Machine Learning and Data Mining Report 2

Algerian Forest Fires

Student no.:	s192322	Student:	Paolo Dalpasso
Student no.:	s206063	Student:	Changzhi Ai
Student no.:	s192624	Student:	Boris Guillerey

Contents

1 Introduction

Project report 2 naturally follows project report 1 on "Algerian Forest Fires prediction". The FWI (Fire Weather Index) indicates the potential intensity of a wild fire and it's calculated integrating the BUI and ISI indexes, which are obtained from the fire weather observations and fuel moisture code, as shown in FIG. 1.1. Thus, we will choose FWI attribute as a prediction object in the whole project 2.

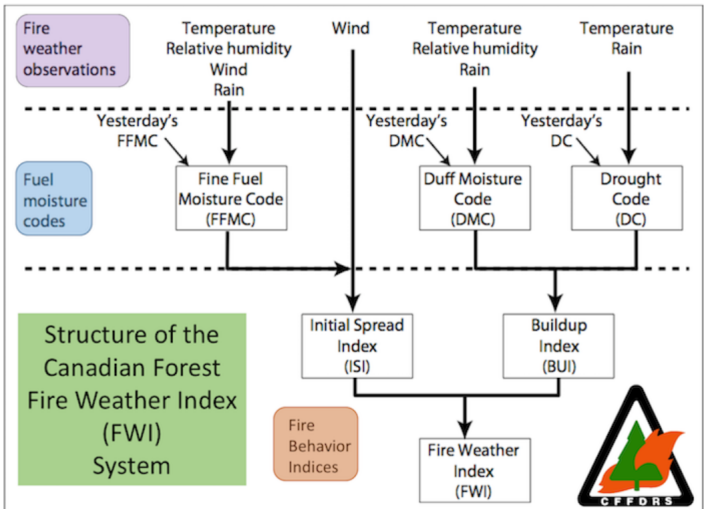

Figure 1.1: FWI calculation

2 Regression part A

2.1 Objective of the regression

The aim of our study is to predict the Fire Weather Index (FWI) for the 2 different regions Bejaia and Sidi-Bel-Abbes. Both regions were monitored for 4 months and meteorological data (Temperature, RH, Ws, Rain) were collected and used to determine other parameters from the Fire Weather Index system (FFMC, DMC, DC, ISI, BUI, FWI), which describes the probability for a wild fire to appear and its intensity. In the article, from which we took inspiration for the report, it's stated that the rain attribute was not considered to train the machine learning algorithm. On the contrary in this report, the FWI index will be first predicted only using meteorological data to test if these are enough to give an acceptable prediction of the FWI.

2.2 Regularization

The FWI is predicted using a linear model, in which a regularization factor λ is introduced to improve the model performance.

The cost function could be calculated and this objective can be solved by computing the derivative and setting it equal to zero. Then the optimal weights ω^* as a function of λ could be obtained as follows:

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

$$\omega^* = \frac{\hat{X}^T \hat{y}}{\hat{X}^T \hat{X} + \lambda I}$$

λ controls the complexity of model. Changing λ, variance/bias ratio of the model would be tuned. Therefore low λ, corresponds to models with high variance and low bias, while high λ, corresponds to models with low variance and high bias. The model is tested on a range of λ and the optimal one, which gives the lowest generalized error can be selected.

The plots of the training and test error of the models trained only with the 4 meteorological attributes and with all the attributes are shown in Fig.2.1.

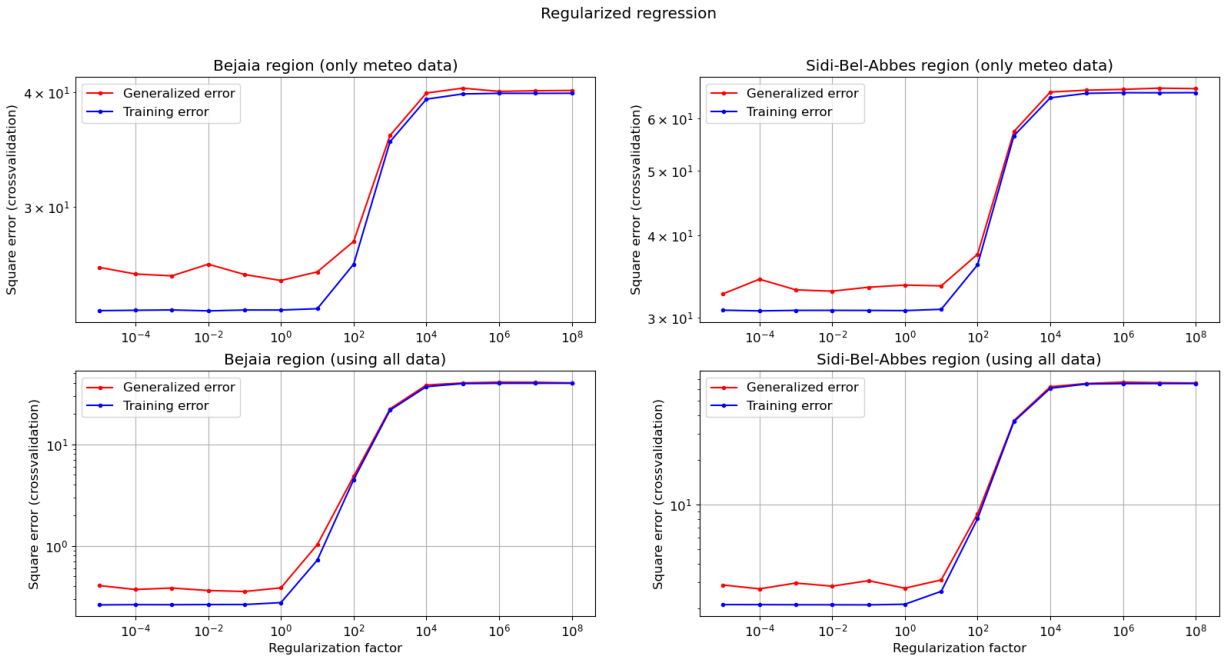

Figure 2.1: Training error and generalized error for regularized regression for both regions

As shown in 2.1 the generalization error is drastically lower when considering all the attributes and not only the meteorological data. For the Bejaia region the error is 3 times lower, while for the Sidi-Bel-Abbes it is 4 times lower.

Therefore, even if it was stated in the previous report, that the regression prediction would have been based only on the meteorological data, to obtain an acceptable prediction of FWI, it is clear that it is necessary to consider all the attributes of the dataset according to these results.

Furthermore, the optimal value of the regularization factor is not clearly identified from the graphs, although it can be said that it is located in all the cases in a range from 10^{-5} to 10. The double cross validation applied later on, will give a better assessment of this value. Finally, we also assess the effect of including quadratic terms in the observation matrix. The reduction of the generalization error is observed on Figure 2.2 and for this case an optimal value of $\lambda = 10$ could be determined.

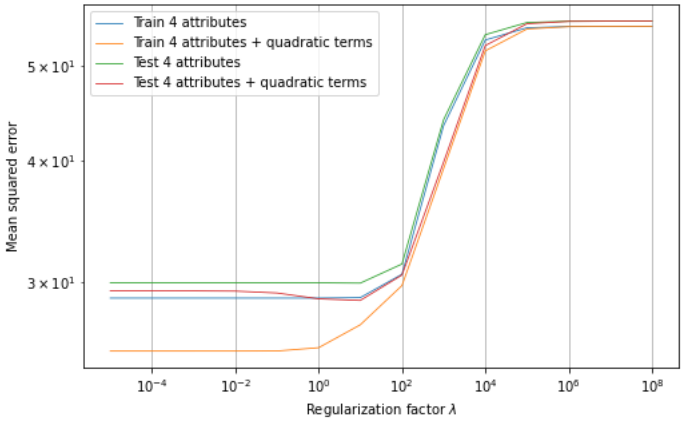

Figure 2.2: Squared error as a function of regularization factor lambda for both regions when applying first four attributes as data matrix X and effect of adding quadratic terms.

2.3 Effects of the attributes

The effect of the attributes in the linear regression can be assessed by representing the variation of their weights with the regularization factor. Figure 2.3 depicts this variation if considering only the meteorological data. In this case, it appears as expected that the temperature and wind speed affect positively the FWI while the rain and relative humidity affect it negatively. On the other hand, if considering the 9 attributes, the results is mostly influenced positively by the ISI, BUI indices, which makes perfectly sense when considering the construction of the FWI index as presented in Figure 1.1.

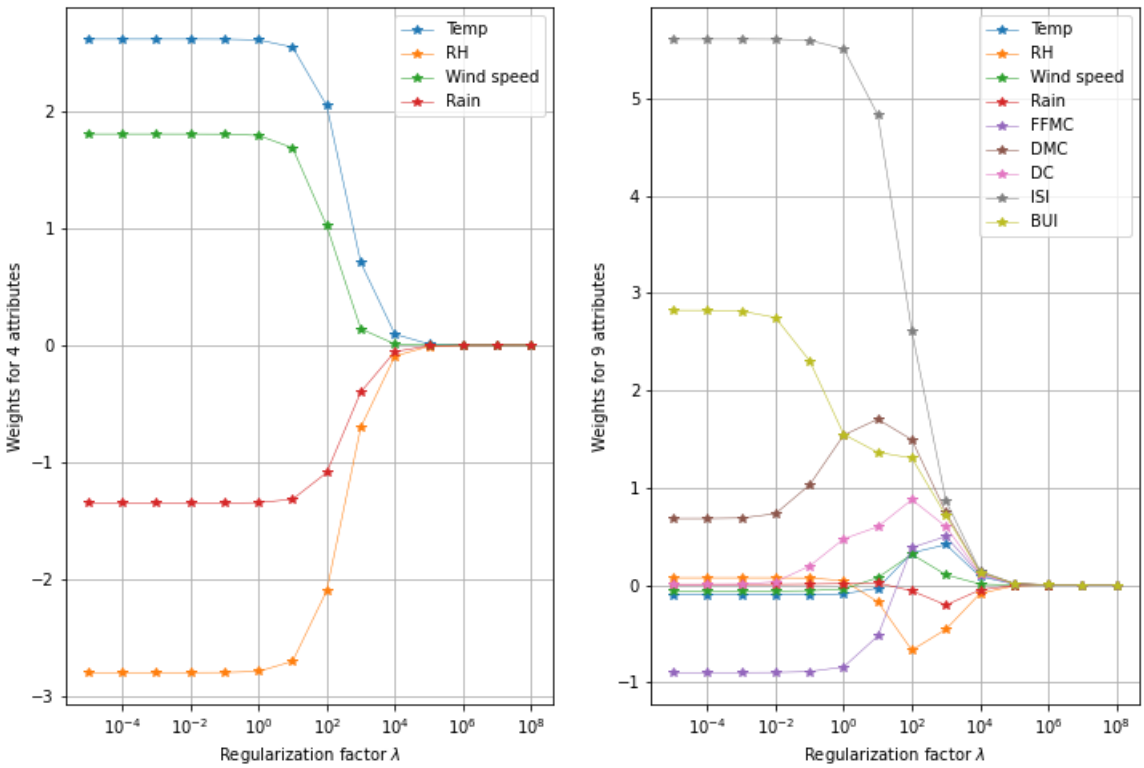

Figure 2.3: Mean coefficient values as a function of regularization factor lambda using data from both regions as inputs.

3 Regression part B

In this section, the FWI is predicted using 3 different models: a regularized linear regression, an Artificial Neural Network (ANN) and a baseline model, which is used as a comparative reference for the other 2 models. The baseline model consists in a linear regression model with no features, i.e. it computes the mean of y on the training data, and use this value as prediction for the test set. A two-layer cross validation (K1 = 10, K2 = 10) is applied to both linear regression and ANN, where each train set of the outer cross validation is splitted in another pair of train and test set, which are used to find the optimal complexity-controlling parameters for both models: respectively λ and the number of hidden units h, for the regularized linear regression and for the ANN.

After a few tests to determine first approximations of optimal values of λ and h, the following range were chosen for the inner loop in double cross validation:

$\lambda = [10^{-5}; 10^{9}], \ \textit{multiples of } 10$

$h = [1; 10]$

In the following part, the results are presented when training the models on both regions together, and by separating them. As we see in the previous part, it is indeed expected that the construction of the models differ for the two regions because of the variability of their climate and vegetation. Before applying the double cross validation with 10 inner folds and 10 outer folds, we also discarded one again using only the 4 meteorological attributes to determine FWI. As seen in Figure 3.1, realized for a simple 2 fold cross validation, the results obtained with only 4 attributes are not satisfactory, as the representation of the estimation versus the true values diverges strongly from the linear line $y = x$.

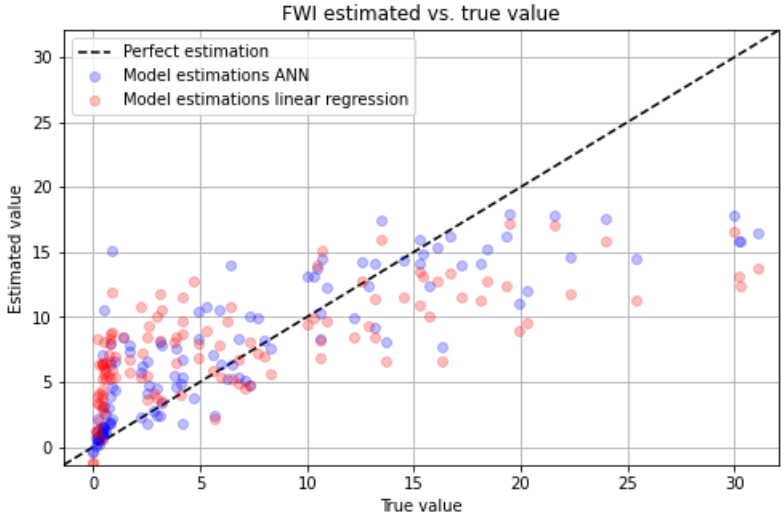

Figure 3.1: Comparison of fitted values with true values for ANN and regularized linear regression considering only meteorological attributes (2 fold cross-validation only.

3.1 Results of the double cross-validation study

3.1.1 Both regions

Applying the double cross validation to the whole data set gives the results depicted in Table 3.1. From this study, we observe that the optimal number of hidden units vary significantly for each test set. The error which is in this case the mean squared error, is considerably lower for ANN and linear regression than for the baseline, which indicates the models behave properly. We also notice that λ_i* varies in a wide range of values, which is consistent with the results of the linear regression part where we showed that the optimal value of λ was not clearly defined when using 9 attributes and no quadratic terms. Finally, we observe that the errors of the ANN are lower than the ones of the linear regression with exception of the third fold. The difference between models will be further assessed in the next part about statistical comparison.

	ANN		**Linear regression**		**Baseline**
Fold	h_i*	E_i^{test}	λ_i*	E_i^{test}	E_i^{test}
1	4	0.12	1	1.08	47.81
2	7	0.32	1.e-05	0.44	46.85
3	4	0.55	1.e-05	0.34	68.96
4	4	0.08	0.1	0.58	41.61
5	4	0.74	1.e-05	1.22	58.55
6	7	0.09	0.1	0.41	63.01
7	9	1.30	1.e-05	1.45	72.94
8	4	0.56	1	0.96	25.49
9	10	0.07	1.e-05	0.59	39.27
10	5	8.97	1	9.34	69.93

Table 3.1: Results of the double cross validation applied to the whole data set (both regions).

3.1.2 Regions separately

As described before, we expect better results when applying the models to the regions separately. Table 3.2 and Table 3.3 present the results of the double cross validation for both regions. We notice that the optimal λ is constantly higher for the Sidi-Bel-Abbès region, which confirms that the model should behave differently on the two regions. We also notice higher errors for the Sidi-Bel-Abbès region which is also associated with a higher variance of the data, as can be seen with the baseline error.

	ANN		**Linear regression**		**Baseline**
Fold	h_i*	E_i^{test}	λ_i*	E_i^{test}	E_i^{test}
1	7	0.41	1e-05	0.42	20.4
2	9	0.07	1e-05	0.108	21.42
3	3	0.04	1e-05	0.12	26.98
4	7	0.24	1e-05	1.11	18.33
5	5	0.51	1e-05	0.39	40.77
6	9	0.03	1e-05	0.13	15.63
7	7	2.45	1e-05	0.52	156.45
8	8	6.23	1e-05	2.29	93.84
9	5	0.05	1e-05	0.55	27.48
10	3	0.02	1e-05	0.15	19.02

Table 3.2: Bejaia region

	ANN		**Linear regression**		**Baseline**
Fold	h_i*	E_i^{test}	λ_i*	E_i^{test}	E_i^{test}
1	5	1.024	1	0.59	51.58
2	4	1.49	1	0.29	56.27
3	4	0.19	1	0.23	26.98
4	3	0.13	1	0.11	24.28
5	8	1.68	1	1.55	93.70
6	7	20.45	0.1	24.19	65.63
7	2	13.61	1	3.2	176.02
8	3	2.77	1	4.5	111.05
9	3	0.72	1	0.39	40.54
10	7	0.58	1	2.77	54.96

Table 3.3: Sidi-Bel-Abbes region

3.2 Comparison of three models: statistical evaluation

The performance among the 3 models is compared statistically, using Setup I, which uses a t-test with $\alpha = 0.05$ and $H_0 : \theta = \theta_A - \theta_B$, where θ_A and θ_B are the performances, i.e. test error, for model A and model B.

The results for the 2 regions are shown, respectively in Table 3.4 and Table 3.5.

3.2.1 9 attributes case

The statistical comparison among the 3 models is similar for both regions as can be seen in the tables below. The comparison between the regularized linear regression and the baseline shows a negative confident interval, indicating that the baseline has an higher test error and the very low p-value, confirms that this result is not due to chance. Therefore, it can be concluded that the regularized regression is a better model than the baseline one. The same conclusion, can be drawn regarding the comparison between the ANN and the baseline. Also in this case the CI is negative and low p-value allows to conclude, with a nice margin of certainty, that the ANN model has a better performance than the baseline one.

The comparison between the ANN model and the regularized linear regression model shows a CI around 0, which indicates that the 2 models have a similar performance. Furthermore, the relatively high p-value indicates

that there is no effect, i.e. the hypothesis $M_A = M_B$ can not be discarded. Therefore, there is not enough evidence to conclude that the ANN model is better than the linear regression one or the other way around.

Model A	**Model B**	θ_L	θ_U	**p-value**
Regularized regression	Baseline	-5.54	-4.16	$2.4e^{-27}$
ANN	Baseline	-5.47	-4.25	$1.8e^{-31}$
ANN	Regularized regression	-0.16	0.14	0.43008

Table 3.4: Confidence intervals and p-values for the pairwise models in Bejaia region

Model A	**Model B**	θ_L	θ_U	**p-value**
Regularized regression	Baseline	-6.82	-5.14	$1.4e^{-27}$
ANN	Baseline	-6.73	-5.26	$2.28e^{-32}$
ANN	Regularized regression	-0.24	0.20	0.4331

Table 3.5: Confidence intervals and p-values for the pairwise models in Sidi-Bel-Abbes region

3.2.2 4 attributes case

For comparison, the statistical assessment was also realized for models trained on only meteorological attributes (this time for both regions together). The main difference in this case was that the ANN gives significantly better results than the linear regression, as can be seen in Table 3.6, where the p-value associated with this comparison is significantly low. This is expected as when considering the 9 attributes, it exists a direct relationship between between the FWI indices and the final FWI index, which is well captured by a linear regression. To the contrary, when considering only the meteorological attributes, the linear regression failed to predict satisfactorily the FWI index whereas the ANN helps significantly to capture the structure of the data.

Model A	**Model B**	θ_L	θ_U	**p-value**
Regularized regression	Baseline	-31.9	-16.9	$3.4e^{-10}$
ANN	Baseline	-48.9	-29.6	$2.0e^{-14}$
ANN	Regularized regression	-19.1	-10.7	$1.0e^{-11}$

Table 3.6: Confidence intervals and p-values for the pairwise models for the whole data set but considering only the 4 meteorological attributes.

4 Classification

In this section 3 different models are used to predict the Fire/not fire class. For this task, a Logistic regression, a K-nearest neighbors and a baseline model are implemented and statistically compared.

As for the Logistic regression, a 2-layer cross validation (K1 = 10, K2 = 10) is used to obtain the optimal lambda for each training set, i.e. the lambda that gives the lowest test error for each training set. Likewise, a 2-layer cross validation is applied to the KNN model and the inner cross validation is used to find the ideal number of k neighbours, which gives the lowest test error for each training set. A simple model, which compute the largest class on the training data, and predict everything in the test-data as belonging to that class is used as baseline model. The results of the classification is shown in Table 4.1 and Table 4.2

	KNN		Logistic regression		Baseline
Fold	k_i*	E_i^{test}	λ_i*	E_i^{test}	E_i^{test}
1	7	0.076	1e-08	0	0.38
2	4	0.153	1.75e-06	0	0.38
3	5	0.166	1e-08	0	0.5
4	10	0.166	2.94e-06	0.083	0.66
5	4	0	7.9e-08	0	0.66
6	8	0.083	6.25e-07	0	0.66
7	8	0	1e-08	0.083	1
8	6	0.166	1e-08	0.083	0.41
9	4	0	1.04e-06	0	0.83
10	4	0.083	1e-08	0	0.66

Table 4.1: Bajaia region

	KNN		Logistic regression		Baseline
Fold	k_i*	E_i^{test}	λ_i*	E_i^{test}	E_i^{test}
1	1	0.15	0.03	0	0.61
2	3	0	0.15	0	0.69
3	3	0.083	1e-08	0	016
4	3	0.083	8.28e-06	0	0.16
5	3	0.41	3.9e-05	0	0.41
6	3	0	1e-08	0	0.16
7	9	0	1e-08	0.083	0
8	3	0.083	2.8e-08	0.083	0.5
9	6	0.33	0.004	0	0.16
10	3	0.16	2.94e-06	0	0.58

Table 4.2: Sidi-Bel-Abbes region

Both KNN model and the Logistic regression predict very accurately the fire/not fire class in both region, respectively with an error of 0.0893 and 0.083 for the Bejaia region and 0.1299 and 0.083 for the Sidi-Bel-Abbes region.

4.1 Comparison with existing works

Existing works have already been performed on the same data set. Especially, [1] presents the results of different works when using a regression tree. We therefore also implemented a regression tree to fit our data. The results are summarized in Table 4.3 where we notice that the accuracy is in every case close to 80%. Here it is worth noticing that those results are considered based on the meteorological attributes, which explains the relative low accuracy of those models. In our study, we chose to focus more in the case where all the attributes are known. This would be the case if applying the FWI system method described in [2] to determine the FWI indices from the knowledge of the 4 meteorological attributes.

Reference	**Accuracy (%)**
F.Abid	82.9
Stojanova	81.2
This work	80.7

Table 4.3: Comparison of results for simple decision trees applied to predict the FWI index from the 4 meteorological attributes.

4.2 Comparison of three models: statistical evaluation

The three models are statistically compared using Setup I, which consists in using the Mcnemar test to compare the performances between 2 models at the time.

Model A	**Model B**	θ_L	θ_U	**p-value**
Logistic regression	Baseline	-0.56	-0.36	$1.4e^{-13}$
KNN	Baseline	-0.6	-0.47	$1.7e^{-16}$
Logistic regression	KNN	-0.039	0.0391	0.13

Table 4.4: Confidence intervals and p-values for the pairwise models in Bejaia region

Model A	**Model B**	θ_L	θ_U	**p-value**
Logistic regression	Baseline	-0.56	-0.38	$3.2e^{-15}$
KNN	Baseline	-0.43	-0.28	$2.07e^{-14}$
Logistic regression	KNN	-0.0007	0.08	0.125

Table 4.5: Confidence intervals and p-values for the pairwise models in Sidi-Bel-Abbes region

The results are, again, similar for both regions. Both comparisons between Logistic regression vs. Baseline and KNN model vs. Baseline, show a negative CI, which indicates that the Baseline error is likely higher than the Logistic regression and KNN one. Moreover the low p-value indicates us that this result is not likely due to chance. Therefore we can conclude that both Logistic regression and KNN have a better performance in predicting the fire/not fire class in both regions. The comparison between Logistic regression and KNN, gives a CI around 0. This could indicates that the two models have a very similar performance, if not identical. Nevertheless, the high p-value indicates that such result is likely due to chance, therefore it's not possible to conclude that the Logistic regression model has a better performance than the KNN one.

5 Summary

For regression, we used three different models (regularized linear regression, artificial neural network and baseline model) implemented two level cross validation to train and predict FWI attribute. It is obvious that ANN and linear regression models have lower test errors than baseline model. Furthermore, we can see that ANN model is even a little better than linear regression. Thus, FWI attribute could be well predicted by ANN and linear regression in our case.

For classification, k-nearest neighbor, logistic regression and baseline models were carried out to train and predict fire or not fire attribute. Both KNN and logistic regression show a good performance on predicting fire situation. Especially, logistic regression have a really low test error for fire prediction.

It is worth pointing out that both regions have a very similar fire prediction situation even though there is a larger error for Sidi-Bel-Abbes region. Therefore, those models may be used as a method to predict forest fire.

6 Question answers

Question 1: answer C

The correct answer can be obtained by considering a separation at $\hat{y} = 0$ and then increasing this value. At $\hat{y} = 0$, TPR and FPR equal to 1 when all attributes are classified positive. From the point (1,1), on the ROC curve, we see that TPR should decrease to 0.75. When the separation increases, one observation is classified negative, and TPR decreases if this observation is in reality positive. This gives answer A or C. When the separation increases again, if the second observation is again positive (as in prediction A), TPR would decrease to 0.5, which is not the case. The right answer is therefore C.

Question 2: answer D

The impurity gain is calculated using:

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

using the classification error as purity function $I(v) = 1 - max_c p(c|v)$ So we calculate I(r), I(v1), I(v2):

$$I(r) = 1 - \frac{5}{14} = \frac{9}{14}$$
$$I(v1) = 1 - \frac{5}{13} = \frac{8}{13}$$
$$I(v2) = 1 - 1 = 0$$

so:

$$\Delta = I(r) - \frac{N(v1)}{N(r)} I(v1) - \frac{N(v2)}{N(r)} I(v2) =$$
$$= \frac{9}{14} - \frac{13}{14} \cdot \frac{8}{13} = \frac{1}{14} = 0,072$$

the solution which is closer to our result is answer D.

Question 4: answer D

When b1 $\leq$ -0.76 is false and b2 $\leq$ 0.03 is also false, we can get congestion level 1. If b1 $\leq$ -0.76 is false and b2 $\leq$ 0.03 is true, we could have congestion level 2. Now, we consider the case that b1 $\leq$ -0.76 is true. If b1 $\leq$ -0.16 is true, we have congestion level 4. Otherwise, if b2 $\leq$ 0.01 is true, we also have congestion level 1, or we have congestion level 3.

Question 5: answer C

The total time is calculated using the equation:

$$5 \cdot [(5 \cdot 4 \cdot 25\,ms) + 25\,ms + (5 \cdot 4 \cdot 9\,ms) + 9\,ms] = 3570\,ms$$

where $5 \cdot []$ indicate the outer cross validation $K_1 = 5$, $5 \cdot 4$ is the loop which applied to each hidden unit/lambda the inner cross validation $K_2 = 4$, once the best paramter is found, another train/test is done in the outer set for both ANN and regularized regression.

References

[1] Faroudja Abid, Nouma Izeboudjen, 06 February 2020, *Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm*

[2] C.E Van Wagner, 1987, *Development and Structure of the Canadian Forest Fire Weather Index System*