Assignment 1 Report

SE755

Bayesian Machine Learning

Hualong Zhu

27/08/2018

Assignment 1 Report

Hualong Zhu
hzhu828 SE755
AucklandUni, Auckland, New Zealand
hzhu828@aucklanduni.ac.nz

I. Introduction

The report mainly introduces the machine learning modeling and testing results of four distinct data sets, including World Cup 2018 data, SH1 traffic volume data, occupancy sensor data, and landsat satellite data. Each type of data has gone through necessary processes to be used to train different machine learning models based on different machine learning algorithms. In addition, in order to evaluate the accuracy of predictions , K-folds, which is a cross-validation method, has been used in the model training and testing process of many models. Grid search as a standard hyperparameter tuning method has been used to find out the suitable hyperparameters of different ML models for different data sets. (* All the analysis code is written by python 2 and can be opened and edited by the Jupyter Notebook.)

II. Data Processing

Features Processing

In order to train the data, the data must be processed firstly to meet the requirements of different ML algorithms, which means that data could be transformed to more value and improve the predictive accuracy of ML models, which is also called *features engineering*.

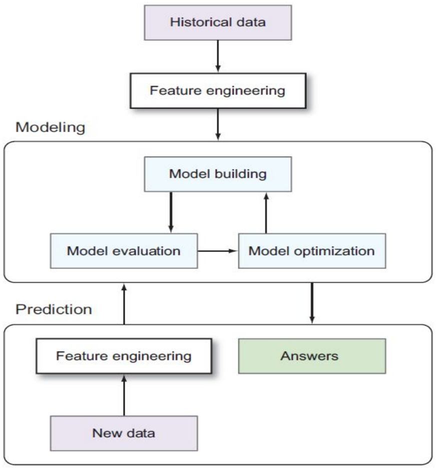

Figure 1 The workflow of model training

Unrelated Data

Before training the data, it is better to figure out and drop the unrelated data. Not only it can reduce the work of calculating, but also decrease the noise of training process, which means it may increase the accuracy of prediction. For example, In the data set of World Cup 2018, obviously the feature "Date" is unrelated to the prediction result, so we should remove its data from the training and testing data, which can save the data processing time and increase the performance of the model. In addition, the feature importance analysis can also help to choose useful features. For example, among the features of occupancy sensor data set, table 1 shows five features and their importance respectively in a decision trees algorithm, it is easy to see that the feature "Temperature" and the feature "HumidityRatio" are not important for the decision tree model. However, the feature importance could be different for different ML algorithms, for instances, the table 2 shows that the feature importance of "Humidity" and "HumidityRatio" is not zero any more.

Data Normalization & Standardization

Some algorithms, such as Linear regression, Logistic regression, and SVM, require data to be normalized or standardized, which means the features should be manipulated to reside on the same numeric scale. Normally, the data should be normalized to be in the range from 0 to 1, or standardized from 1 to -1. In this report, all the related features in the four data sets except the target feature have been normalized in the range from 0 to 1.

III. Model Training and Testing

Machine Learning Algorithms

Machine learning algorithms are distinguished from rule-based systems in which they can create their own models based on data. Supervised ML systems generalized by learning form the features of examples with known results. Table 3 shows the ML algorithms which have been used in this report.

Evaluation Metrics

Evaluation Metrics can be seen as performance indicators of a training model, which are used to evaluate the performance of the training model. Different ML algorithms may use different evaluation metrics based on the types of the task of data set.

1. Mean squared error (MSE), which is a standard metric for evaluation for regression machine algorithms, is the average squared difference between the model-predicted value and the true value of the target variable. The smaller the value, the better the ML algorithm is. Available in Data set: World Cup 2018 and SH1 traffic volume.
2. R2 score, which is a regression metric, 1.0 is the best. Available in Data set: World Cup 2018 and SH1 traffic volume.
3. F1 score, a classification metric, which can be seen as a weighted average of the precision and recall, and 1 is the best value, and 0 is the worst. Also, the relative contribution of precision and recall to the F1 score are equal. In the multi-class and multi-label case, this is the weighted average of the F1 score of each class. Available in Data set: World Cup 2018 and landsat satellite.
4. Accuracy classification score, a classification metric, which computes subset accuracy in multilabel classification. The best value is 100%, the worst value is 0%. Available in Data set: Occupancy sensor, World Cup 2018 and landsat satellite.

5. Area Under ROC, which computes area under the Receiver Operating Characteristic Curve from prediction scores, is a classification metric which only supports binary classification task or multilabel classification task. Available in Data set: Occupancy sensor.

Cross-validation

In order to evaluate the performance of model on new data, cross-validation has been done to make the user confident about the accuracy of the new model when it was used to make predictions on new data.

One method is commonly used is k-fold cross-validation. K-fold cross-validation randomly splits the data into k subsets which is also called folders. For each subset, a model is trained on all the data except the data from that subset and made predictions for the data from that subset.

As the figure 2 shows, after all the subsets are cycled through, the predictions for each subset are aggregated to compared to the true target value to assess accuracy of the predictions.

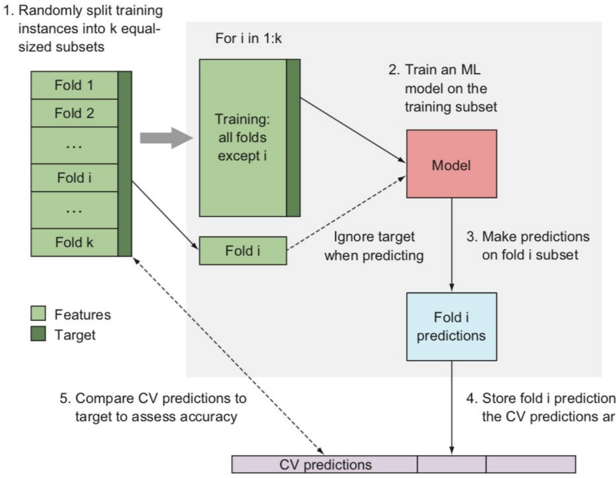
1. Randomly split training instances into k equal-sized subsets
For i in 1:k
Fold 1
Fold 2
...
Fold i
...
Fold k
Training: all folds except i
2. Train an ML model on the training subset
Model
Fold i
Ignore target when predicting
3. Make predictions on fold i subset
Fold i predictions
Features
Target
5. Compare CV predictions to target to assess accuracy
4. Store fold i predictions in the CV predictions array
CV predictions

Figure 2 K-fold cross-validation

In the assignment, all the models generated by the algorithms use the k-fold cross-validation to evaluate the accuracy of predictions, k is 3 as default.

Hyperparameters Tuning - Grid search

The standard way to tune parameters for an machine model is through grid search. The steps are shown below:

1. Choose a evaluation metric.
2. Choose the machine learning algorithm.
3. Select the parameter should be tuned
4. Define the grid as the Cartesian product between the arrays of each tuning parameter.
5. The tuning parameters should pass cross-validation.
6. Finally, select the best set of tuning parameters corresponding to the largest value of the evaluation metric.

IV. Model and Algorithm Analysis

Data Set: World Cup 2018

For regression tasks, the score attribute is the target, table 5 shows the performance. Linear regression and Right as the machine learning algorithms have been used to train the model. Ridge behaves better because it has smaller MSE and better R2 score which is nearer to 1.

Table 5 The Performance of Regression Task

Algorithms	MSE	R2 Score
Linear regression	5.29	-1.57

Ridge	3.98	-0.93

For classification tasks, the status attribute is the target, which table 6 shows the performance. Decision tree, random forest, naive bayes, perceptron, K-nearest neighbors and SVC are used to train the model. Naive bayes is is the most suitable ML algorithm among them.

Table 6 The Performance of Classification Task

Algorithms	**Accuracy**	**F1 Score**	**Hyperparameter Tuning Results**

Decision tree	53.85%	0.45	max_leaf_nodes=2,
			min_samples_split=2

Random Forest	38.46%	0.36	max_leaf_nodes=8,
			min_samples_split=4

Naive Bayes	61.54%	0.51	-
Perceptron	23.08%	0.09	-

K-nearest neighbors	15.38%	0.08	leaf_size=10,
			n_neighbors=7

SVC	23.08%	0.10	C=1.0

Data Set: SH1 traffic volume

This is a regression task for the data set. So linear regression, Ridge, and Decision tree have been used to train the model. The performance of Linear regression and Ridge is nearly the same with each other.

Table 7 The Performance of Regression Task

Algorithms	**MSE**	**R2 Score**
Linear regression	843.81	0.98
Ridge	834.91	0.98
Decision tree	2018.72	0.94

Data Set: Occupancy sensor

The "Occupancy" attribute is the target attribute. This is a binary classification problem which can use Area Under ROC as the evaluation metric, so that classification ML learning algorithms should be used to train the model. Table 8 shows the training results.

Table 8 The Performance of Classification Task

Algorithms	**Accuracy**	**Area Under ROC**	**Hyperparameter Tuning Results**

Decision tree	99.56%	1.00	max_leaf_nodes=4,
			min_samples_split=2

Random Forest	99.56%	0.99	max_leaf_nodes=7,
			min_samples_split=2

Naive Bayes	97.14%	0.99	-
Perceptron	99.65%	0.99	-

K-nearest neighbors	95.30%	0.98	leaf_size=10,
			n_neighbors=9

SVC	95.06%	0.98	C=1.0

Data Set: landsat satellite

This is a multi-class classification problem which cannot use Area Under ROC, instead F1 score is used to evaluate the model.

Table 9 The Performance of Classification Task

Algorithms	**Accuracy**	**F1 Score**	**Hyperparameter Tuning Results**

Decision tree	82.42%	0.82	max_leaf_nodes=51,
			min_samples_split=2

Random Forest	87.00%	0.87	max_leaf_nodes=89,
			min_samples_split=2

Naive Bayes	61.83%	0.60	-
Perceptron	52.00%	0.48	-

K-nearest neighbors	88.33%	0.88	leaf_size=10,
			n_neighbors=8

SVC	88.08%	0.88	C=1.0

V. Conclusion

The report mainly shows the modeling workflow of machine learning based on four different data sets by using different machine learning algorithms. The selections of machine learning algorithms and evaluation metrics could be different based on the type of the taske and the type of target features. Totally, there is no ML algorithm which can suit any kind of data sets. The ML algorithms should be selected based on the concrete issues.

Appendix

Source code address:

https://github.com/Long315/755_Assignment1

Table 1 Feature Importance Table By Using Decision Trees

	Feature	Importance
0	Light	0.995019
1	CO2	0.003514
2	Humidity	0.001467
3	Temperature	0.000000
4	HumidityRatio	0.000000

Table 2 Feature Importance Table By Using Random Forest

	Feature	Importance
0	Light	0.620348
1	Temperature	0.196150
2	CO2	0.166454
3	Humidity	0.011278
4	HumidityRatio	0.005770

*Table 3 Machine Learning Algorithms Used In This Report

ML Algorithm Name	Type	Linear or Nonlinear	Data Set Applying
Linear regression	Regression	L	World Cup 2018, SH1 traffic volume
Ridge	R	L	World Cup 2018, SH1 traffic volume
SVC	Classification	L	World Cup 2018, Landsat, Occupancy sensor
K-nearest neighbors	C/R	N	World Cup 2018, Occupancy sensor, Landsat
Decision trees	C/R	N	World Cup 2018, SH1 traffic volume, Occupancy sensor, Landsat
Random forest	C/R	N	World Cup 2018, Occupancy sensor, Landsat
Perceptron	C	N	World Cup 2018, Occupancy sensor, Landsat
Naive Bayes	C	N	World Cup 2018, Occupancy sensor, Landsat
...	...	...	...