# Prediction of wine sensorial quality: a classification problem

Maurizio Carpita, Silvia Golia

## 1. Introduction

When dealing with a wine, it is of interest to be able to predict its quality based on chemical and/or sensory variables. There is no agreement on what wine quality means, or how it should be assessed and it is often viewed in intrinsic (physicochemical, sensory) or extrinsic (price, prestige, context) terms (Jackson, 2017). For example, in Golia et al. (2017) it was measured by a global score of quality, ranging from 0 to 100, produced by Altroconsumo, an Italian independent consumer's association, and based on a large set of variables including chemical and sensory variables, as well as variables of context. Cortez et al. (2009) used an indicator, ranging from 0 to 10 with 0 meaning very bad and 10 excellent, obtained from the evaluations of experienced judges who scored the wines.

In this study we started from the Cortez et al. (2009) paper, but we maintained the categorical nature of the variable measuring the wine sensorial quality. The approach to the prediction of this categorical variable followed by Cortez and coauthors makes use of the observed wine quality, but it suffers from the fact that it is necessary to know the wine quality measure. Instead, in this paper we started from the predicted probabilities' record of the categories of the target variable, obtained from the application of the Cumulative Logit Model, and then we applied a classifier in order to predict the final category. This last step is the one of interest for this paper; in fact we will compare the predictive performances of the default method (Bayes Classifier), which assigns a unit to the most likely category, and other two methods (Maximum Difference Classifier and Maximum Ratio Classifier). In order to do that, we will use the data analysed in Cortez et al. (2009) concerning both the white and red variants of the Portuguese "Vinho Verde" wine.

The paper is organized as follows. Section 2 discusses the categorical classifiers used in this study, whereas Section 3 reports the results concerning the prediction of the wine sensorial quality. Conclusions follow in Section 4.

## 2. The categorical classifiers

As stated in the introduction, the statistical problem of this study refers to the way in which the record of the predicted occurrence probabilities of each of the categories of the categorical target variable is transformed into a single value. The default method is the *Bayes Classifier* (BC), which assigns a unit to the most likely category. BC has the property to minimize, on average, the test error rate (James et al., 2013), so it is the optimal criterion when the accuracy of the classification is the main goal. Nevertheless, BC favors the prevalent category most and when there is not a category of interest but all the categories have the same relevance, it can not be the best choice.

Starting from this observation, in Golia and Carpita (2018, 2020) we have investigated the performances of different categorical classifiers (some of them take into account also the ordinal nature of the target variable) and we have found the so-called *Maximum Difference Classifier* (MDC) promising. In this study we considered MDC and a new classifier denoted as *Maximum*

Maurizio Carpita, University of Brescia, Italy, maurizio.carpita@unibs.it, 0000-0001-7998-5102
Silvia Golia, University of Brescia, Italy, silvia.golia@unibs.it, 0000-0003-0015-8126

*Ratio Classifier* (MRC). Both classifiers are based on the comparison between the predicted probabilities and the sample frequencies and they are defined as follows.

Let $pr_i$ be the predicted probability of the category $c_i$ ($i = 1, 2, \ldots, k$) of the categorical variable $C$, and let $fr_i$ be the corresponding frequency computed from observed data. The MDC computes the deviations of $pr_i$ from $fr_i$ and takes the category corresponding to the maximum difference, that is:

$$MDC : \arg \max_{i \in (c_1, c_2, \ldots, c_k)} (pr_i - fr_i).$$

This classifier represents the extension of what proposed by Cramer (1999) for the dichotomous case.

The MRC computes the relative deviations of $pr_i$ from $fr_i$ and takes the category corresponding to the maximum ratio, that is:

$$MRC : \arg \max_{i \in (c_1, c_2, \ldots, c_k)} (pr_i / fr_i).$$

## 3. The prediction of wine quality

The data under study concern the sensorial quality of the white and red variants of the Portuguese "Vinho Verde" wine (Cortez et al., 2009). The wine quality was measured by a sensory preference variable, from now on denoted as SPV, using a 0-10 scale. For each wine, eleven of the most common physicochemical variables were recorded; they represent the explanatory variables for the SPV, which is the target variable. Table 1 reports the frequencies of SPV scores observed in the white and red wine data sets; not all the available scores were used and some of them own a low frequency.

Table 1: Frequencies of the sensory preferences observed in the white and red wine data sets

| Scores | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| White wines | 0.004 | 0.033 | 0.297 | 0.449 | 0.180 | 0.036 | 0.001 |
| Red wines | 0.006 | 0.033 | 0.426 | 0.399 | 0.124 | 0.011 | – |

The model used to study and predict the occurrence probabilities of each of the categories of the SPV, is the *Cumulative Logit Model* (CLM) (Agresti, 2010), defined as follows. Let $Y$ be a categorical target variable with $k$ ordinal categories $\{1, 2, \ldots, k\}$, and let $\{X_1, \ldots, X_p\}$ be a set of explanatory variables; for the statistical unit $s$, the CLM has the following form:

$$\text{logit}[P(Y_s \leq i)] = \log \frac{P(Y_s \leq i)}{1 - P(Y_s \leq i)} = \alpha_i + \sum_{m=1}^{p} \beta_m x_{sm}, \quad \text{for } i = 1, 2, \ldots, k - 1.$$

Once estimated the parameters, it is possible to use the model for predictive purposes, so the CLM gives the $k$ predicted probabilities that are passed to the categorical classifier.

In order to evaluate the predictive performance of a classifier, some indicators computed from the confusion matrix can be used. In this study they are: the *Sensitivity* (Sen) of each category, the *Maximum Distance Between Sensitivities* (MDBSen), the *Overall Accuracy* (OvAc), the *Macro Average F1 score* (MAF1) and the *Kappa statistic* (Kappa) (Raschka and Mirjalili, 2019). $\text{Sen}_i$ expresses how well the classifier recognizes a unit belonging to the category $c_i$. MDBSen, defined as:

$$\text{MDBSen} = \max_{i \neq j} |\text{Sen}_i - \text{Sen}_j|,$$

highlights the balanced or unbalanced ability of the classifier to assign a unit to the right category, the lower the MDBSen, the more balanced the classification. The OvAc is the rate of correct classification and it is the indicator maximized by BC. The MAF1 is another indicator to measure the accuracy of the classifier and it is obtained as the average of the F1 scores class-by-class. The choice of MAF1 instead of the weighted average F1 score, is linked to the will to attribute the same relevance to all classes. Kappa is used to measure the agreement between the actual and the predicted classifications of a dataset, while correcting for agreement occurred by chance.

Table 2 reports the value of these statistics computed on the base of the in-sample prediction of the SPV of all the available wines. For the sake of clarity, we added the percentage variation of OvAc, MAF1 and Kappa with respect to the value obtained applying BC in the last three rows.

Table 2: Performance Indicators for white and red wines

| | White wines | | | Red wines | | |
|---|---|---|---|---|---|---|
| | BC | MDC | MRC | BC | MDC | MRC |
| $Sen_3$ | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 | 0.600 |
| $Sen_4$ | 0.031 | 0.049 | 0.258 | 0.000 | 0.019 | 0.208 |
| $Sen_5$ | 0.493 | 0.668 | 0.386 | 0.739 | 0.720 | 0.355 |
| $Sen_6$ | 0.757 | 0.444 | 0.342 | 0.614 | 0.527 | 0.392 |
| $Sen_7$ | 0.222 | 0.522 | 0.301 | 0.271 | 0.482 | 0.357 |
| $Sen_8$ | 0.000 | 0.000 | 0.091 | 0.000 | 0.000 | 0.556 |
| $Sen_9$ | 0.000 | 0.000 | 0.800 | | | |
| MDBSen | 0.757 | 0.668 | 0.709 | 0.739 | 0.720 | 0.392 |
| OvAc | 0.527 | 0.493 | 0.336 | 0.593 | 0.577 | 0.369 |
| MAF1 | 0.215 | 0.229 | 0.200 | 0.272 | 0.285 | 0.253 |
| Kappa | 0.230 | 0.252 | 0.131 | 0.329 | 0.327 | 0.161 |
| Var. OvAc | – | -6.429 | -36.251 | – | -2.740 | -37.829 |
| Var. MAF1 | – | 6.499 | -6.912 | – | 4.781 | -6.997 |
| Var. Kappa | – | 9.617 | -43.184 | – | -0.363 | -50.891 |

In the face of an expected but limited reduction in OvAc (6.4% for white wines and 2.7% for red wines), MDC performs better than BC with respect to MAF1 and Kappa and shows more balanced values of the sensitivities, especially for the white wines. MRC outperforms both BC and MDC in terms of balancing the sensitivities, but loses a lot in terms of OvAc and Kappa.

Given that the lowest and highest sensory preferences have low frequency, we merged the first two and the last two categories for both the two varieties of wine, obtaining a SPV on a 5-category ordinal scale for white wine and on a 4-category ordinal scale for red wines.

Table 3 reports the indicators of Table 2 with the exception of the sensitivities of the single categories. The results show the same behaviour observed in Table 2. It is of interest to note that also in this case there are some categories with a low frequency and others that absorb the majority of the statistical units.

## 4. Conclusions

In this paper we investigated the impact of different classifiers in the capability to predict the wine sensorial quality of the Portuguese "Vinho Verde" wine. We have studied this variable applying the CLM for prediction purposes. We have transformed the prediction of the occurrence probabilities of each of its categories into a single sensory preference through three

Table 3: Performance Indicators for white and red wines after merging some categories

| | White wines (5 categories) | | | Red wines (4 categories) | | |
|---|---|---|---|---|---|---|
| | BC | MDC | MRC | BC | MDC | MRC |
| MDBSen | 0.757 | 0.650 | 0.157 | 0.740 | 0.686 | 0.370 |
| OvAc | 0.528 | 0.494 | 0.358 | 0.602 | 0.585 | 0.432 |
| MAF1 | 0.302 | 0.332 | 0.304 | 0.422 | 0.445 | 0.400 |
| Kappa | 0.232 | 0.254 | 0.153 | 0.343 | 0.340 | 0.225 |
| Var.OvAc | – | -6.381 | -32.173 | – | -2.703 | -28.170 |
| Var.MAF1 | – | 9.940 | 0.659 | – | 5.394 | -5.242 |
| Var.Kappa | – | 9.798 | -33.956 | – | -0.969 | -34.502 |

different classifiers, the BD, the MDC and the MRC. The results have shown that, despite an expected but limited reduction of the overall accuracy, the MDC seems to be the suitable categorical classifier in an unbalanced context (that is when some categories absorb almost all the statistical units) and when all the categories have equal importance (i.e. different types of mis-classification do not involve different costs).

# References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd ed*. John Wiley & Sons, Hoboken, New Jersey.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47**, pp. 547–5.33.

Cramer, J.S. (1999). Predictive performance of the binary logit model in unbalanced samples. *The Statistician*, **48**(1), pp. 85–94.

Golia, S., Brentari, E., Carpita, M. (2017). Causal reasoning applied to sensory analysis: The case of the Italian wine. *Food Quality and Preference*, **59**, pp. 97–108.

Golia, S., Carpita, M. (2018). On classifiers to predict soccer match results, in *ASMOD 2018: Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data*, eds. S. Capecchi, F. Di Iorio and R. Simone, FedOAPress, pp. 125–132.

Golia, S., Carpita, M. (2020). Comparing classifiers for ordinal variables, in *Book of short papers SIS 2020*, eds. A. Pollice, N. Salvati and F. Schirripa Spagnolo, Pearson, pp. 1160–1165.

Jackson R.S. (2017). *Wine Tasting, 3rd ed*. Academic Press.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.

Raschka, S., Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, Birmingham.