University of Liège

Project 3 - Supervised classification

MATH2021-1 - High-dimensional data analysis

Yann Claes (s161317)
Gaspard Lambrechts (s161826)
François Rozet (s161024)

MSc in Data science and engineering

Academic year 2019-2020

Data

The 500 measures of pollutant concentrations and atmospheric conditions from the Shunyi station used in project 1 and 2 having a very unbalanced binary indicator (`rain`), it has been decided to restart from the 35 000 measures made at this station, and work on the mean daily pollutant concentrations. From this new data, the binary variable will be defined as "exceeding the weighted WHO[1] recommendations for *at least* one of the pollutants". This binary variable will be detailed below.

Taking the average of the 24 (one per hour) concentrations for each day, and considering only the rows without missingness (complete case analysis), the data is now composed of 1461 observations and 13 variables, described in the table 1.

year	year
month	float month[2]
PM2.5	daily mean PM2.5 concentration
PM10	daily mean PM10 concentration
SO_2	daily mean SO_2 concentration
NO_2	daily mean NO_2 concentration
O_3	daily mean O_3 concentration
temp	daily mean temperature
pres	daily mean pressure
dewp	daily mean dew point
rain	daily mean rain
wspd	daily mean wind speed
wdir	daily dominant wind direction[3]

Table 1 – Variables of the new dataset

Hence, `day` has been removed because it is contained in `month` and `CO` has been removed because no WHO recommendations are made for this pollutant. All other variables except `wdir` have been aggregated by mean of means. `wdir` has been taken as kind of the most frequent wind direction over the day, taking the wind speed into account. This data is saved in the file `products/csv/data.csv`.

Averaging over the days has consequences since all peaks will be smoothed, but for the needs of this project (predicting if one of the daily mean pollutant concentration will surpass a given value), it does not seem too problematic to work on the daily mean of each variable. Furthermore, as a sensor is often down for a certain number of time (and thus a series of consecutive measures), some of the daily means will be measured on only

[1] World Health Organization

[3] Mapping of each day (at noon) of the year to a float between 0 and 12.

[3] The most common wind direction, with frequency of appearance weighted by the wind speed. Expressed as an angle (rad) oriented counterclockwise from east direction.

on very few measures, being certainly biased. It will be considered as acceptable given the low missing rate of the data.

Binarization of the data

For the following questions, the pollutant concentrations will all be replaced by one single binary indicator: *"Does one of the pollutant mean concentrations overpass the WHO recommendations over a day weighted by some margin ?"*, represented by the variable `alert`.

The WHO recommendations can be found in Table 2. It has been chosen to use the maximum mean concentration over 8 h as the daily mean O_3 concentration threshold (as no data is provided for one day). For other pollutants, we used the actual one-day threshold given by the WHO.

The binary variable has been defined as

```
alert = any(mean_pollutant_concentrations > margins * recommendations)
```

where the `margins` have been artificially tuned to `[3 3 1 1 1]` and the recommendations taken to `[25 50 20 200 100]` as said above. We justify this arbitrary margins by the fact that the particle matter (PM) concentrations are around 80% of the time above the WHO recommendations at Shunyi station in Beijing, it can thus seem not useful to alert for a danger each time that the PM concentrations are above these thresholds, but alert when these concentrations reach a less frequent but more dangerous threshold, such as three times the WHO recommendation.

Pollutant	Max. mean exposition over			
	1 y	1 d	8 h	10 min
PM2.5	$10\,\mu g\,m^{-3}$	$25\,\mu g\,m^{-3}$		
PM10	$20\,\mu g\,m^{-3}$	$50\,\mu g\,m^{-3}$		
SO_2		$20\,\mu g\,m^{-3}$		$500\,\mu g\,m^{-3}$
NO_2	$40\,\mu g\,m^{-3}$	$200\,\mu g\,m^{-3}$		
O_3			$100\,\mu g\,m^{-3}$	

Table 2 – Recommendation for maximum mean pollutant exposition over a certain period

Feasibility of the classification

On the one hand, from the analysis conducted in part 1 of the project, we know that there exists a strong correlation structure between the pollutant concentrations and the atmospheric conditions (`temp`, `dewp`, `pres`). Our binary variable being based on the pollutant concentrations, achieving a classification based on all the other variables (including the atmospheric ones) seems promising.

On the other hand, the PCA analysis conducted in part 2 of the project indicated that the total variance of the dataset was mainly contained in two independent (by construction) principal components (gathering 68 % of the total variance), the first one being roughly composed of the pollutant variables, the second one being roughly composed of the atmospheric variables. We thus see that a non negligible part of what we would like to be in our explanatory variables (atmospheric variables) is uncorrelated with what has constituted, after binarization, our dependent variable (pollutant variables).

Also, from project 1, we know that the `year` doesn't have much of an influence on other variables. Therefore, it shouldn't have any significant influence over the `alert` variable.

1 Classification using the logistic regression model

1.1 Finding a logistic regression model

Using as explanatory variables all the remaining quantitative variables (i.e. `year`, `month`, `temp`, `pres`, `dewp`, `rain`, `wspd` and `wdir`), we built a first logistic regression model to predict the binary value of `alert`. This yielded the coefficients summary of Table 3.

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	
Intercept	426.099 86	110.895 46	3.842	0.000 122	***
year	−0.174 30	0.055 17	−3.159	0.001 581	**
month	−0.127 90	0.021 13	−6.053	1.42×10^{-9}	***
temp	−0.241 25	0.021 52	−11.210	$< 2 \times 10^{-16}$	***
pres	−0.070 32	0.012 89	−5.456	4.88×10^{-8}	***
dewp	0.150 57	0.016 33	9.220	$< 2 \times 10^{-16}$	***
rain	−1.173 92	0.327 61	−3.583	0.000 339	***
wspd	−0.513 87	0.111 19	−4.622	3.81×10^{-6}	***
wdir	0.327 87	0.039 44	8.312	$< 2 \times 10^{-16}$	***

Table 3 – Coefficients summary of the full logistic model

The corresponding AIC value is equal to 1572.6.

From the summary, we see that all explanatory variables seem to have an important influence on the logistic model. The last column ($\Pr(>|z|)$) represents the probability that a coefficient that should be zero has a higher absolute value than the estimated coefficient. This value being extremely low for all coefficients, we can reject the null hypothesis of having a zero coefficient.

We then decided to draw a correlation plot between the explanatory variables, in order to spot a potential multicollinearity between them, leading to Figure 1. From this plot, we can see that `temp`, `pres` and `dewp` are strongly correlated, which is not a surprise if we keep in mind the correlation analysis that was conducted in the first project.

To emphasize this behavior, we computed the associated VIF scores, which can be found in Table 4.

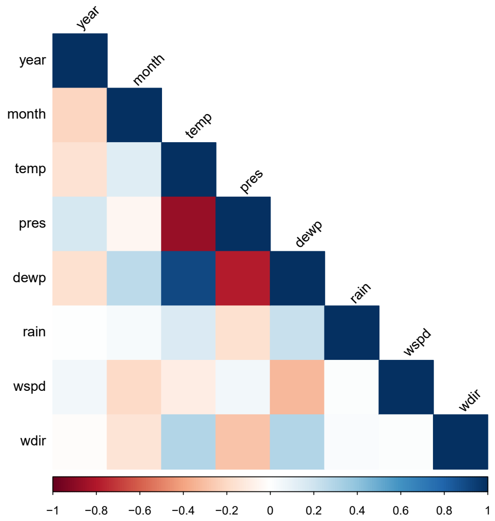


Figure 1 – Correlation plot of the explanatory variables

From these VIF scores, we can see that the temperature and dew point have very big VIF scores compared to all other variables, and that the pressure has a greater VIF score than the other explanatory variables (except `temp` and `dewp`). We should thus keep these scores in mind when looking at the estimated coefficients of Table 3, as they might lead to

year	month	temp	pres	dewp	rain	wspd	wdir
1.091	1.405	14.553	4.316	12.649	1.166	1.462	1.232

Table 4 – VIF scores of the explanatory variables of the full model

a wrong insight of the interpretability of the influence of the explanatory variables due to this multicollinearity behavior. In order to check whether a major issue is introduced, we constructed three different models where each one contained all the other variables and either `temp`, `pres` or `dewp`. Then, we compared the coefficients of these three variables (in terms of sign and absolute value), leading to Table 5.

	`temp`	`pres`	`dewp`
Full model	-0.24125	-0.07032	0.15057
Reduced models	-0.023704	-0.0001157	-0.0005504

Table 5 – Coefficients of `temp`, `pres` and `dewp`

From this table, we can observe that the parameters differ a lot from the full model to the reduced models, emphasizing the impact of multicolinearity.

To get rid of this problem, we decided to replace these three variables by the first principal component of the data restricted to these three variables, and to build a model with this principal component as well as the remaining variables. This yielded the coefficients summary of Table 6, with an AIC of 1717.

	Estimate	Std. Error	z value	Pr($>|z|$)	
Intercept	317.876628	103.695226	3.065	0.00217	**
year	-0.157033	0.051459	-3.052	0.00228	**
month	-0.081628	0.018506	-4.411	1.03e-05	***
rain	-0.231231	0.261069	-0.886	0.37577	
wspd	-1.071795	0.103444	-10.361	$<$ 2e-16	***
wdir	0.321574	0.036226	8.877	$<$ 2e-16	***
PC1	-0.004691	0.003484	-1.346	0.17817	

Table 6 – Coefficients summary of the modified logistic model

Deriving this modified model does not come with a gain of interpretability of the model.

In order to select the optimal set of explanatory variables, we decided to perform a backward AIC selection of the variables using the function `stepAIC`. This resulted in the optimal model corresponding to the model where the variable `rain` is removed, as the AIC is minimized in that case. We also performed a forward variable selection, which yielded the same results[4].

[4] These results can be found in the source code `logistic_regression.R`.

Interpretation of the model From the coefficients of Table 6, we can see that the principal component resuming the three colinear variables will have very little impact on the classification rule. This result was suggested in the preliminary discussion about the feasibility of the prediction. Most of the other variables (except `rain` which is removed from the model) have a negative coefficient, while only the intercept and `wdir` have positive coefficients. For example, the bigger the wind speed, the smaller the probability of success, and thus the smaller the probability of having a pollutant concentration surpassing the recommendations.

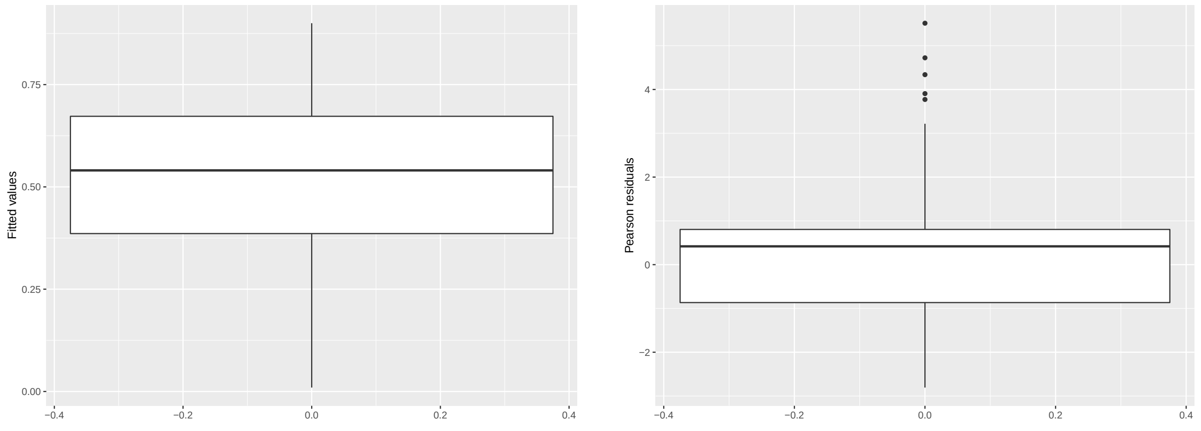
We have

$$\begin{aligned}\operatorname{logit}(\pi) &= 317.88 - 0.157\texttt{year} - 0.082\texttt{month} - 1.072\texttt{wspd} + 0.321\texttt{wdir} - 0.005\texttt{PC1} \\ \pi &= \frac{\exp(317.88 - 0.157\texttt{year} - 0.082\texttt{month} - \ldots)}{1 + \exp(317.88 - 0.157\texttt{year} - 0.082\texttt{month} - \ldots)}\end{aligned}$$

The classification rule is

$$\pi > 0.5$$

Fitted values and residuals As far as the fitted values are concerned, we get the boxplot of Figure 2a. The mean value of the fitted values is equal to 0.515 647, which is quite close to the median, and 75 % of the fitted values lie between 0.386 870 and 0.674 243. We can thus observe that the model does not predict extreme values but rather moderate values for the probability of success, which corroborates what has been concluded in the correlation structure of project 1. The computed residuals correspond to the Pearson residuals, and are summarized in the boxplot of Figure 2b. The mean of these residuals is equal to 0.009 266, meaning that residuals are, on average, close to zero, and thus that the fitted model is quite good on the data we provided it with.



(a) Boxplot of the fitted values

(b) Boxplot of the Pearson residuals

The error rate of the predictions is equal to 0.310 099 6, which means that the model does not seem to be too bad at determining whether a pollutant concentration will surpass the recommendations or not. We should however keep in mind that this rate was computed on the same dataset as the one used to build the model. It is thus normal that the error rate is not too high.

1.2 Performance of the classification model

As a reminder, the optimal model is composed of the following variables: `year`, `month`, `wspd`, `wdir` and `PC1`. The leave-one-out cross-validation yielded the confusion matrix of Table 7. From this confusion matrix, we can compute the total error rate which is around 50 %. This rate is much higher than the one obtained previously, which shows that the prediction model is maybe not as good as expected.

		Estimated	
		0	1
True	0	400	281
	1	440	285

Table 7 – Confusion matrix for the leave-one-out technique.

The sensitivity and specificity values have also been computed, and are respectively equal to 0.393 103 4 and 0.587 371 5. This means that the model fails at predicting a surpassing pollutant concentration around six times out of ten, and correctly predicts the non-surpassing behavior about 60 % of the time.

To get more insight about the performance of the classification model, we decided to draw the ROC curve corresponding to the leave-one-out technique, yielding Figure 3.

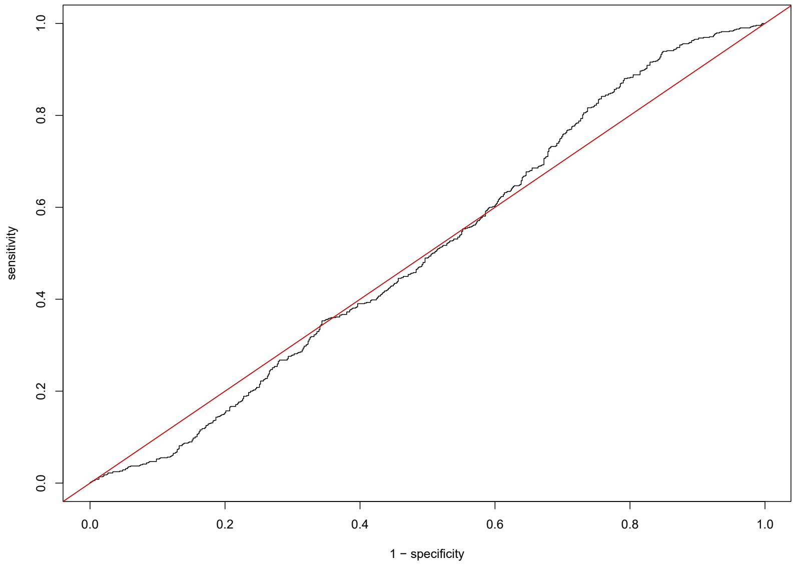


Figure 3 – ROC curve of the leave-one-out technique

We can indeed see that the ROC curve follows a curve quite similar to the random model, *i.e.* the model predicting 0 or 1 with equal probabilities. To corroborate this observation, we computed the area under the ROC curve, which is equal to 0.5053643.

2 Classification based on LDA scores

2.1 Canonical variable and scores

Initially using all the quantitative variables (centered and scaled), the LDA procedure resulted in the most discriminant direction described by the coefficients of the Table 8.

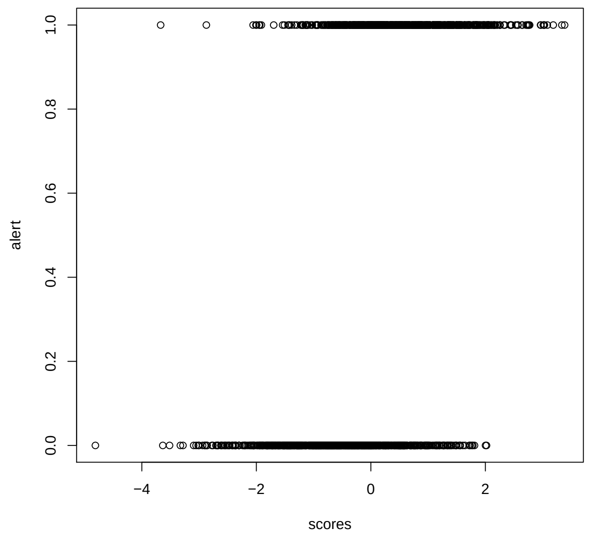
year	month	temp	pres	dewp	rain	wspd	wdir
−0.180	−0.375	−2.234	−0.605	1.721	−0.201	−0.330	0.516

Table 8 – LDA canonical projection coefficients

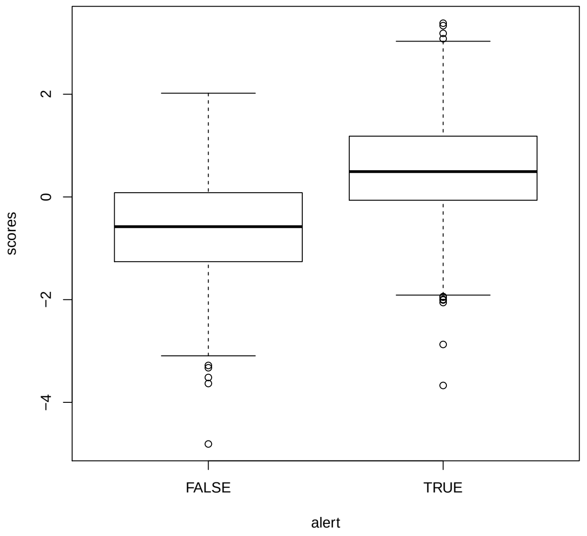
As a result, the canonical variable expression is

$$Z_1 = -0.180\texttt{year} - 0.375\texttt{month} - 2.234\texttt{temp} - 0.605\texttt{pres} + \dots$$

The scores are then obtained by projection, i.e. by evaluating Z_1 for each row of the dataset.



(a) Scatter plots of the scores w.r.t `alert`



(b) Boxplots of the scores w.r.t `alert`

As one can see, the LDA procedure didn't succeed greatly at separating both distributions, which is corroborated by its quite low discriminant power $\gamma_1 = 0.2406$. This result means that both distributions overlap a lot within the $\mathbb{R}^p$ space, it is therefore not possible to completely separate them. The reason could simply be that the quantitative variables are not sufficiently linked to the constructed `alert` values.

2.2 Suppressing less discriminant variables

In table 8, one can observe that the variables `year` and `rain` are the one with the lowest influence on the canonical variable and, therefore, the score. As far as `year` is concerned, this observation sticks with the preliminary analysis. For `rain`, we can see a parallel with the previous section (Logistic regression) where `rain` was removed from the dataset to increase the AIC.

Indeed, by looking at Table 9, we can see that removing either `year` or `rain` doesn't significantly reduces the discriminant power of the method, conversely to any other variables.

out	year	month	temp	pres	dewp	rain	wspd	wdir
γ	0.2351	0.2206	0.1527	0.2242	0.1872	0.2340	0.2278	0.1992

Table 9 – Leave-one-variable-out discriminant powers

By removing both, the cardinal variable becomes

$$Z_1' = -0.326\texttt{month} - 2.137\texttt{temp} - 0.642\texttt{pres} + 1.555\texttt{dewp} - 0.389\texttt{wspd} + 0.543\texttt{wdir}$$

and its associated discriminating power decreases to 0.2281.

2.3 Classification using the scores

Using the simplified model (without `year` and `rain`), we computed, as before, the scores z of each individual x. The perfect classification rule would be to assign to an individual the group with the highest probability knowing the individual. However, in practice, it is not possible to compute exactly this probability.

Yet, we can approximate the result of such rule by selecting the group with greatest posterior probability[5] that the individual belongs in it. Moreover, under the normality assumption and the assumption of *homoscedasticity* of the covariance matrices in each group, one can show that finding the greater posterior probability is the same as finding the group whose center (mean) is the closest to the individual.

Therefore if we wish to classify the rows based on their scores (obtained previously) using that rule we have

$$\text{group} = 1\left(|z - \bar{z}_1| \leq |z - \bar{z}_0|\right)$$

where $\bar{z}_1$ and $\bar{z}_0$ are respectively the mean scores of the group 1 and 0.

Applying such rule on our dataset yields table 10.

		Estimated	
		0	1
True	0	477	204
	1	206	519

Table 10 – Confusion matrix of leave-one-out LDA classification.

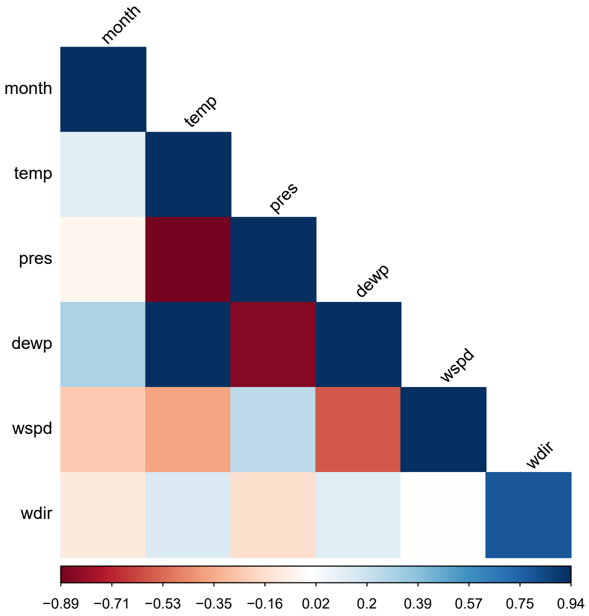
Surprisingly, the results aren't that bad with 70.8 % of correct classification !

[5] If the prior probability was (very) unbalanced we should have considered it in the computations, fortunately it is not the case (we are almost perfectly balanced). Also, potential misclassification cost haven't been considered.

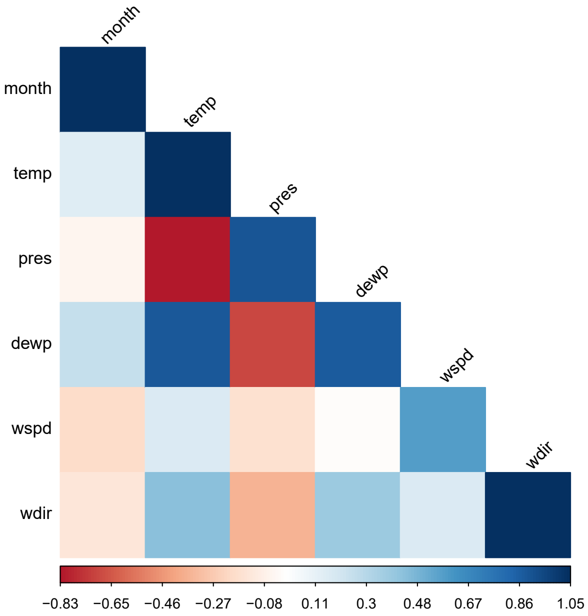
2.4 Homoscedasticity assumption

As stated in the previous section, the classification rule has been partially derived from the assumption of *homoscedasticity* of the covariance matrices in each group. Yet, this assumption has still to be validated, i.e. to prove that the variability of the quantitative variables is (more or less) the same in each group.

First of all, we computed the scores' variance in both groups : 1.101 for the group 0 and 0.905 for the group 1. These are fairly close. We also computed the covariance matrices in both groups.



(a) Covariance matrix of group 0



(b) Covariance matrix of group 1

As we can see, both matrices aren't totally similar, yet not dissimilar enough to invalidate the homoscedasticity assumption.