A Comparison of Naïve Bayes and Random Forest on airline customer satisfaction prediction

Yuichi Kuriyama

Glossary

Term	Definition
Accuracy	Percentage of the correct prediction from all of predictions
Bays Theorem	Mathematical equation to calculate the conditional probability given prior information
Classification	Machine learning task to categorise the label correctly
Confusion Matrix	N×N table to summerise the correct label and prediction to see the general tendency
Correlation	Type of measurement to see how one data is related to the other in the range of -1 to 1
Cross Validation	Validation method to iterate the training and validation data to see the generalisation to the unseen data
Decision tree	Non-parametric machine learning method by creating a series of branch.
Ensemble	Technique by combining multiple algorithms to create a better result
F1 Score	Scoring metrics based upon the weighted average of precision and recall

Grid-Search	Hyper parameter optimisation technique by conduct a complete search over a subset of the hyperparameters space
Hyper Parameter	Parameter to control the learning process of the model
Kernel density estimation	Non-parametric way to estimate the probabilistic density
Naïve Bayes	Algorithm to compute a posteriori for each observation and output the most probable one as a result given a prior information
Overfitting	Concept that statistical model fits into the training data whereas shows bad performance against unseen data
Precision	The number of true positive divided by the number of positive predictions
Random Forest	Ensemble learning method by constructing several decision trees
Recall	The number of true positives divided by the number of true positives and the number of false negatives.
Regression	Statistical modelling task to predict a continuous value based upon features.
Underfitting	Similar to the concept of overfitting but also shows bad performance to the training data
Zero-frequency Problem	The phenomena that If some information in the test dataset is not appearing in the train data, Naïve Bayes classifier will output a zero probability for that feature in the test data

Intermediate results including any negative results

I. Computational capacity

Since the original training data set was huge (103594 rows), creating the machine learning model with original data takes a huge amount of time. For instance, in the hyper-parameter optimisation phase, it roughly takes more than 3 hours to compute. Hence, dataset used for the poster is modified by decreasing the dataset from 103594 to 5000 rows. This modification allows to manageably implement each machine learning technique within the time frame. Although this might seem to be a huge data loss, the result I obtained (94.6%) was similar to the level where previous studied (95.4%). That final result shows that this model holds validity and reliability for this study.

II. Reproducibility

In the initial modelling process, the output was not exactly the same for each trial although hyper parameter was the same. This is because both Naïve Bayes and Random Forest model did not specify the random state. In order to tackle this, random seed were set, resulting in producing the same output for every iteration.

III. No improvement for Naïve Bayes after hyper parameter tuning

Figure 1 shows the confusion matrix for Naïve Bayes classifier for training dataset.
In the comparison to the Naïve Bayes before hyper parameter tuning, the model with the most optimised parameter did not improve the score and produce the same prediction. One assumption from critical evaluation is that independent variables are mostly normally distributed ordinal variables, leading to the result that normal distribution for hyper parameter (default) was the best result.

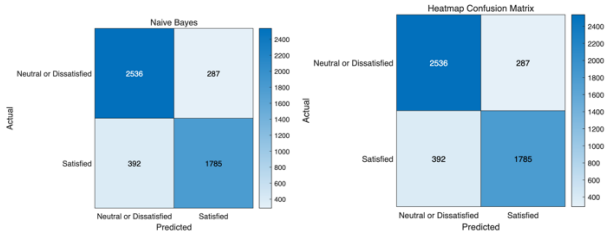

Figure 1

Implementation details including a brief description of main implementation choices

For the sake of visualisation and convenience for exploratory data analysis, pre-processing was done by Python. In the implementation stage, the focus for this study is look at how Naïve Bayes and Random Forest model have impacts on the result and the implication. Furthermore, by implementing 10-fold cross-validation, this study also attempts to examine the effect for hyper-parameter by grid-search. These methodologies enable to provide a sound understanding of each algorithm and critically evaluate two machine learning models.

Each implementation details including MATLAB function is as follows:

Naïve Bayes

- Fictnb

Random Forest

- Template Tree
- Fictcensumble
- Surf (Three-dimensional visualisation for learning process of hyper parameter against objective)

Confusionmat and perfcurve functions were also applied in both models to examine the general tendency to the predictive model.