Danmarks Tekniske Universitet

Introduction to Machine Learning and Data Mining

Georgios Arvanitidis

Project 2: Supervised learning: Classification and Regression

---

Alexandre del Barco (s232518)
Jakub Oszczak (s233577)

November 26, 2023

Contents

List of Figures

List of Tables

Student ID	Regression	Classification	Discussion	Exam Questions
s232518	80%	20%	50%	50%
s233577	20%	80%	50%	50%

1 Regression

In this first section a regression model will be applied to help describe, understand and predict unknown values of the dataset selected. Later on, the the results using different predicting models will be statistically evaluated.

By using a regression model is expected to predict the refractive index based on other attributes. As discussed in the previous project, a correlation between the value in hte composition of the glasses of silicon (Si) and calcium (Ca) were related to their value of refractive index (RI). Therefore, initially, these two attributes (Ca and Si) will be used to predict the values of refractive index (RI), and more attributes will be tested to help the RI prediction and check if a better result is got. The first thing is done to the data is to apply a feature transformation to it, this consist of normalising all features in it to achieve a value equal to 0 for their mean and equal to 1 for the standard deviation. Looking at the results at Table 1, it can be seen that can not be guaranteed that the values for the mean and the standard deviation are exactly 0 and 1 due to numerical precision, but they are close to them.

Column name	**Mean**	**Standard Deviation**
RI	-2.877e-14	1.002
Na	2.179e-15	1.002
Mg	-2.801e-16	1.002
Al	-3.434e-16	1.002
Si	9.966e-16	1.002
K	7.470e-17	1.002
Ca	-3.137e-16	1.002
Ba	-1.763e-16	1.002
Fe	-6.121e-17	1.002

Table 1: Summary mean and standard deviation after feature transformation

1.1 Part A: Regression problem and statistical evaluation

1.1.1 Best attribute to be predicted

As it was discoursed in the first assignment, it was intended to predict the value of the refractive index (RI) of glasses based on the vales of their composition of calcium (Ca) and silicone (Si), as a it was found a strong positive correlation between the amount of calcium oxide in the glass and the value of its refractive index, whereas it was also found a medium strong negative correlation between the share of silicone in the glass its value of refractive index.

As it is know that the attributes RI, Na and Ca are correlated, theoretically, any of this suitable to be predicted by the others. However, it is desired to predict the attribute that shows a better response to a linear regression, as it will be the one with best results so the more suitable. Thus, a OLS linear regression model is applied to predict each of Si, Ca and RI variables based on the other two attributes, and the average MSE of the residuals is computed to check which attribute is more suitable to regression analysis.

Attribute	RI	Na	CA
Average MSE	2.64	-9.13	-7.83

Table 2: Average MSE od residuals for RI, NA, and CA

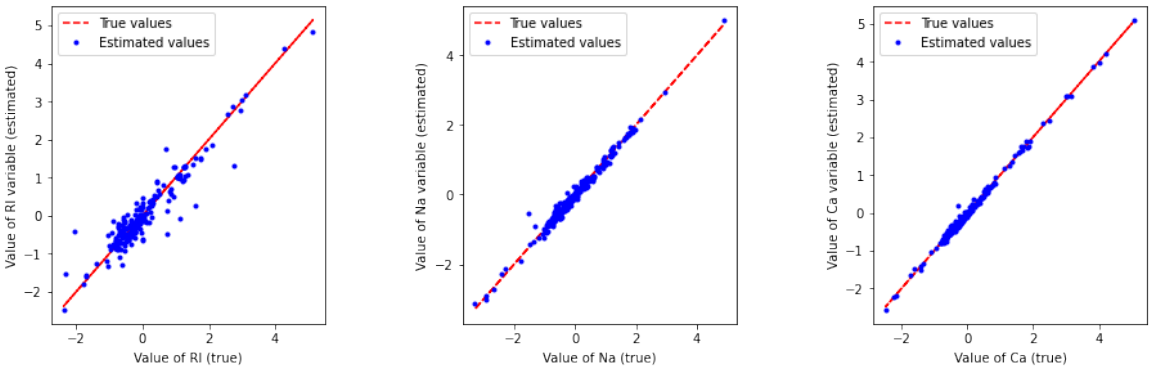

Figure 1: OLS regressions for RI, Na and CA

As can be seen in Figure 3, all three possible estimation attributes seem to align well with the true fit. Nevertheless, RI is the one that shows the lowest value of MSE residual, as can be seen in Table 2. Therefore, is decided to stick with the idea of predicting RI based on the values og the Ca and Na attributes.

1.1.2 Linear Regression to predict RI

Now the decision do be figured out is if a better result is got by using only the the Si and Ca attributes or a better fit is gotten using a combination of all attributes. To answer this question a both linear regression to predict RI based on on all other attributes and linear regression based on CA ans Si are computed.

LR based on Ca and Si

When performing the linear regression based to predict values of RI based on the Ca and Si attributes is calculated, the results are quite promising, as the behaviour of plotting the results comparing the estimated value of RI and the true value of RI (Left plot in Figure 2) intend to follow a straight line with slope equal to one. Moreover, the residuals (difference between each real value with its prediction) follow a normal distribution with its highest values gather all together at the zero value, so the bast majority of the residual values all close to zero (Right plot Figure 3). The final equation got is Equation 1, where $Si = X_1$ and $Ca = X_2$.

$$y = -0.3899X_1 + 0.729X_2 \tag{1}$$

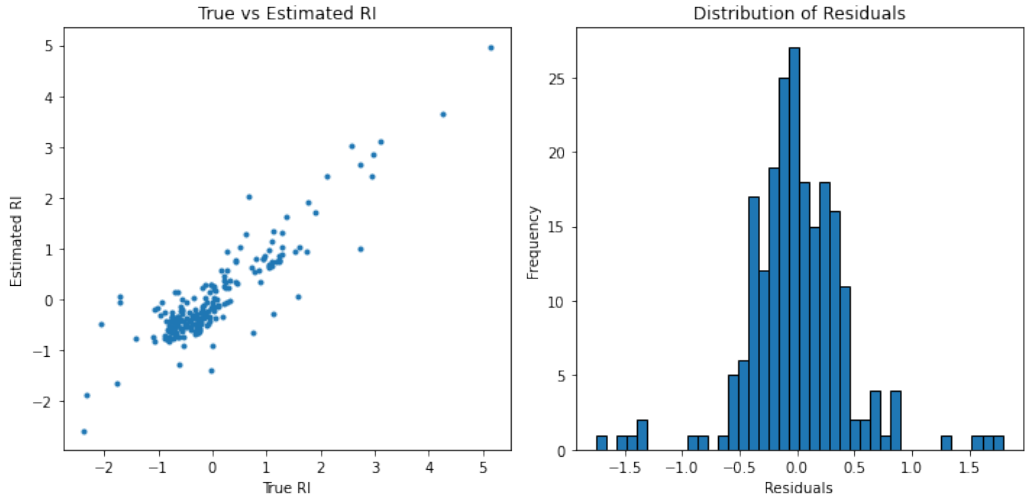

Figure 2: OLS regressions for RI, Na and CA. Left: True vale vs Estimated. Right: Distribution of residuals

LR based on all attributes

Then the same procedure as before is repeated, but in this time the values of RI are predicted using all the data available for all attributes. The result are again really accurate as as can be in Figure 3. The final result is even improved, plotting the true values of RI vs their estimated values is seen that the sahep of it is even more gather to the x=y line, which ensures better result. On the other hand, in the left plot of Figure 3 is observed that the residuals are more concentrated around the 0 value of them, which give us again the same information: the regression performs better if using more attributes than just Ca and Si to predict RI values. The final equation is Equation 2.

$$y = 0.043X_{Si} + 1.461X_{Ca} + 0.375X_{Na} + 0.876X_{Mg} + 0.005X_{Al} + 0.297X_{K} + 0.488X_{Ba} + 0.014X_{Fe} \quad (2)$$

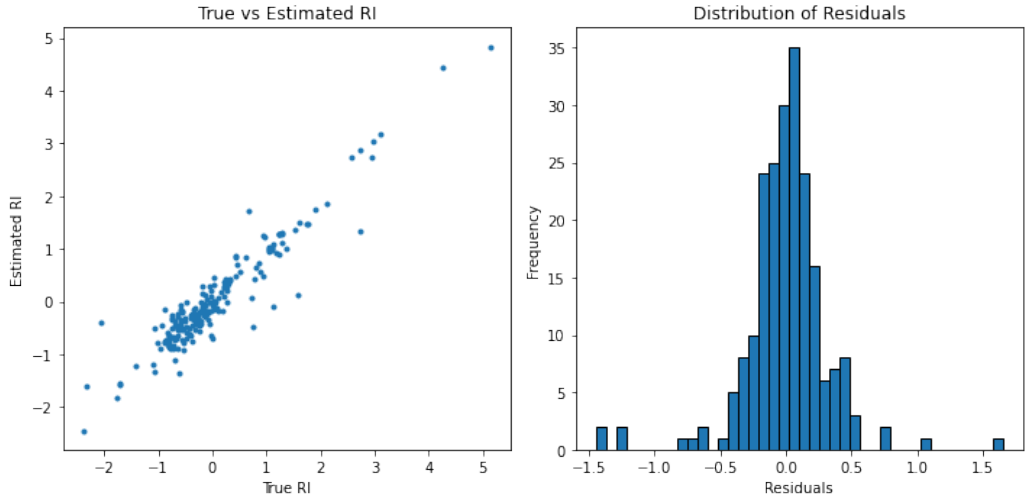

Figure 3: OLS regressions for RI, using all attributes. Left: True vale vs Estimated. Right: Distribution of residuals

As the results obtained using this method are more accurate, this one is the one chosen.

1.1.3 Regularisation

Now a regularisation parameter(λ) is introduced in the linear regression to try to minimise the overall sum of square errors (SSE) and also try to significantly reduce the variance in the data set then affording overfitting without substantially increasing the bias. The regularisation parameter will be set to different values: from $\lambda = 10^{-4}$ to $\lambda = 10^{8}$. Then to properly estimate the generalisation error, a two level cross-validation will be performed with $K_1 = K_2 = 10$ folds in both layers. This means that the generalisation error is first calculated by taking the average test error given by each possible value of the regularisation error within the folds, thus the λ with less error is chosen.

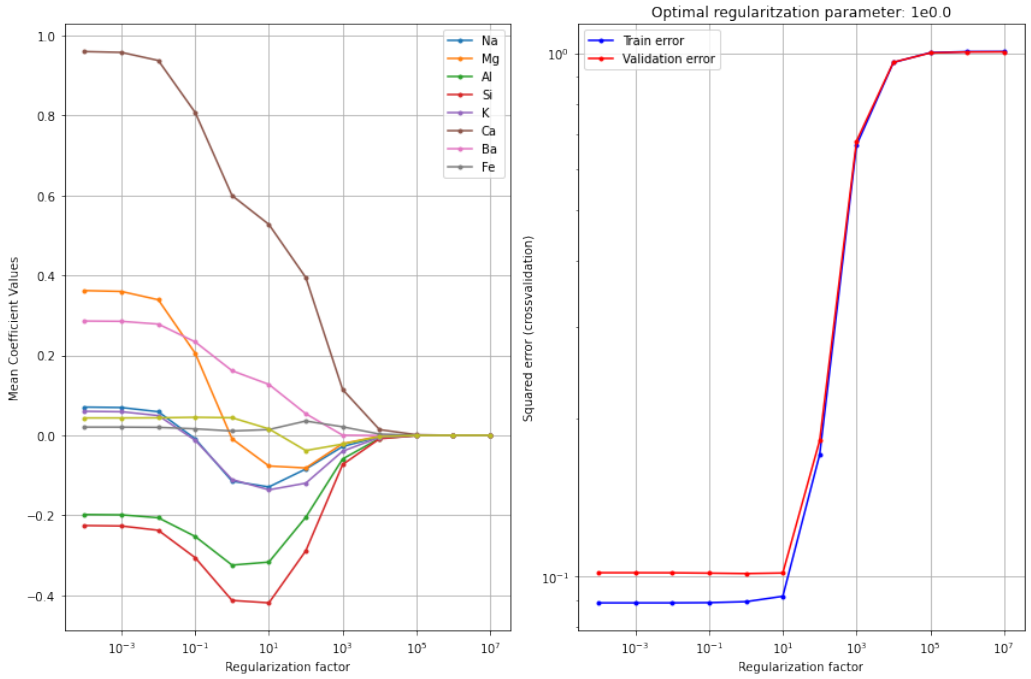

Figure 4: Left: Mean linear model coefficient as a function of regularisation parameter. Right: squared error as a function of regularisation error

By watching the results plotted in Figure 4 Right, it can be seen the regular trend for the squared error of going down and eventually increase λ increases. The optimal value of regularisation parameter to minimise the most the squared error is $\lambda = 1$. This means that setting λ to this parameter it is reached the best trade-off between bias and variance. If a higher value is set, the bias'd be higher and the variance smaller; and the other way around it the λ were smaller.

On the other hand, in Figure 4 Right, can be seen some different behaviours depending on the the attribute, some decrease to eventually stabilise and converge to 0 as λ increase, some decrease to then increase to then stabilise and converge to 0. The weight at the last fold, then final equation are represented in Equation 3.

$$y = -0.11x_{Na} - 0.32x_{Al} - 0.41x_{Si} - 0.11x_K + 0.61x_{Ca} + 0.16x_{Ba} + 0.01x_{Fe} \tag{3}$$

Then it is possible to conclude that if a glass have high values of Ca and Ba and low values of Si, K and Na it has a high change to have high value of RI.

1.2 Part B: Comparison and statistical evaluation of 3 prediction models

In this section, it will be compared the performance of three different prediction models:

- Regularised Linear Regression
- Artificial Neural Network (ANN)
- Baseline

1.2.1 Comparison of 3 models

The Regularised Linear Regression will be the one done in the previous section, and the ANN and BL will be computed. Again, to estimate the generalisation error, it will be used a two-level cross-validation with $K = 1 = 2 = 10$ for each model. In this regard, the outer fold is used to test the model performance, whereas

the inner fold is used to compute a complexity control parameter design to minimise the generalisation error. The ranges for λ are the same as in last case: from $\lambda = 10^{-4}$ to $\lambda = 10^{8}$.

For the ANN, the networks are trained using a linear hidden layer and an output layer with an hyperbolic tangent activation function in the hidden layer. The complexity controlling parameter is set form 1 to 3 throughout cross-validation.

Outer fold	ANN		Linear regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	1.0	3.75	0.01	0.09	1.95
2	1.0	0.79	1e-05	0.03	0.24
3	1.0	2.43	0.01	0.19	0.98
4	1.0	1.31	0.1	0.10	1.35
5	1.0	0.81	0.001	0.05	0.85
6	1.0	0.16	0.1	0.054	0.22
7	1.0	1.25	0.01	0.05	0.85
8	1.0	0.93	0.01	0.23	0.92
9	1.0	3.69	10.0	0.30	1.4
10	1.0	1.23	0.1	0.07	1.24

Table 3: Comparison final results of two-level cross-validation three models

The Table 3 gathers together the comparison of the results got for each model. The model with the lower error is the Linear Regression. Although the result got from ANN and Baseline are quite close to each other, Baseline is slightly better than ANN.

1.2.2 Statistical Evaluation

Finally, a statistical evaluation is run to evaluate if there is a significant performance difference between the three models used. It is chosen the use the **Setup 1**, which is characterised by the training set being fixed, also $\alpha = 0.05$.

	ANN vs RLR	**ANN vs Baseline**	**RLR vs Baseline**
CI	(0.86, 3.31)	(0.024, 2.39)	(-1.19, -0.56)
p value	0.0039	0.0462	0.0001

Table 4: Comparison final results of two-level cross-validation three models

The results shown in Table 4 state that Regularised linear regression differ notably comparing to the other two models, while ANN and Baseline ate aligned, as a high p-value is found in ANN vs Baseline but not in other cases. Among the confidence interval, the ones computed from ANN vs RLR and ANN vs Baseline overlap in a certain part, what means that there is not a big difference in their metrics.

2 Classification

2.1 Introduction

As for the classification task we have chosen to predict the type of glass based on the oxide content of different elements in the glass. The choice of classification task was obvious since glass type prediction is the primary purpose of our dataset. It is a multiclass classification problem since there are 7 possible glass types to be predicted.

2.2 Models comparison

We have decided to compare three models: multinomial logistic regression, K-Nearest Neighbours (KNN) and a baseline model. Multinomial logistic regression is a variance of logistic regression used for multiclass classification problems. It uses cross-entropy loss function instead of log loss as binomial regression does. The regularization parameter λ is used to prevent overcomplicating of the model. Due to the specifics of Scikit learn library for Python which we are using, λ is an inverse of regularization strength, so smaller values specify stronger regularization. The set of λ values checked is in the range from 0.0001 to 1.0 on a log scale:

$$\{0.0001, 0.001, 0.01, 0.1, 1.0, 10, 100, 1000\}.$$

Second model used is K-Nearest Neighbours, which assess the predicted class based on number (K) of nearest data points. In this model the complexity controlling parameter is "K" the number of nearest neighbours taken into account. To help avoid ties we have chosen to only consider odd numbers for K:

$$\{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21\}$$

Lastly the baseline model simply assigns any considered point to the biggest target class present in the training data. That means a baseline model does not care about dataset features, the majority class gets predicted every time.

Once again for the comparison of models the two-level cross-validation was performed in order to obtain the most accurate results. The number of folds are K1 = K2 = 10, where K1 is outer loop and K2 inner loop. Folds for both levels of cross validation were generated with the same seed (random state value) to provide the same evaluation conditions for every model. The data used was standardized such that each column has mean 0 and standard deviation 1.

Table 5: **The results from two-level cross-validation ($K_{outer} = K_{inner} = 10$); the optimal values of k_i^* (number of neighbours in KNN) and λ_i^* (regularization) along with the corresponding error (E_i^{test}).**

Outer fold		KNN	Multinomial regression		Baseline
i	k_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	0.348	1000	0.369	0.646
2	3	0.253	1	0.369	0.641
3	3	0.321	100	0.316	0.651
4	11	0.332	10	0.337	0.656
5	3	0.310	1000	0.362	0.647
6	11	0.331	1	0.355	0.636
7	3	0.341	1	0.387	0.636
8	3	0.320	10	0.393	0.647
9	3	0.316	10	0.368	0.621
10	3	0.289	10	0.352	0.658
Generalization Error		0.316		0.361	0.644

The optimal values of k_i^* turned out to be the same for almost all outer fold iterations. The weird behaviour is that the extreme low value of the range (3) is almost always chosen. For KNN algorithm it seems that the smaller K the better. It would suggest that the data is quite complex and needs a more complicated model. As for the Multinomial Logistic Regression λ_i^* is much more diverse. It does not stick to extreme values of the range. It seems that lower error occurs for mid range values like 1 and 10. Some weird and unexpected results may be due to the imbalanced classes. Class balancing techniques like SMOTE could be applied, but it was not in the course's scope and the baseline model wouldn't make much sense then.

2.3 Statistical evaluation

In order to gain a deeper understanding of the performance distinctions among the three examined models, we conducted a statistical assessment. In line with the project's guidelines, we utilized the outer validation splits to facilitate the statistical assessment. We employed Statistical Setup II, a correlated t-test, to conduct pairwise comparisons of the models [1]. This approach assesses two models by estimating the disparity in their generalization errors. Unlike the Setup I this takes into account the randomness of the training sets. The significance level (α) was established at 0.05. If the equation E_i - $E_j = 0$ is true, then performance of two models is the same. Therefore null hypothesis is:

$$H_0 : \text{Models } M_i \text{ and } M_j \text{ have equal performance.}$$

	KNN vs Multinomial Regression	KNN vs Baseline	Multinomial Regression vs Baseline
Confidence Interval (CI)	$[-0.080, -0.010]$	$[-0.358, -0.298]$	$[-0.312, -0.255]$
p-value	0.017	$1.338e^{-9}$	$3.252e^{-9}$

Table 6: Results of statistical evaluation.

As one can see from the above table every p-value is below threshold value (0.05), which means that results are statistically significant. It means that for every comparison null hypothesis got rejected. Therefore in every case two models differ in terms of performance. Obviously in the comparisons of KNN and Multinomial Regression with Baseline model the performance gap is much more significant than in the third comparison. However it is still clear that KNN did a bit better than Multinomial Regression.

2.4 Training a logistic regression model

While simple linear regression is about fitting a line to data points, in multiple linear regression it is fitting the plane to data points. In simple logistic regression it is about fitting a logistic function to data in a way that the error is minimized. Multinomial logistic regression expands the binary regression to handle more than two classes. Instead of having one set of coefficients for one class, we have multiple sets of coefficients, one for each class. The model calculates the probability of each class for a given observation. To convert these class probabilities into a valid probability distribution (probabilities summing to 1), we use the softmax function which takes the exponential of each class probability and normalizes them. Therefore ensuring that the sum of probabilities across all classes equals 1.

We trained a Multinomial Logistic Regression model with optimal parameter value $\lambda = 10$. The dataset was normalized (mean 0 and standard deviation 1) and split into training set (80%) and test set (20%). Standard "L2" penalty was used. We achieved following results:

- Accuracy: 0.7209
- Precision: 0.6910
- Recall: 0.7209
- F1 Score: 0.6933

The confusion matrix is presented below:

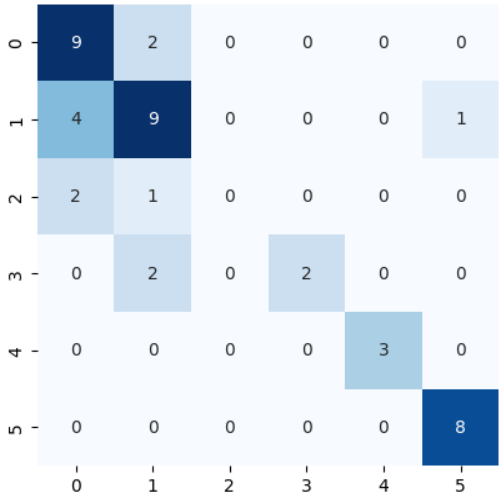

Figure 5: Confusion matrix for trained Multinomial Logistic Regression model

When it comes to feature importance here is the table of model's coefficients:

Class	0	1	2	3	4	5
RI	-0.0793	0.2406	-2.9255	-1.0863	0.0445	3.8059
Na	-0.3688	-1.4627	-1.3613	-1.4275	2.3146	2.3057
Mg	3.1917	-1.5723	1.2033	-1.7689	0.5118	-1.5656
Al	-2.1551	-1.0742	-2.1365	2.1085	-0.6804	2.2181
Si	0.7227	-1.0744	-2.3373	-0.6804	0.8038	2.5656
K	0.7020	0.0172	-1.7266	2.1663	-3.6560	2.4971
Ca	1.1710	-1.5965	1.2625	0.9138	0.5794	-2.3302
Ba	1.5435	-0.1369	-0.1165	-0.2919	-1.7246	0.7263
Fe	0.8906	1.1421	0.8040	0.5232	-2.3394	-1.0206

Table 7: Model's coefficients for every feature and class

In logistic regression, the importance of a feature is associated with the magnitude of its coefficient. Larger magnitude coefficients indicate a stronger influence of the corresponding feature on the prediction. Additionally, the sign of the coefficient (positive or negative) indicates the direction of the influence. Here are some observations:

- Class 0: Feature Mg (with a coefficient of 3.19) has the highest positive influence.
- Class 1: Feature Ca (with a coefficient of -1.60) has the highest negative influence.
- Class 2: Feature RI (with a coefficient of -2.93) has the highest negative influence.
- Class 3: Feature K (with a coefficient of 2.17) has the highest positive influence.
- Class 4: Feature K (with a coefficient of -3.66) has the highest negative influence.
- Class 5: Feature RI (with a coefficient of 3.81) has the highest positive influence.

3 Discussion

Writing a report on the comparison of regression and classification models has provided us with a deeper understanding of the importance of parameter tuning and model selection. The inclusion of a baseline model

underscored the necessity of benchmarking against simple, intuitive approaches. Performing a correlated t-test for statistical analysis improved my ability to rigorously assess and compare model performances. The examination of logistic regression coefficients offered a practical understanding of feature importance and how specific attributes influence classification outcomes. Overall, crafting this report has equipped us with a holistic view of the classification process, thoughtful model selection, robust evaluation methodologies, and the interpretability of model outcomes.

We found a scientific paper [2] of Polish researchers where they performed a classification task on the Glass Dataset we are using. They have also used a KNN classifier and got simmilar results. In this study the number of nearest neighbours K=3 was used, which aligns with our optimal value of K=3 obtained from two-level cross-validation process. The accuracy score of about 0.72 for classification task for KNN model is also very similar to what should have got since the performance of KNN and Multinomial Logistic Regression was very similar and for MLR we obtained accuracy score of 0.72.

4 Exam Problems

Question 1
Answer C - It is fairly simple, when you look at the formulas for FPR and TPR you can see that when you move the threshold of $\hat{y}$, passing past every red cross changes the TPR rate by 0.25 and passing past the black circle changes FPR by 0.25. Now just match the pattern of ROC curve with the ordering of red crosses and black circles.

Question 2
Answer C - We are considering classification error impurity measure which is represented by following equation: $1 - max\ p(c|v)$. To calculate impurity gain of split $x_7 = 2$ we do the following: $1 - (\frac{134}{135}) = 0,0074$. Where 135 is number of all samples and we have 134 after subtraction of one sample for $x_7 = 2$.

Question 3
Answer A - So, we have 7 features and an bias on the input so 8 nodes and then we have 10 nodes in hidden layer and then we have those 10 hidden nodes plus bias and 4 output nodes. Every layer is fully connected. It means every node has connection to every other node. That gives us 8*10 + 11*4 = 80 + 44 = 124 parameters.

Question 4
Answer D - to figure out which answer is the correct one we need to follow the tree for all answers. If we look at congestion 2 we can see that answers B and C are false because to get to cong.2 we need node B to be True, both those answers have B: b1>=-0,16 and b>=-0,76 which eliminates them. Use the same principal to eliminate other and get correct answer.

Question 5
Answer C - in order to obtain the table we need to do 4 trainings (K2 = 4), per every parameter value (5 times) and then we do it 5 times (K1 = 5) with 1 additional training for each outer fold. So (4*5*5*25ms + 1*5*25ms) + (4*5*5*9ms + 1*5*9ms) = 3570ms.

Question 6
Answer B - When we put values of b into $\hat{y}$ vector, transpose and calculate values for all 3 k's. Then we just plug those values into final softmax equation for k=4. We get a 0.730 which is 73% probability of class 4 for Answer B.

References

[1] Mikkel N. Schmidt Tue Herlau and Morten Mørup. *Introduction to Machine Learning and Data Mining.* Technical University of Denmark, 2023.

[2] K. et. al. Adamiak. "Object Classification Using Support Vector Machines with Kernel-based Data Preprocessing." In: *Image Processing and Communications* 21(3) (2016), pp. 45–53.