Machine Learning and Data Mining project: leaf analysis

Guglielmo Padula

AY 2021-2022 - Data Science and Scientific Computing

1 Problem statement

The problem is to identify a leaf based on some of its attributes. The structure of the problem is the following:

- It takes as input numerical vectors.
- The output of the problem consists in a categorial variable.

The structure of the data is the following:

- the numerical vectors consists in the leaf attributes (length, height, ecc.)
- the output categorical variable is the leaf name (a string)

2 Proposed solution

Based on the problem statement we want to solve a multiclassification problem. Specifically, we studied the performance of the following techniques:

- Bayes classifier
- K nearest neightbours classifier
- SVM classifier
- Random Forest classifier

3 Assessment and performance indexes

To assess the validity of the procedure I use two performance indexes:

- the missclassification error rate
- the AUC index

The AUC index is also considered so we can assess quality also in presence on unbalanced data. The AUC index is defined only for binay classification case. We apply it to multiclass classification in the following way: we measure the AUC for every single class taking it as positive (one vs all approach) and then we take the average.
These two indexes are estimated on both the training set and on the test set.

4 Experimental evaluation

4.1 Data

The full dataset that I used for studying the problem is avaiable here: `https://archive.ics.uci.edu/ml/machine-learning-databases/00288/`. It contains data about 40 different plant species, however only 30 species have numerical data. The data of the others will be discarded. There is also a feature regarding the implict ordering of the observations, which has no informative content, so it will be discarded. What remains of data is composed of $n = 340$ observations with $p = 14$ numerical features (the attributes of the leaf) and a categorial variable (the label of the leaf). The data is unbalanced.

4.2 Specific setup for each technique

4.2.1 Bayes classifier

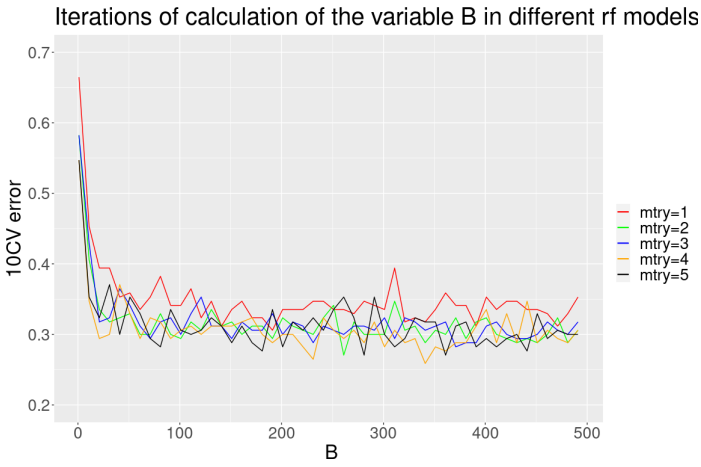
We to use the naive bayes classifier, because we can assume that the features are independent.

4.2.2 KNN classifier

About the Knn classifier we considered k implicit based on the minimization of the $10CV$ error on the training data.

4.3 Random Forest classifier

For the Random Forest classifier we tested different values of $mtry(1, 2, 3, 4, 5)$ and to the choice of B implicit based on the minimization of the $10CV$ error on the training data.
The following image shows the 10CV errov with respect to B and $mtry$.

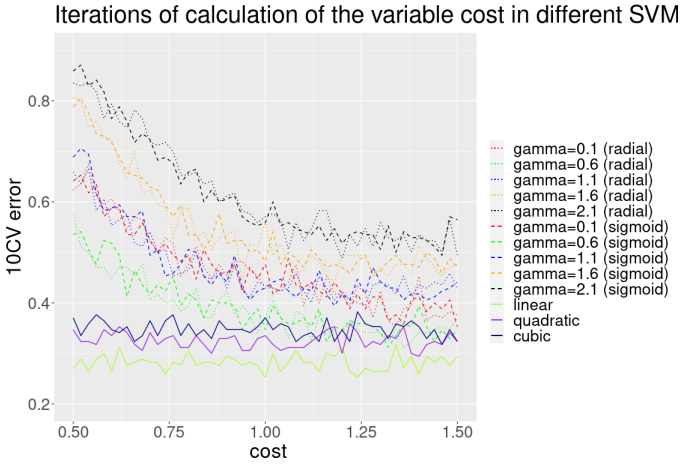


4.4 SVM

For the SVM classifier we tested multiple kernels with different parameters, with the choice of C based on the minimization of the $10CV$ error on the training data. The following combinations have been tested:

- Radial kernel with gamma=0.1,0.6,1.1,1.6,2.1
- Sigmoid kernel with gamma=0.1,0.6,1.1,1.6,2.1
- Linear kernel
- polynomial kernel with degree 2 or 3

The following image shows the 10CV errov with respect to C and the various combinations of gamma and kernel tested.



4.5 Results and discussion

We use a random classifier as a baseline. Follows a table of the performances of every model tested.

	training error rate	training AUC	test error rate	test AUC
RF (mtry=1)	0	1	0.23	0.89
RF (mtry=2)	0	1	0.24	0.88
RF (mtry=3)	0	1	0.24	0.87
RF (mtry=4)	0	1	0.25	0.87
RF(mtry=5)	0	1	0.25	0.86
Naive bayes	0.14	0.93	0.32	0.84
Knn	0	1	0.29	0.78
Radial svm (gamma=0.1)	0.23	0.88	0.31	0.83
Radial svm (gamma=0.6)	0.08	0.96	0.32	0.84
Radial svm (gamma=1.1)	0.03	0.98	0.39	0.79
Radial svm (gamma=1.6)	0.02	0.99	0.44	0.76
Radial svm (gamma=2.1)	0.02	0.99	0.49	0.73
Sigmoid svm (gamma=0.1)	0.59	0.68	0.69	0.63
Sigmoid svm (gamma=0.6)	0.86	0.54	0.86	0.54
Sigmoid svm (gamma=1.1)	0.88	0.53	0.89	0.53
Sigmoid svm (gamma=1.6)	0.87	0.54	0.89	0.53
Sigmoid svm (gamma=2.1)	0.87	0.54	0.89	0.53
Linear svm	0.11	0.94	0.27	0.71
Quadratic svm	0.0132	0.99	0.27	0.86
Cubic svm	0	1	0.27	0.85
Random classifier	0.97	0.50	0.95	0.51

Our ideal objective is to maximize the test AUC and minimize the test error, so the best classifier is RF with mtry=1, which significantly surpass the random classifier. About the radial and sigmoid svm classifier, with the increasing of gamma there is a behaviour of increasing overfitting.