Sin Sze Li 210035720

12 December 2021

INM 430 Machine Learning
Final coursework

A comparison of Decision Tree and k-Nearest Neighbours on Classification of Foetal Health

Supplementary material

I. Glossary

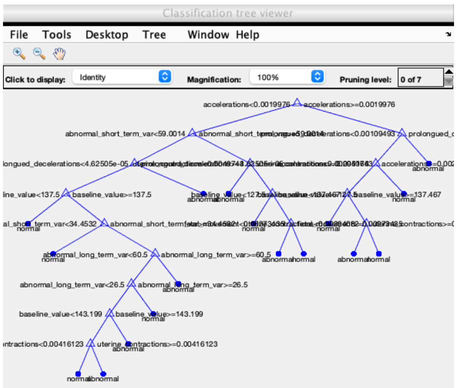
Accuracy	Total number of corrected classified samples out of all classified sample.
AUC	Area under receive operating characteristic (ROC).
Bayesian optimization	An approach that uses Bayes Theorem to direct the search in order to find the optimal hyperparameters of a model.
Bias	Simplifying assumptions made by a model to make the target variable easier to predict.
Cardiotocogram	A technique to monitor the heartbeat of a foetus and the uterine contractions during pregnancy and labour.
CART algorithm	Classification And Regression Tree algorithm.
Classes	The output variable of a sample after being passed to a classification algorithm.
Classification	An algorithm to predict a distinct outcome.
Computation cost	The execution time per time step during a simulation.
Confusion matrix	A table that shows the performance of a classification model.
Correlation heatmap	A 2D matrix to represent correlation between different variables.
10-fold Cross validation	In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained for testing the model and the remaining k-1 subsamples are used as training data. The training data is used to train the model and the test set is used to evaluate it.
Decision tree	A machine learning model for classification and regression. To classify a sample, the sample is passed from the root node, assessed by its feature, and passed to another node via one of the branches. The end node determines the class of the sample.
Exhaustive search algorithm	Algorithm that searches for all possible combination of parameters.
F1 score	Harmonic mean of precision and recall. $$F1\ score = 2\ \times \frac{Precision \times Recall}{Precision + Recall}$$
False negative rate	Also called miss rate. It is the measure of incorrectly identified positive out of all actual positive samples. $$False\ negative\ rate = \frac{False\ negative\ (FN)}{True\ positive\ (TP) + False\ negative\ (FN)}$$
False positive rate	Also called fall-out. It is the measure of incorrectly identified negative out of all actual negative samples.

	$False\ positive\ rate = \frac{False\ positive (FP)}{False\ positive\ (FP) + True\ negative\ (TN)}$
Feature selection	A process to reduce the number of input variable when training a model.
Gini's Diversity Index	Also known as Gini impurity, Gini's (Diversity) Index calculates the probability of a particular feature classified incorrectly when it is randomly chosen. It ranges from 0 to 1. When building a decision tree, the feature with least Gini Index is chosen as the root node.
Hyperparameters	Parameters of a model that controls its learning process.
k-Nearest Neighbours	A machine learning model for classification and regression which searches for similar datapoints of a particular sample to be predicted.
Node	A point of a decision tree that branches lead to.
Noise	Additional meaningless information of data.
Nonparametric	If a model is nonparametric, no assumption about the distribution of data is made.
Normalization	Rescaling the data such that all data resembles a normal distribution of mean 0 and standard deviation 1.
Outliers	Datapoints that differs significantly from other observations.
Overfitting	Occurs when the model fits exactly on the training data. If a model overfits, it cannot perform accurately on the test data.
Precision	The measure of correctly identified positive out of all identified positive samples. $Precision = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ positive\ (FP)}$
Precision-Recall (PR) curve	A graphical plot that illustrates the tradeoff between precision and recall for different threshold.
Recall	Also called sensitivity or true positive rate. It is the measure of correctly identified positive out of all actual positive samples. $Recall = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ negative\ (FN)}$
Regression	An algorithm to predict a continuous outcome.
ROC curve	Receiver operating characteristic – a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied.
Scaling	Data transformation technique such that the data fits within a specific scale.
Stratification	Dividing a dataset into subgroups (strata) such that each strata contains the same proportion of each class label. It is used to guarantee that the test set is representative of the entire dataset.
Strip plots	Graphical data analysis technique to summarize a univariate data set.
Supervised learning method	A method that uses datasets with target features (classes) to train algorithms that can classify or predict outcomes.

Synthetic Minority Oversampling Technique (SMOTE)	A tope of data augmentation for the minority class. New data are synthesized from the existing data such that the total number of data for each class is the same.
Test set	Subset of data that is used to test the model for generalisation.
Training set	Subset of data that is used to train a model.
Variance	Amount that the estimate of class will change if different training data was used.
References	https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8 https://patient.info/pregnancy/cardiotocography https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/ https://medium.com/analytics-vidhya/stratified-sampling-in-machine-learning-f5112b5b9cfe https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

II. Intermediate results including any negative results

- Illustration of decision tree



Classification tree viewer

Click to display: Identity

Magnification: 100%

Pruning level: 0 of 7

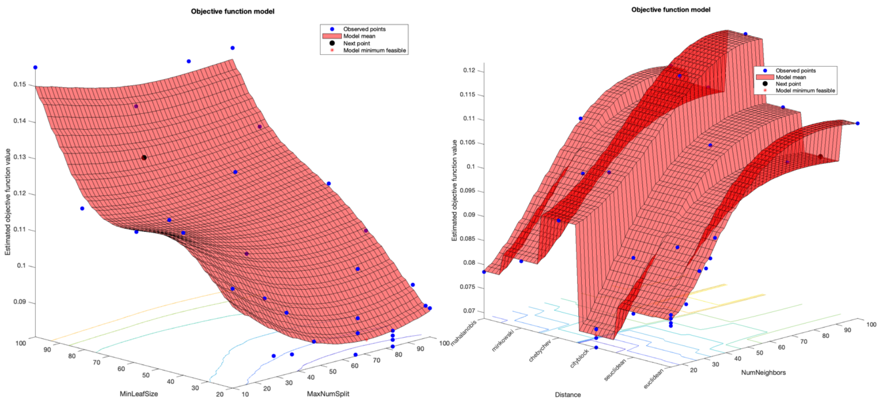
File
Tools
Desktop
Tree
Window
Help
accelerations<0.0019976
accelerations>=0.0019976
abnormal_short_term_var<59.0014
prolongued_decelerations<0.00109493
prolongued_d
longued_decelerations<4.62505e-05
accelerations>=0.002
line_value<137.5
baseline_value>=137.5
baseline_value>=137.467
al_short_term_var<34.4532
abnormal_long_term_var<60.5
abnormal_long_term_var>=60.5
abnormal_long_term_var<26.5
abnormal_long_term_var>=26.5
baseline_value<143.199
baseline_value>=143.199
ntractions<0.00416123
uterine_contractions>=0.00416123
normal
abnormal

- Performance of models on training test

	DT	kNN
Accuracy	0.935	0.940
Recall	0.938	0.973
Precision	0.932	0.912
F1 score	0.935	0.942
AUC	0.965	0.990

The performance of kNN is much better on the training set than on test set. This is because SMOTE was applied to training set but not on test set, so there is no class imbalance in the training set.

- Bayesian optimization for decision tree (left) and k-Nearest Neighbour (right)



III. Implementation details including a brief description of main implementation choices

All initial data analysis including the strip plots and correlation matrix was done in Python for convenience. The dataset was split into 70% training set and 30% test set by scikit-learn. SMOTE was done using imbalanced-learn.

The main MATLAB functions used were:

- fitctree()	To train decision tree
- fitcknn()	To train k-Nearest Neighbours
- bayesopt()	To optimize hyperparameters using Bayesian optimization
- normalize()	To rescale the numerical features into the range 0 to 1
- confusionmat()	To compute confusion matrix
- confusionchart()	To create confusion matrix chart
- perfcurve()	To create ROC

When implementing Bayesian optimization, a function to calculate F1 score was used. Details of the function can be found here:
https://uk.mathworks.com/matlabcentral/fileexchange/71000-to-optimise-hyperparameter-of-ml-model-using-f1?s_tid=prof_contriblnk