CS 559 (Machine Learning) Research Project; Spring 2019

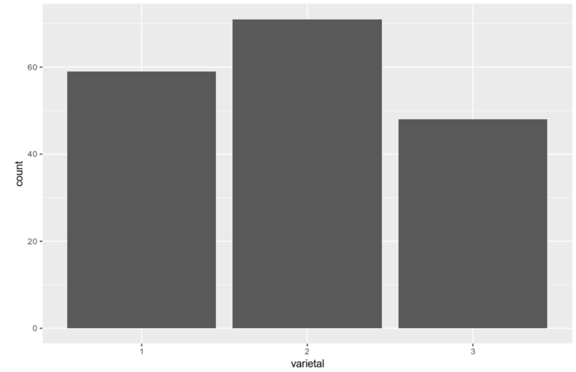
Patty Langasek

Abstract

Wine flavor can be very distinctive, and those who enjoy wine usually have a favorite grape they will select when choosing a specific wine. Though grapes can have a distinctive flavor, the quality of the wine is not entirely dependent on the specific grape. As the Jill Crecraft, the owner of Sip D’vine in Portland, OR will caution, “Don’t blame the grape.” I have evaluated 178 different wines (unfortunately, the data, not the wine itself), to determine if quantitative attributes can be used to identify the grape varietal. Overall, of the 3 varietals in the dataset, 1 was almost entirely isolated based on 2 attributes. Classifying the other two, however, proved more of a challenge.

Introduction

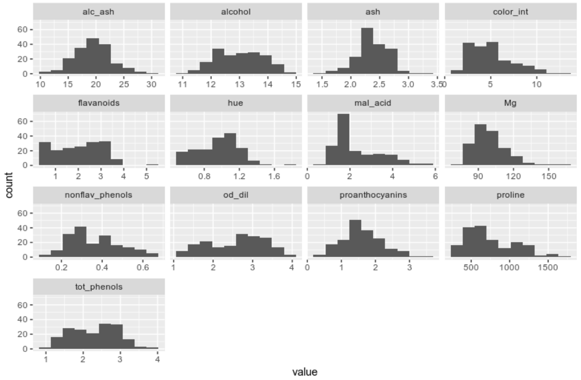
Data was collected from a Kaggle repository created by Bryn Humphreys 2 years ago (Version 1) at https://www.kaggle.com/brynja/wineuci. The dataset was created in July 1991 by Forina, M. et al, in Genova Italy, and last updated 21 September 1998, curated for classifier experimentation. The data are analyses of 3 cultivars in 1 regional wine-growing region in Italy. There are 13 attributes of 178 observations with a distribution of Class 1: 59, Class 2: 71, Class 3: 48 (Graph 1).



Graph 1: Varietal Counts

EDA

Initial evaluation of the data revealed generally normal(ish) distribution curves per attribute (Graph 2), with one interesting note of delineation between 2 varieties (specifically 1 and 3) in the flavanoid attribute (Graph 3). There were no missing values, and no obvious outliers in any of the attribute distributions. The only attribute showing clear distribution variation when separated by class was Flavanoids.



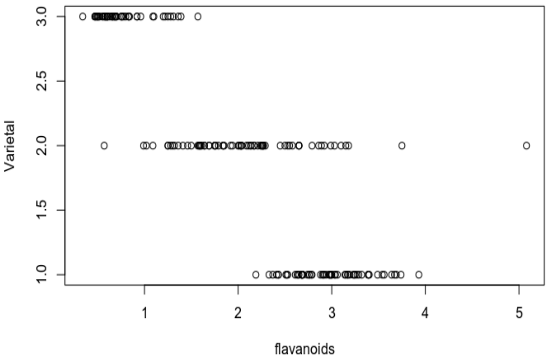
Graph 2: Distributions of Attributes

Methods

All evaluation and code was done in R Studio with R version 3.4.4 on x86_64-pc-linux-gnu. For classification, I chose to use decision trees, with the intention of building a random forest. Unfortunately, due to skill limitation, it ended up being a random grove instead.

The full dataset was randomly split into a training set with 138 observations, and a testing dataset with the remaining 40 observations was reserved for final prediction testing. The testing dataset remained blinded until the final testing procedure. The training dataset was then analyzed with 6 trials of cross-validation. The sub-training sets were 120 observations with the remaining 18 reserved for cv testing.

Due to discovery of the clear division between varietals 1 and 3 with the flavanoid attribute in the full dataset, prior knowledge was used to manually split the dataset along a flavanoid branch. This perfectly divided varietals 1 and 2 from 2 and 3, and there was a small gap where a few varietal 2 observations could be 100% purely classified, so they were classified manually as the first “leaf” of every tree root. Impurity for the remaining 2 branches was fairly consistent for each trial, around 30% impure on the 1st branch, and 20% on the 2nd branch.



Graph 3: Flavanoid Measurements by Varietal

Each branch was evaluated for the parameter that would most accurately divide the remaining 2 classes. I artificially filtered both branches based on the given (true) varietal class, then calculated the mean and standard deviation of each attribute. Using dnorm(), I calculated the probability that an observation's attribute value would correspond to each class in that specific branch based on the mean and standard deviations produced by those class attributes. I then compared each of the probabilities for every attribute and selected the highest performing attribute. For the CV cases, I calculated the accuracy with which each CV test branch performed, then selected the parameters that were most consistent in accuracy performance to test the final testing dataset. After classifying the final testing dataset, I evaluated the results using an accuracy calculation, an impurity calculation, and the confusion table below.

Results

The first branch perfectly and consistently divided classes 1 and 3, with each initial leaf scoring a 0 impurity score for variety 2. Cross validation on 6 trials indicated attributes alcohol and Magnesium perform well to separate varietals 1 and 2, with alcohol being most consistent with 80-83% accuracy, and alcalinity (sic) of ash was most consistent in accurately predicting varietals 2 or 3 between 90-93%. Hue also performed well in the trials, but not as consistently well as alcalinity of ash.

For the final testing set, I kept the 1st branch split with flavanoids, and selected alcohol to split 1 and 2, and alcanlinity of ash to split 2 and 3. impurity for the branch1 (with varietals 1 and 2) for varietal 1 was 0.2857, and for branch2 (with varietals 2 and 3) was 0.2667. The alcohol parameter performed exceptionally well in the final classification, with one leaf scoring an impurity of 0 for varietal 1, and the other leaf scoring an impurity of .1429 for varietal 2 (1 of the 3 leaves for varietal 2). Unfortunately, alcalinity of ash performed poorly for the first time in this dataset, with only 60% accuracy. The leaf for varietal 2 scored a 1 for impurity (!), miss-classifying both observations on that leaf, and varietal 3 scored .3077 for impurity on its leaf.

T	**Classification/Prediction**		
	1	2	3
1	14	1	0
2	0	10	4
3	0	2	9

Table 1: Confusion Matrix for Final Test

Discussion

Overall, classification tree could be better tuned with many more repetitions of the CV process. Unfortunately, my coding skills are limited in automating enough to the process to successfully create a forest that could be scanned for better attribute predictors, or better attribute cut-off-ranges to create more branches. I fear creating too many more branches with the limited number of observations, attributes, and classes, could result in over-fitting, that while it performs well on this dataset, would not be generalizable. It would be interesting to combine this with another classification method that could evaluate co-linearity between parameters. Clearly 1 and 2 are well separable, and 2 and 3 may have closer similarities that are difficult to distinguish with this method.

References

Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

Gordon, J. 13 Sept. 2017 [Internet] Accessed from: https://www.youtube.com/watch?v=LDRbO9a6XPU&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal&index=9&t=0s. Accessed 22 Aug 2019.

Humphreys, B., 2017 [Internet] Accessed from: https://www.kaggle.com/brynja/wineuci. Accessed 8 Aug 2019.