0$, $H_0 : E^{gen}_{ANN} - E^{gen}_{Linear} = 0$ vs. $H_1 : E^{gen}_{ANN} - E^{gen}_{Linear} \neq 0$ and $H_0 : E^{gen}_{Linear} - E^{gen}_{Baseline} = 0$ vs. $H_1 : E^{gen}_{Linear} - E^{gen}_{Baseline} \neq 0$. Given below are the computed p-values and confidence intervals:

	p-value	confidence interval
ANN vs. Linear Reg.	0.00039	[0.523, 1.259]
ANN vs. Baseline	0.99	[-0.00167, 0.00165]
Linear Reg. vs. Baseline	0.0039	[0.523, 1.259]

From the above table, we see that the p-value for the paired t-test between ANN and the regularized linear regression model is very low. Hence we need to reject the null-hypothesis, and therefore accept the alternative hypothesis, namely $H_1 : E^{gen}_{ANN} - E^{gen}_{Linear} \neq 0$, and therefore that the is a large difference in model performance. We can also see that the p-value is very large for the comparison between ANN and the baseline. Hence, we have to accept the null hypothesis, namely $H_0 : E^{gen}_{ANN} - E^{gen}_{Baseline} = 0$. This ultimately means that the 2 models have the same model performance. We also observe that the p-value for the comparison between the linear regression model and the baseline is very low, which signifies that we have to reject the null-hypothesis and thereby conclude that the 2 models do not have the same performance. All in all, we can conclude by looking at Table. 1. and the p-values and confidence intervals, that the regularized linear regression model outperforms both the ANN and baseline model. We can also conclude that the ANN and baseline have the same performance. The baseline model has proven to be quite successful. This is certainly due to the range of refractive index values being very narrow, namely in the interval of $[1.51115, 1.53393]$. A model that computes the average within each fold, within this narrow range, will therefore predict values that lie close to the true value. Based on the above analysis, for predicting the refractive index, we recommend using a regularized linear regression model with $\lambda = 0.01$. This model ultimately had the greatest performance.

4 Classification

4.1 The Classification Problem

In this section, we wish to perform classification on the Glass Identification data-set in order to solve the aforementioned classification problem - namely predicting the class given its forensic content, i.e. the 8 different attributes. We therefore want to classify a piece of glass as belonging to one of the classes: "building windows 1", "building windows 2", "vehicle windows 1", "vehicle windows 2", "containers", "tableware", or "headlamps."We therefore wish to compare 3 different

classification models, namely regularized logistic regression (RLR), a neural network (ANN) and a baseline model.

4.2 Description of The Three Models

The number of hidden units was used as the complexity-controlling parameter for the ANN. We again decided to use hidden units, h, in the range from 1 to 10. This would allow us to see if the optimal number of hidden units would converge. As our classification problem is multi-class, the cross entropy loss function was used to train the ANNs. The regularization parameter, λ, was used as the complexity-controlling parameter for the regularized linear model. The number of miss classifications was used as an error measure for all 3 models. The chosen range for λ was in the range of 10^{-5} to 10^{9}. This again would ensure that all regularization parameters were covered. The baseline classifier was modelled such that it predicted the class according to the most frequent class in y_{train} set for each cross-validation fold.

4.3 Results

Given below is a table displaying the results obtained by performing 2 level cross-validation on the 3 classification models:

Outer fold, i	ANN, h_i	ANN, E_i	RLR, λ_i	RLR, E_i	Baseline, E_i
1	5	0.59	1.00	0.58	0.86
2	9	0.27	1.00	0.49	0.59
3	4	0.32	10^{-5}	0.42	0.73
4	2	0.36	0.01	0.51	0.55
5	2	0.43	10	0.47	0.86
6	2	0.43	0.01	0.56	0.76
7	5	0.24	0.01	0.46	0.67
8	4	0.38	10^{-6}	0.59	0.76
9	3	0.23	10^{-6}	0.50	0.57
10	1	0.33	10	0.59	0.67

Table 2: A Table displaying the classification results for the 3 different models. Note that i denotes the fold i in the $K_1 = 10$ folds.

From the above table, it becomes apparent that the most frequent number of optimal hidden units is $h = 2$. In addition, the most frequent value for the regularization parameter for the logistic regression classifier is $\lambda = 0.01$. This is the same λ value as for the optimal regularized linear regression model.

4.4 Statistical Evaluation

In order to determine if the three models have significant performance differences, a McNemar test is to be pairwise computed. Our null-hypotheses for the 3 McNemar tests are: $H_0 : E^{gen}_{ANN} - E^{gen}_{Baseline} = 0$ vs. $H_1 : E^{gen}_{ANN} - E^{gen}_{Baseline} \neq 0$, $H_0 : E^{gen}_{ANN} - E^{gen}_{Logistic} = 0$ vs. $H_1 : E^{gen}_{ANN} - E^{gen}_{Logistic} \neq 0$ and $H_0 : E^{gen}_{Logistic} - E^{gen}_{Baseline} = 0$ vs. $H_1 : E^{gen}_{Linear} - E^{gen}_{Baseline} \neq 0$. Given below are the computed p-values and confidence intervals:

	p-value	confidence interval
Logistic Reg. vs. ANN.	0.860	[-0.00167, 0.0610]
ANN vs. Baseline	0.033	[-0.174, -0.0124]
Logistic Reg. vs. Baseline	0.049	[-0.168, -0.000137]

From the above table, we notice a high p-value for the comparison between the logistic regression and ANN models. We therefore have to accept the null-hypothesis and conclude that both models have the same model performance. However, we can conclude that the both the ANN and Logistic regression have significantly better performance that the baseline model, as their respective p-values are both smaller than the chosen significance level, $\alpha = 0.05$. Based on the above analysis, we recommend either using a logistic regression model with $\lambda = 0.01$ or an ANN with $h = 2$ and cross entropy loss function, as these models clearly outperformed the baseline model. We do not have enough statistical evidence to claim that one is better than the other.

4.5 Training a Logistic Regression Model

We fitted a multi-class logistic regression model to our dataset $\mathbf{X}$ and class labels, $\mathbf{y}$. The logistic regression model had regularization parameter, $\lambda = 0.01$. Given below is a table displaying the weights of each attribute in the regularized logistic regression model:

Intercept	Na	Mg	Al	Si	K	Ca	Ba	Fe
4.42	-1.55	-1.34	5.23	-3.15	-0.24	2.90	3.52	2.21

Table 3: A Table containing the weights of each attribute, as well as the intercept, of the regularized logistic regression model.

We notice that the most positive weight belongs to the attribute Al. It is therefore the Al content that has the most positive impact on the class prediction. Contrarily, the most negative weight belongs to the attribute Si. Hence the Si content has the most negative impact on the class prediction. We also see that the attribute K has the least affect on the class prediction. For the regularized linear regression model, we saw that Ca, Mg and Ba had the strongest positive weights and thereby the strongest effect on the predicted refractive index values. Therefore, the same features, i.e. attributes are not deemed relevant for both the linear regression model and the logistic regression model. The multinomial regularized logistic regression model makes a

prediction by weighting the data observation (the values of the 8 attributes) according to the weights shown in the figure above and thereby computes the probabilities - from 0 to 1 of the data observation belonging to each of the classes. The class which has the greatest computed probability is then assigned to the data observation.

5 Discussion & Summary

Having performed linear regression in order to determine the refractive index based on the rest of the attributes (Section 2), we found that the Ca attribute had the strongest positive weight. This aligns with our findings in Project 1, which showed that the correlation between refractive index and Ca was strongly positive. Additionally, the regularization parameter that minimized the generalization error was found to be $\lambda_{opt} = 10^{-2}$.

Having trained and compared the performance of three regression models (Linear regression, ANN, and baseline) using two level cross-validation, a statistical evaluation showed that the performance of the linear regression was different from the performances of ANN and the baseline model However, it was found that the performances of the baseline model and ANN were the same. This was due to the narrow range of values of refractive index in the dataset, meaning that the baseline was able to predict values that lie closely to the true value. All in all, the p-values and confidence intervals show that the regularized linear regression model outperforms both the ANN and baseline models.

For classification, the performances of three models (Logistic regression, ANN, and baseline) were compared using the McNemar test. This statistical evaluation showed that the baseline model performed differently from the logistic regression and ANN models, and that the logistic regression and ANN models performed similarly. Based on the p-values and confidence intervals, it was concluded that the ANN and logistic regression models clearly outperformed the baseline model, however, we cannot conclude which model was best out of the three.

Although the Glass Identification dataset has been analyzed previously, none of the studies shown on the website are available online. Therefore, we cannot compare our findings to previous studies.

6 Exam Problems

6.1 Question 1

Answer: B. We see that for our first chosen threshold, the $FPR = 0$ and the $TPR = 0.25$. Hence our first 2 points need to be correctly classified. This only happens in Prediction B, where $\frac{2}{8} = 0.25$.

6.2 Question 2

Answer: C. We construct the split according to $x_7 = 2$. At the root we have $37 : 31 : 33 : 34$, at the left branch we have $0 : 1 : 0 : 0$ and at the right branch we have $37 : 30 : 33 : 34$. We then compute the class error at each partition, i.e. $ClassError(r) = 1 - (\frac{37}{135}) = \frac{98}{135}$, and then compute the purity gain. We obtained the result of 0.00741

6.3 Question 3

Answer: C. In going from the input layer to the hidden layer with 10 units, there are $7 * 10 = 70$ parameters that need to be trained. When going from hidden layer to output layer there are $10 * 4 = 40$ parameters to be trained. Hence in total there are $40 + 70 = 110$.

6.4 Question 4

Answer: D. We know that congestion level 3 lies between $b_1 = [-0.76, -0.16]$ and $b_2 = [-3, 0.03]$. Only options B and D correctly classify this. We also know that congestion level 2 lies between $b_1 = [-3, -0.76]$ and $b_2 = [0, 3]$. Out of B and D, only D correctly classifies this. Therefore it must be D.

6.5 Question 5

Answer: C. The number of models trained for the ANN and regression model is $K_1(K_2S + 1) = 105$ each, where $S = 5$. Therefore the total time is $105(20 + 5 + 8 + 1) = 3570$.

6.6 Question 6

Answer: B. By inserting $k = 4$ and computing the outcome of $P(y = k|\hat{y})$, we see that the observation $b = [-0.6, -1.6]^T$. Therefore the answer is B.

7 References

UCI Machine Learning Repository, Glass Identification Data Set, Data donated in 1987, `https://archive.ics.uci.edu/ml/datasets/glass+identification`