Danmarks Tekniske Universitet

DTU

Project 2 - Supervised learning: Classification and regression

Authors

Eleni Kiachaki - s222777
Anna Melidi - s222757
Kledi Salla - s230702

April 17, 2023

Contents

1 Contribution

	1 Regression	2 Classification	3 Discussion	4 Exam Problems
1. **s222777**	50%	0%	33.3%	33.3%
2. **s222757**	0%	75%	33.3%	33.3%
3. **s230702**	50%	25%	33.3%	33.3%

Table 1: Contribution Table

2 Regression

2.1 Part A

The present section includes the solution of a relevant regression problem for the Breast Cancer data set, as well as the statistical evaluation of the subsequent result.

2.1.1 Target variable and data preprocessing

Since linear regression is going to be applied, we cannot use the class of the data set, which is the target variable of it. For this purpose, a continuous variable must be selected that is highly correlated to a number of independent to each other attributes so as to use the latter ones for the prediction of the former. In order for these requirements to be met, we decided to predict HOMA, that stands for Homeostasis Model Assessment, based on the attributes insulin, that shows an almost ultimate correlation (0.93), leptin, resistin and MCP-1, that are all positively correlated to HOMA with a coefficient of 0.3 on average. Glucose was discarded, as it is highly correlated to insulin (0.7).
In order to perform the linear regression, the values were standarized by subtracting the mean and dividing with the standard deviation, obtaining, in this way, values with mean 0 and standard deviation 1 for each attribute.

2.1.2 Regularization

Regularization was achieved by using the L2 regularization method or in other words Ridge regression. Ridge regression works by introducing a penalty term into the cost function which is proportional to the sum of the squares of the coefficients. As a result, the optimization issue gets easier to solve and the coefficients shrink. This penalty term pushes the model to strike a compromise between effectively fitting the training data and being simple. The amount of the penalty can be fine-tuned using the regularization parameter lambda. It is vital to choose a decent value for lambda. When lambda is zero, the penalty factor has no impact, and ridge regression produces the standard least squares coefficients. However, as lambda increases to infinity, the effect of the shrinkage penalty increases, and the ridge regression coefficients approach zero. In order to perform this kind of regularization the Ridge function was used from the sklearn library while seven lambdas of equal distance were chosen between the range of 10^{-3} and 10^{3}. The results can be seen in figure 1. One thing to note is that higher ranges were checked for lambda and they gave off the same optimal lambda.

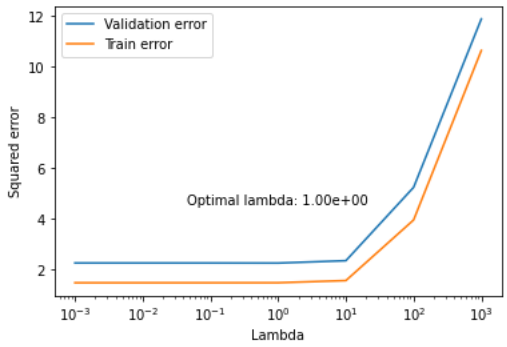

Figure 1: Training error and generalization error for different values of the regularization parameter λ

As we can see, the graph is stable up until lambda reaches 10 after which the MSE starts to increase.This means that the model is not over fitting the data and is achieving a good balance between bias and variance, however after lambda=10, the model becomes too constrained and the MSE starts to increase due to an increase in bias.

2.1.3 Selected Model

Attribute	wi
Intercept (w_0)	2.69
Insulin	3.22
Leptin	0.15
Resistin	0.22
MCP.1	0.30

Table 2: Mean coefficients of the model with $\lambda = 1$

From the above table we can see the weights of each attribute along side the intercept of the linear equation we get from the regression process. As we can see, all the weights are positive with the highest one being that of Insulin. That being said, we can see that these attributes have a positive impact on predicting the target attribute and are in agreement with the correlation heat-map of project 1 and so if a new dataset was to be tested, Insulin, Leptin , Resistin and MCP.1 would play an important role in predicting HOMA.

2.2 Part B

In this part, three different Machine Learning models are going to be compared: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline in the regression problem defined previously. The aim is to investigate whether one model is better than the other or if either model performs better than a trivial baseline. In order to answer these questions, two-level cross-validation was applied, followed by statistical evaluation of the difference observed among the models' performance.

2.2.1 Preprocessing

Before proceeding to the models' construction and comparison, it is necessary to determine which values are going to be used for the regularization parameter λ , with regard to the linear regression model, as well as for the hidden units number, with regard to the Artificial Neural Network. As far as the linear regression model is concerned, according to the previous results, the minimum generalization error was observed when:

$$\lambda = 10^0$$

Therefore, all the values close to 1 were collected, and our target lambdas were obtained:

$$\lambda = [10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}]$$

For investigating the suitable hidden units number (h) of our ANN, we used 3-fold cross validation on a set of values for h, while changing the max iterations parameter. After these tests we concluded on 10000 max iterations, as we noticed that at this iteration value, the loss, which refers to the error or cost of the neural network model on the validation data, takes its minimum value and remains relatively stable in even bigger iterations numbers. The number of hidden units used in the model was chosen to be between the size of the input layer (6) and the output layer (1), resulting in the following set of values:

$$h = [1, 2, 3, 4]$$

Furthermore, we defined our baseline model as a model that simply produces the mean of the target values of the training set as estimation. Subsequently, we implemented 2 level cross validation to evaluate and compare our models. The inner fold was used to determine the optimal hyperparameters of the models, namely the λ and hidden units number, whereas in the outer fold these two optimal values tuned in the inner fold, were used for the training of the linear regression model and the ANN in order to compute the generalization error. Additionally, the generalization error was computed by the mean squared loss.

2.2.2 Results

Outer fold	ANN		Linear Regression		Baseline
i	hi	E_i^{Test}	λ_i	E_i^{Test}	E_i^{Test}
1	4	26.77	1.00	0.20	14.42
2	4	4.80	1.00	0.88	13.33
3	3	54.53	0.10	0.64	14.4
4	2	7.02	0.01	0.68	14.54
5	3	2.44	1.00	0.61	14.49
6	3	17.96	1.00	0.40	14.39
7	1	7.73	0.01	5.64	9.89
8	3	11.65	1.00	1.57	14.49
9	2	48.27	1.00	8.29	9.73
10	2	22.27	1.00	3.80	13.20
Average		20.34		2.27	13.30

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

From the table provided above, it can be inferred that both the baseline model and the ANN have a higher generalization error and therefore most probably worse average performance compared to the linear regression model. Another worth mentioning observation is that in 4 out of the 10 outer folds, the 3 hidden units seem to be giving the optimal results. Lastly, having investigated the linear regression model's parameters, it is apparent from the table that the optimal regularization parameter is $\lambda = 1$, appearing in 5 out of 10 folds. This result comes in accordance with the result derived in the previous section, as the optimal value found here ($\lambda = 1$) is included in the optimal range discovered previously ($10^{-}3 < \lambda < 10^{1}$).

The bigger mean squared error (MSE) obtained for each outer fold when applying 2-level cross-validation on the artificial neural network (ANN) compared to the linear regression model may be attributed to the fact that ANN models are more complex and have a larger number of parameters compared to linear regression models, which makes them more prone to overfitting.

More specifically, when performing 2-level cross-validation, we divide the dataset into training and test sets for each outer fold, and then further divide the training set into training and validation sets for each inner fold. If the ANN is overfitting to the training set during the inner loop, then it may select a model with too many parameters, which will perform poorly on the validation set. This will result in a higher error rate on the test set during the outer loop. In contrast, a linear regression model is less complex and has fewer parameters compared to an ANN, which makes it less prone to overfitting. Therefore, it may perform better in situations where the dataset is small, as the one investigated in this report. However, it may not capture complex relationships between the features as well as an ANN would, especially in situations where the relationship between the input and output variables is non-linear.

2.2.3 Performance evaluation

This section is focused on the statistical evaluation of the difference in the performance among the three models investigated in the part of the report. For this purpose, the Setup I was used, as described in the Box 11.3.4 in the course's book [1]. This process was followed in order for us to draw clear conclusions related to the performance of each model in the the Breast Cancer Coimbra Dataset, which could be used in future analysis.
10-fold cross validation was applied to determine the generalization error for each model. For ANN, the hidden units number was set to $h = 3$, since based on the information provided in the table 1, i is the optimal value in 4 out of the 10 outer folds, which suggests that it is a good value to use for the model. As far as the linear regression model is concerned, the value $\lambda = 1$ was kept, given that it appears to be the most commonly occurring optimal value across the 10 outer folds. Finally, the confidence parameter was set to $\alpha = 0.05$.

Null-Hypothesis (H0)	Estimated difference ($\hat{z}$)	CI Lower	CI Upper	*p-value*	Conclusion
$E_{LN}^{Gen} = E_{BL}^{Gen}$	-5	-10,53	-0,36	0,04	Rejected
$E_{BL}^{Gen} = E_{ANN}^{Gen}$	-3,54	-6,15	-0,94	0,02	Rejected
$E_{ANN}^{Gen} = E_{LN}^{Gen}$	-4,36	-18,25	9,52	0,4	Not Rejected

Table 4: Setup I Statistical Test Results: *LR* stands for Linear Regression, *ANN* for Artificial Neural Network and *BL* for the Baseline Model

For the comparison between the Baseline Model and Linear Regression Model, the estimated difference was -5, meaning that on average, the Linear Regression Model's predictions were 5 units closer to the true values than the Baseline Model's predictions. The 95% confidence interval was [-10.53, -0.36], suggesting that we can be 95% confident that the true difference in performance between the two models falls within this range. The p-value was 0.04, validating the statistically important difference between the models.
The second comparison is between the baseline model and ANN. The estimated difference was found -3.54, meaning that on average, the ANN Model's predictions were 3.54 units closer to the true values than the baseline Model's predictions. The 95$ confidence interval was [-6.15, -0.94], suggesting that the true difference in performance between the two models falls within this range, with 95% certainty. The p-value was 0.02, indicating that the difference between the models is statistically significant.
The last comparison refers to the ANN and linear regression models. In this case, the estimated difference was -4.36, indicating that the Linear Regression Model's predictions were 4.36 units closer to the true values than the ANN Model's predictions, on average. The 95% confidence interval was [-18.25, 9.52], suggesting that we can be 95% confident that the true difference in performance between the two models falls within this range. The p-value was 0.4, proving that the models do not differ statistically significantly.
In summary, the statistical tests show that the LR and ANN models outperformed the Baseline model significantly, but there was no significant difference in performance between the LR and ANN models.

3 Classification

In this section we are going to address the main binary classification problem, the prediction of breast cancer choosing the possible biomarkers. We are going to employ three different methods to do that, logistic regression (LR), decision trees (DT) and a baseline. Finally, we will utilize two level cross validation to evaluate their performances and we will compare them using statistical testing.

3.1 Data pre-process and feature selection

Data were normalised by subtracting the mean and divided by the standard deviation, as described in the previous report. Additionally feature selection was an essential step because it helped to improve model performance by selecting the most relevant and with predictive value features from the dataset. Without

feature selection, the model may had become overly complex and struggled to identify patterns or make accurate predictions.

Attribute selection was done for each model separately. For the LR we used L2 regularization (Ridge), also described in 2.1.2. It performed feature selection during training by forcing some of the coefficients of the less important features to be small. The parameter C (C=$1/\lambda$) controls the strength of the regularization, with smaller values of C resulting in more regularization and more coefficients being set close to zero.

For the DT classifier we calculated the Gini importance score (Gini coefficient) for each feature in the dataset. The Gini importance score reflects the relative contribution of each feature to the overall impurity reduction of the DT during the training process. The selected features for a given alpha value (control complexity parameter) were determined based on whether the feature importance was greater than 0.1 for all folds. If the feature importance was not greater than 0.1 for all folds, the feature was not selected for that alpha value. This ensured that only features that consistently demonstrated high importance across all folds were selected, which can potentially lead to better generalization performance of the model.

3.2 Logistic Regression Classifier

The first ML model we used for our classification task was LR which has λ as its regularization parameter. After several trials with 5 fold cross validation with different λ intervals we found that the range of λ = [0.1, 20] gave the optimal results. Furthermore, we used L2 regularization (Ridge) as penalty term as previously described.

Figure 2 shows the accuracy of the LR model on the training and test sets for different values of the regularization parameter λ. As can be seen, increasing λ leads to a decrease in training accuracy, but an increase in test accuracy, indicating that a moderate level of regularization helps to prevent overfitting. The optimal value of λ, chosen based on the test set performance, was 4.

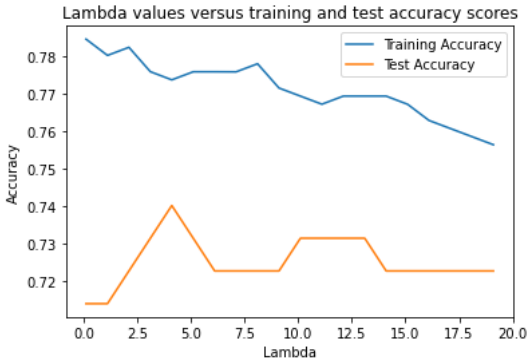

Figure 2: Performance of the Logistic Regression model on the training and test sets for different values of λ. The blue line represents the accuracy on the training set, and the orange line represents the accuracy on the test set.

We evaluated the performance of the LR model with the best λ using a confusion matrix as shown in Figure 3. As can be seen in the plot, the model achieved a high accuracy of 74% and a precision of 75% and recall of 67% for the cancer class.

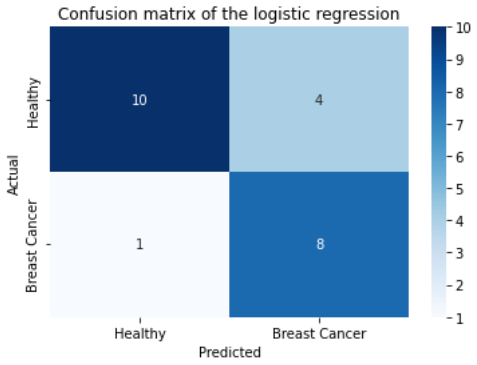

Figure 3: Confusion matrix for the logistic regression model trained using 5 fold cross validation. The matrix shows the number of true positives, false positives, false negatives, and true negatives for the test set predictions. Overall, the model achieved an accuracy of 74% on the test set.

Examining now the attribute coefficients of the LR model with the best lambda we generated Figure 4. The coefficients represent the contribution of each attribute to the prediction of the class. The size of the bars indicate the magnitude and direction of the coefficients, respectively. Resistin, HOMA, glucose and BMI therefore seem to have strong predictive values and may are promising biomarkers.

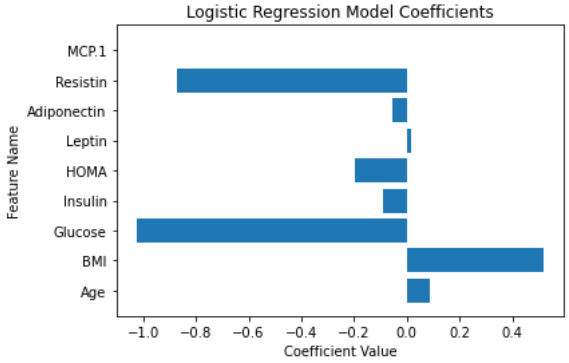

Figure 4: Attribute coefficients of the logistic regression model.

3.3 Decision Tree Classifier

The second ML model we used for our classification task was DT which has α as its complexity controlling parameter. After several trials with 5 fold cross validatation with different α intervals we found that the range of $\alpha = [0, 1]$ gave the optimal results. Furthermore feature selection was performed using Gini coefficient as describes in 3.1.

The optimal value of the alpha parameter was selected based on the highest accuracy on the test set, as shown in Figure 6.

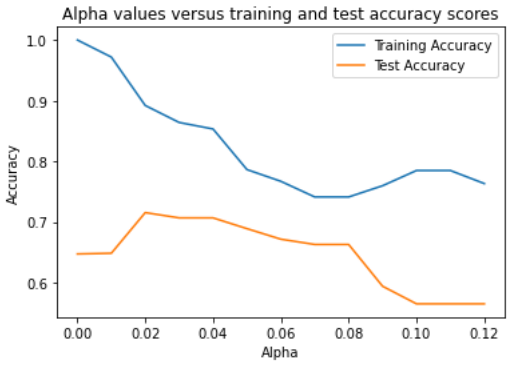

Figure 5: Performance of the Logistic Regression model on the training and test sets for different values of λ. The blue line represents the accuracy on the training set, and the orange line represents the accuracy on the test set.

The alpha that maximizes the test accuracy at 72% is equal to 0.02 and was selected. The confusion matrix below at Figure 7 displays the number of true positives, false negatives, true negatives, and false negatives for each class in the test set.

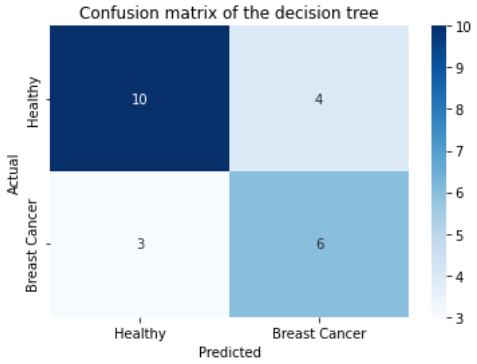

Figure 6: Confusion matrix plot for the Decision Tree model with optimal alpha value. The plot shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class in the test set.

The Decision Tree Classifier with the selected alpha value= 0.02 is shown in Figure 8.

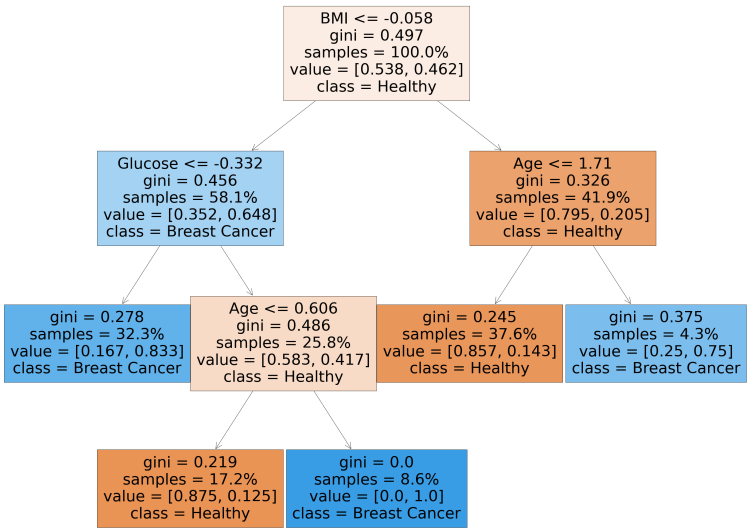

Figure 7: Decision tree visualization with the selected alpha=0.02 for classification of breast cancer. The leaves indicate the predicted class and the number of samples that end up in each class. The classification accuracy on the test set is 72%.

3.4 Baseline

The third model was a trivial baseline model which makes its prediction according to the majority class of the training data and does not have any complexity controlling parameter. The cross-validation accuracies of the baseline model were evaluated using 5-fold outer cross-validation. The highest test accuracy is 56% as shown in Figure 9.

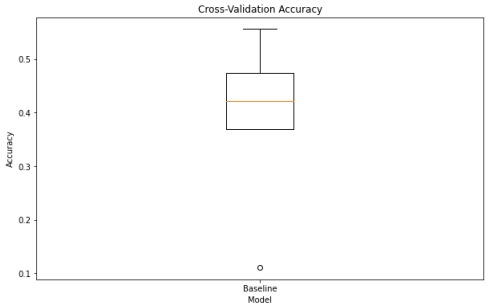

Figure 8: Cross-validation accuracies of the baseline model. The plot shows the average cross-validation accuracies of the baseline model over 5 outer cross-validation splits. The error bars represent the standard deviation of the mean. The baseline model simply predicts the most frequent class in the training set.

3.5 Evaluation of the models using two-level cross-validation

For the evaluation of the models a two-level Cross Validations is performed with 10 folds in both outer and inner loop. In the inner loop the optimal values of complexity controlling parameter λ and α are being calculated. Whereas, in the outer loop the these optimal values are utilized for the training of the Logistic Regression and Decision Tree models in order to calculate the generalization error of their performance. The generalization error is given by the formula below:

$$E_{gen} = \sum_{i=1}^{10} \frac{|D_{i,test}|}{N} E_{i,test}$$

In table 5 below there are the results of this evaluation. What can be firstly observed is that the Baseline model performs worse than the other two models in terms of average E_i^{Test} having an average value of 0.48. Regarding the Logistic Regression and the Decision Tree Classifier we see that both models achieved about the same E_i^{Test}.While the average alpha value for all the folds is the same as previously predicted 0.02, we notice a slight increase in lambda being equal to 0.5 instead of 0.4 that was previously computed. However, it is relatively close and it may be due the fact that in 3.2 we used accuracy as a metric to measure performance, while now we used the generalization error.

Outer Fold	Logistic Regression		Decision Tree Classifier		Baseline
i	λ_i^*	E_i^{Test}	α^*	E_i^{Test}	E_i^{Test}
1	2	0.17	0.02	0.08	0.75
2	1E-10	0.25	0.01	0.25	0.25
3	1E-10	0.42	0.03	0.33	0.50
4	1	0.42	0	0.17	0.58
5	1E-10	0.25	0	0.42	0.50
6	1	0.42	0.02	0.42	0.58
7	1E-10	0.27	0.02	0.27	0.27
8	1E-10	0.36	0.03	0.27	0.27
9	1	0.18	0.02	0.45	0.45
10	1	0.18	0.02	0.18	0.27
Average	0.5	0.26	0.02	0.27	0.48

Table 5: Two level cross-validation table used to compare the three models

3.6 Statistical evaluation using McNemar's test

In this part we will try to determine if there is a significant difference in the performance of the models when evaluated on the same test set by using the Setup I as in 2.2.3.

As can be seen in Table 6 in the comparison between the Decision Tree and the Baseline Model we can see that the estimated difference in generalization error is z = -0.1818 which means that on average the Decision Tree model estimations are by 0.1818 units closer to the true values compared to baseline model estimations. Since the interval includes zero, it indicates that there is a possibility that the true difference between the generalization errors of the two models could be zero. However, since the p-value is less than the significance level (α=0.05), we reject the null hypothesis that there is no difference in the generalization errors of the two models.

Regarding the Logistic Regression and the Baseline Model now we can also assume that Logistic Regression outperforms the Baseline.

Lastly, comparing the Decision Tree and the Logistic Regression we can not reject the Null-Hypothesis since the p-value was highter than the significance level, meaning that these two models do not have a significant difference in performance when predicting breast cancer.

All in all we can see that both the Decision Tree and the Logistic Regression outperform the Baseline, but again we can not be sure if one of them is best.

Null-Hypothesis (H_0)	Estimated Difference($\hat{z}$)	CI Lower	CI Upper	*p-value*	Conclusion
$E_{DT}^{Gen} = E_{Bl}^{Gen}$	-0.1818	-2.5456	2.1820	0.0009	Rejected
$E_{LR}^{Gen} = E_{BL}^{Gen}$	-0.3636	-2.5748	1.8475	0.0055	Rejected
$E_{LG}^{Gen} = E_{DT}^{Gen}$	-0.1818	-2.2289	1.8653	0.1306	Not Rejected

Table 6: Setup I Statistical Test Results (DT stands for Decision Tree, LR for Logistic Regression and Bl for the Baseline Model)

4 Discussion

Linear regression was executed using Ridge regularization(L2) where it was shown that Insulin, Leptin, Resistin and MCP.1 all played an important role in predicting HOMA as expected from project 1 when taking into consideration the correlation heatmap.

Subsequently, 3 ML models for the above mentioned regression task were constructed and the statistical evaluation performed on them, revealed that the regularized linear regression outperformed the Artificial Neural Network and both of them outperformed the baseline model.

Regarding the classification part we tested 3 different machine learning models, LR Classifier, DT Classifier and a Baseline, to predict breast cancer based on specific biomarkers. Both LR Classifier and DT Classifier had the same performance and outperformed the Baseline.

Previous study that exploited the same data set to also predict breast cancer, proposed as well logistic regression model using resistin, glycose, age and BMI and achieving sensitivity and specificity values of 82-88% and 84-90%, respectively.[2]

Our feature selection for the LR Classifier agrees with that study proposing resistin, glycose, age, BMI and additionally HOMA as possible biomarkers of breast cancer. However, it is important to note that further validation and testing of the models with larger and more diverse datasets are needed before they can be implemented in clinical settings.

5 Exams questions

5.1 Question 1 : Option D

A clear separation between the black circles and the red crosses indicates that the model is performing well in distinguishing between the two classes. Taking into account that the ROC curve is close to the diagonal line we can assume that its performance is not good.

5.2 Question 2 : Option A

We know that the impurty gain is given by

$$\Delta = I(r) - \sum_{k=1}^{K} \frac{N(v_k)}{N(r)} \, I(v_k) \quad (1)$$

We wish to determine the impurity gain of the split x7 = 2 using the classification error impurity measure. The impurity of a node is given by:

$$I(r) = 1 - \max_k p_{r,k}$$

where $p_{r,k}$ is the proportion of observations in class k at node r. Therefore, the impurity of the initial node can be computed as follows:

$$I(r) = 1 - \max_k \frac{N_k}{N} = 1 - \frac{30}{96} \approx 0.6875$$

where N_k is the number of observations in class k and N is the total number of observations.

The impurity of each child node can be similarly computed using the counts of observations in each class. For the left child node (x7 = 2), we have:

$$I(v_1) = 1 - \max_k \frac{N_{k,1}}{N_1} = 1 - \frac{33}{33 + 28 + 30 + 29} \approx 0.3968$$

where $N_{k,1}$ is the number of observations in class k at the left child node and N_1 is the total number of observations at the left child node.

For the right child node ($x7 \neq 2$), we have:

$$I(v_2) = 1 - \max_k \frac{N_{k,2}}{N_2} = 1 - \frac{4+2+1+3+5}{33+4+28+2+1+30+3+5+29} \approx 0.6719$$

where $N_{k,2}$ is the number of observations in class k at the right child node and N_2 is the total number of observations at the right child node. Finally, we can compute the impurity gain of the split as follows:

$$Delta = I(r) - \frac{N_1}{N} I(v_1) - \frac{N_2}{N} I(v_2) \approx 0.0195$$

Therefore, the impurity gain of the split x7 = 2 is approximately 0.0195.

5.3 Question 3 : Option A

The question refers to an ANN having a single hidden layer with 10 units, and using the softmax activation function for multi-class classification and the sigmoid non-linearity for the hidden layer.

For a single hidden layer neural network, as the one in the exercise, including a softmax activation function,s with n_h units, input dimension M and output dimension C, the number of trainable parameters is determined by:

$$N = (M+1)n_h + (n_h+1) * C$$

where the +1 corresponds to the bias term for each layer. The given ANN, includes 7 input features (x_1 to x_7) and 4 output classes (y=1,2,3,4). Therefore, the input dimension M is 7 and the output dimension C is 4. The hidden layer has 10 units. Thus, the number of trainable parameters is:

$$N = (7+1) * 10 + (10+1) * 4 = 124$$

5.4 Question 4 : Option D

The correct assignment of rules to the nodes in the decision tree based on the predicted label assignment in Figure 4 is:

- A: In order to split the labels in the groups (1,2) and (1,3,4), we should set: $b_1 >= -0.76$. If this is true, then we will end up in Congestion 3,1 or 4 as it can be seen on the tree.
- B: From node B, to end up to Congestion level 2, we should set $b_2 >= 0.03$
- C: From C node, to end up to level 4 Congestion, $b_1 >= -0.16$
- D: From D, $b_2 >= 0.01$ would give congestion level 1.

Therefore, the correct answer is D.

5.5 Question 5 : Option C

First of we have five folds for each of the models in the outer level for a total of ten. Moreover, there are four folds for each of the five outer folds making this a total of twenty. Also there are five values of lamda and hidden units which are used for the models, and so the total models trained and tested are:

$$20 * 100 = 200$$

Therefore, the total number of trained and tested models is 210.
Now, we know that the total time to train and test one neural network is 20+5=25 ms, thus by dividing in half the total models we can compute the total time for the ANN:
105∗25= 2625 ms
Same applies for the regression: 105∗(8+1)=945 ms
Finally we get :

Total time = 3570 ms

5.6 Question 6 : Option B

First we compute the product of each y with the weights given.

- For y= [1, -1.4, 2.6] we get: 12.46, 11.03, 8.89
- For y= [1, -0.6, -1.6] we get : -2.66, -2.42, -1.56
- For y= [1, 2.1, 5.0] we get : 12.79, 12.13, 9.99
- For y= [1, 0.7, 3.8] we get : 11.8, 11.3 and 8.89

Now we calculate the probability for each result using :

$$P(y = 4) = \frac{1}{1 + e^{y_1 * w_1} + e^{y_2 * w_2} + e^{y_3 * w_3}}$$

and we get respectively the values : 0.000003025, 0.73045, 0.000001767, 0.000004656
From these values we can conclude that B is the correct answer since it has the highest probability.

List of Figures

List of Tables

References

[1] T. H. Mikkel Schmidt and M. Morup, *Introduction to Machine Learning and Data Mining*. 1 ed., 2022.

[2] M. Patrício, J. Pereira, J. Crisóstomo, M. R. Alves, B. Carvalho, A. Oliveira, A. Santos, L. Raposo, C. Lopes, F. Caramelo, *et al.*, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, no. 1, p. 29, 2018.