Chikaze Mori
6 December 2020

A Comparison Of Naïve Bayes And Random Forest Applied To The Red Wine Dataset

Glossary

Accuracy	Percentage of correct predictions made by the model.
Attribute	Synonym for feature. In fairness, attributes often refer to characteristics pertaining to individuals.
Bias	A learner's tendency to consistently learn the same wrong thing.
Boosting	A machine learning technique that iteratively combines a set of simple and not very accurate classifiers into a classifier with high accuracy by up-weighting the examples that the model is currently misclassifying.
Categorical Variables	Variables with a discrete set of possible values. Can be ordinal (order matters) or nominal (order doesn't matter).
Class	One of a set of enumerated target values for a label.
Classification	Predicting a categorical output.
Confusion Matrix	Table that describes the performance of a classification model by grouping predictions into 4 categories.
Cross-validation	A mechanism for estimating how well a model will generalize to new data by testing the model against one or more non-overlapping data subsets withheld from the training set.
Decision Tree	A model represented as a sequence of branching statements.
Ensemble	A merger of the predictions of multiple models.
Feature	With respect to a dataset, a feature represents an attribute and value combination.
Hyperparameters	Higher-level properties of a model such as how fast it can learn (learning rate) or complexity of a model.
Label	In supervised learning, the "answer" or "result" portion of an example.

Model	A data structure that stores a representation of a dataset (weights and biases). Models are created/learned when you train an algorithm on a dataset.
Multi-class classification	Predicts one of multiple possible outcomes.
Naïve Bayes (NB)	A collection of classification algorithms based on Bayes Theorem which is an equation for calculating the probability that something is true if something potentially related to it is true.
Normalisation	Restriction of the values of weights in regression to avoid overfitting and improving computation speed.
Objective	A metric that your algorithm is trying to optimise.
Overfitting	Occurs when your model learns the training data too well and incorporates details and noise specific to your dataset.
Oversampling	A technique used to solve imbalance of data. It selects more samples from one class than another.
Parameters	Properties of training data learned by training a machine learning model or classifier. They are adjusted using optimisation algorithms and unique to each experiment.
Principal Component Analysis (PCA)	Simply looks at the direction with the most variance and then determines that as the first principal component.
Random Forest (RF)	An algorithm used for regression or classification that uses a collection of tree data structures.
Smoothing	A technique used to create an approximating function for categorical data. It allows Naïve Bayes to handle categorical data better.
Test Set	A set of observations used at the end of model training and validation to assess the predictive power of your model.
Training Set	A set of observations used to generate machine learning models.
Variance	How much a list of numbers varies from the mean (average) value.
Zero-frequency problem	A problem that Naïve Bayes predicts zero probability when a dataset has no occurrence of a class label.
References	https://developers.google.com/machine-learning/glossary https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html http://www.datascienceglossary.org

Intermediate Results

-Negative results-

Since the classes of the red wine dataset is imbalance (Table 1), it is reasonable to apply techniques such as SMOTE to the dataset. SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling techniques to solve imbalance of data. It selects examples that are close in the feature space, draw a line between the examples, and generate a new sample at a point along that line. We apply SMOTE to the dataset, and oversample the classes of 3 and 8 which its number in the dataset is small compared to others until its number reaches 50.

Table 2 shows the classification errors of each model in each phase that were trained and tested on the dataset applied SMOTE. It showed that SMOTE did not improve the performances, and even led significantly worse performances for both models, especially the NB model even though the number of oversampled data was small. It caused an overfitting problem in the RF model, and it is worse than the one without SMOTE.

Number of Classes

Class	5	6	7	4	3	8
Numbers	681	638	199	53	18	10

Table 1

Classification error

	NB	RF
Train	57.5%	34.6%
Validation	57.4%	34.0%
Test	56.8%	42.2%

Table 2

Implementation Details

-Brief description of main implementation choices-

For the convenience, all the data preprocessing process was managed in Python. Among the variables generated in the process of PCA, only 5 of them were selected for the new dataset as they explained more than 99% of the original 11 attributes. During the data preprocessing process in Python, the dataset was normalised in a range between -1 to 1 to bring all the attributes to the same range and improve data integrity.

The MATLAB functions used for training the models are listed below:

- fitcnb (NB)
- fitcensemble (RF)

For RF models, the name value pair ‘Reproducible’ is set to be true for the reproducibility. However, fitcensemble function for NB does not have this name value pair as an option. Hence random numbers are controlled by using rng function.

All the hyperparameters used for training the best models were directly extracted from the results.