Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	2	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 4

(Submitted 30.05.2022)

Task 1

Suppose you have ten observations of a binary response $Y \in a, b$ with the true class labels

$$(y1, ..., y10) = (a, a, a, a, a, b, b, b, b, b) \tag{1}$$

where a denotes the positive class.
Construct two vectors of estimates of the response, $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$, so that, when comparing to y,

- **Accuracy of both is greater than 0.6.**
- **F1-score and Matthews' correlation coefficient give a different ranking of estimates, that is,**

$$F1(\hat{y}^{(1)} > F1(\hat{y}^{(2)} \text{and} MCC(\hat{y}^{(1)} < MCC(F1(\hat{y}^{(2)}) \tag{2}$$

or vice versa.
Explain how you found your solution.
We found the solution that fulfills all conditions by starting with two vectors with no false negatives and no false positives respectively, but three entries in the other category were false:

$$\hat{y}^{(1)}_{guess1} = (a, a, a, a, a, b, b, a, a, a) \tag{3}$$

$$\hat{y}^{(2)}_{guess1} = (b, b, b, a, a, b, b, b, b, b) \tag{4}$$

We used the following formulas to calculate precision, recall, F_1-score and Matthews' correlation coefficient (MCC):

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

Which yielded the following values:

	$\hat{y}^{(1)}_{guess1}$	$\hat{y}^{(2)}_{guess1}$
TP	5	2
FP	3	0
TN	2	5
FN	0	3
precision	$\frac{5}{8}$	1
recall	1	$\frac{2}{5}$
F_1	$\frac{10}{13}$	$\frac{4}{7}$
MCC	$\frac{1}{2}$	$\frac{1}{2}$

Table 1: Precision, recall, F_1-score and MCC for the first iteration of vectors.

We can see in Figure 1 that the MCC is the same for both vectors. This implies thta the proportion of Positives and Negatives in both categories is important. We decreased the number of false negatives in the second array by one, which yielded the following vectors and values:

$$\hat{y}^{(1)}_{guess2} = (a, a, a, a, a, b, b, a, a, a) \tag{9}$$

$$\hat{y}^{(2)}_{guess2} = (b, b, a, a, a, b, b, b, b, b) \tag{10}$$

	$\hat{y}^{(1)}{}_{guess2}$	$\hat{y}^{(2)}{}_{guess2}$
TP	5	3
FP	3	0
TN	2	5
FN	0	2
precision	$\frac{5}{8}$	1
recall	1	$\frac{2}{5}$
F_1	$\frac{10}{13}$	$\frac{4}{7}$
MCC	$\frac{1}{2}$	$\frac{\sqrt{21}}{7}$

Table 2: Precision, recall, F_1-score and MCC for the first iteration of vectors.

In this iteration we can see (Figure 2) that $F_1^1 > F_1^2$ and $MCC^1 < MCC^2$, while also fulfilling the condition that the accuracy should be greater than 0.6.

Task 2

a) Download and load the phoneme data set (phoneme.csv) from moodle. Split the dataset into training and test set according to the speaker column. Be sure to exclude the row number, speaker and response columns from the features. How many data points are in the training and test set, respectively? Create a subset of the original data that contains only data points with a response value aa or ao. What is the size of training and test set now? Useful functions: strsplit(), grepl()

Number of observations in the data set:

- training set: 3340 observations
- test set: 1169 observations
- subset training set: 1278 observations
- subset test set: 439 observations

b) Fit a logistic regression model on the data subset from the previous task and report train and test accuracy assuming a classification threshold of 0.5. Useful functions: glm()

		test data response	
		aa	ao
glm prediction	aa	121	52
	ao	55	211

The test accuracy of glm models is: $\frac{121+211}{121+211+55+52} = \frac{332}{439} = 0.756$ The accuracy of 76% is quite good. So the fitted model on the train data predicts "good" response results for the test data.

The train accuracy of glm-model is: 0.907 = 91%

c) Repeat step (b) using LDA and report your findings. Useful functions: lda() from the MASS package

		test data response	
		aa	ao
lda classification	aa	121	39
	ao	55	224

The test Accuracy of the lda-models is: $\frac{121+224}{121+224+39+55} = \frac{345}{439} = 0.786$ The accuracy is even a bit better than the accuracy from the glm-model.

The train accuracy of lda-model is: 0.894 = 89%. There is a very small difference (2%) in the train accuracy of both models.

d) Generate confusion matrices for the logistic regression and the LDA model for aa and ao. Which differences can you observe between the models?

		lda.class	
		aa	ao
glm.pred	aa	148	25
	ao	12	254

Table 3: Confusion matrices for glm-model and lda-model

Table 3 displays the confusion matrices and compares the both models and their classification. For the class 'aa' we can identify 148 matches in the GLM- LDA-model for 'ao' we have 254 matches. Overall 37 classifications were different in the models. Accuracy of both models compared is: $\frac{148+254}{148+25+254+12} = \frac{402}{439} = 0.916$ this means both models come in 92% to the same classification results.

If we look on the confusion matrices compared each model with the test result:

		test data response	
		aa	ao
glm prediction	aa	121	52
	ao	55	211

		test data response	
		aa	ao
lda classification	aa	121	39
	ao	55	224

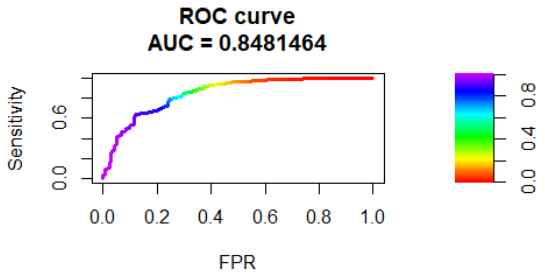
The test accuracy of the GLM model is: $\frac{121+211}{121+211+55+52} = \frac{332}{439} = 0.756$ This means the GLM model predicts to 76% the same response values as the test subset.

The test accuracy of the LDA models is: $\frac{121+224}{121+224+39+55} = \frac{345}{439} = 0.786$ This means the LDA model predicts to 79% the same response values as the test subset.

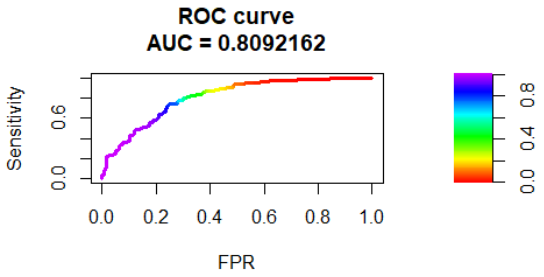
Both models have the same "true positive" counts and "false negative" counts. They only differ in counts of "true negative" and "false positive". The accuracy of model the LDA model is better than the one of the GLM- model.

e) **For both learned models, plot ROC and PR curves for evaluating the performance on test data without specifiying a particular threshold. Treat ao as the positive class. Give the respective area under the curves. Given all results obtained so far, would you prefer logistic regression or LDA in this example? Why? Useful functions: roc.curve and pr.curve from the PRROC package**

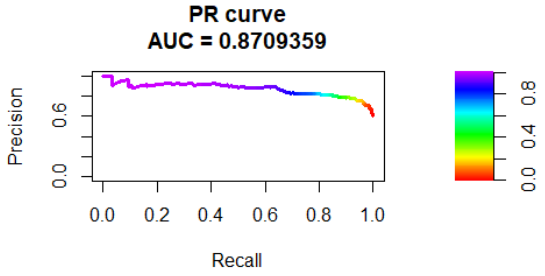
We plotted both the ROC- and PR-curves for LDA and logistic GLM. With the ROC-curve area under the curve for LDA was 0.85 (Figure 1a), the area under the curve for GLM was 0.81 (Figure 1b). This meant a better tradeoff of false-positive-rate and sensitivity for LDA. For the PR-curve, the area under the curve was 0.87 for LDA (see Figure 1c) and 0.84 for GLM (see Figure 1c). This meant that there was a better trade-off between low false negative and low false positive rate for LDA. Because both measures have better values for LDA and the accuracy was also higher, we would choose LDA.



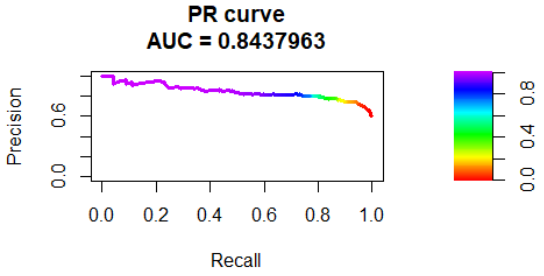
(a) ROC curve for binary LDA



(b) ROC curve for logistic GLM



(c) Precision-Recall-curve for binary LDA



(d) Precision-Recall-curve for logistic GLM

Figure 1: Plots of ROC and PR curves for glm-model and lda-model

f) Fit an LDA model using the full training data set (all five phonemes). Report train accuracy, test accuracy, and the entire confusion matrix. Compare with the result for the two-class LDA from Task c). What do you observe? Which factors are primarily responsible for possible differences? We obtained a train accuracy of 94.4% and a test accuracy of 92.0% for the LDA model trained on the full training data. This implies even better performance when taking into account all classes of phonemes. We got the following confusion matrix for the real labels for the train set and the predicted labels.

	aa	ao	dcl	iy	sh
aa	424	82	0	0	0
ao	95	677	0	0	0
dcl	0	0	553	1	0
iy	0	0	8	851	0
sh	0	0	1	0	648

Table 4: Confusion matrix between predicted train labels and real train labels

	aa	ao	dcl	iy	sh
aa	129	39	0	0	0
ao	47	223	0	0	0
dcl	0	0	190	2	0
iy	0	1	5	309	0
sh	0	0	0	0	224

Table 5: Confusion matrix between predicted test labels and real test labels

You can see that there is quite some misclassification between "aa" and "ao", but the misclassification is much lower for the other phonemes. This applies for the train confusion matrix (Table 4) as well as for the test confusion matrix (Table 5). The results imply that "aa" and "ao" have similar parameters (and thus sound similar) and might be harder to distinguish than the other phonemes. This could be the main reason for the higher accuracies with the full model.