Protein Report

by Abdul Rehman

Submission date: 19-Sep-2022 06:22PM (UTC+0500)
Submission ID: 1903550272
File name: Assessment_Task_3.docx (113.13K)
Word count: 1591
Character count: 8395

Assessment Task 3: Problem Solving Task

Q1: What are the differences between hyperparameter and parameter of a machine learning (ML) model. Explain your answer using at least two machine learning models that you have learned in this unit.

Answer: Model parameters are internal to the models that are usually used to configured the machine learning models. The parameters of the model are adjusted during the training of the model internally and did not allow to set these parameters to set externally. The weights and coefficient in linear regression model and centroid the K Nearest Neighbor are the examples of machine learning parameters.

While the hyper-parameters are external to the machine learning models and they are specified by the users externally. The hyper parameters are manually set by the machine learning expert base on the different scenarios like dataset dimension and complexity. Kernel type in SVM, and the maximum-depth in decision tree and random forest model are the examples of hyperparameters in machine learning.

Q2: Prove that Elastic net can be used as either LASSO or Ridge regulariser.

Answer:

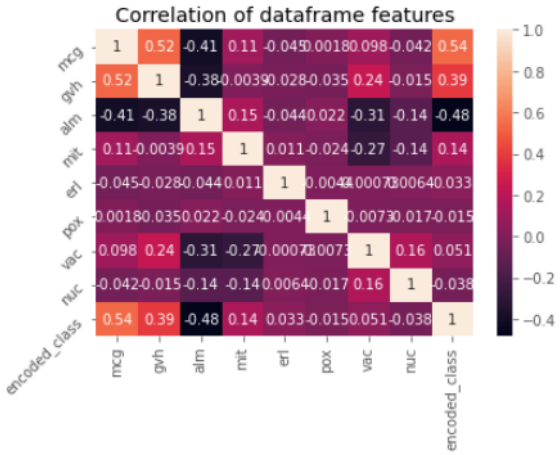
- Lasso uses the L1 regularization and Ridge uses the L2 regularization.
- Lasso set the irrelevant value to 0 while Ridge set them to lower.
- Lasso remove irrelevant features while Ridge minimizes their impact.
- Elastic net is the weighted combination of L1 and L2. As it is the combination of both, it behaves like the Lasso and Ridge.
- Elastic net can be used as Lasso by only changing the L1_wt hyperparameter value.
- It decided how much weight goes to L1 and how much goes to L2.
- If the L1-wt is zero then the regularization act as pure Ridge and if the L1_wt is one then the regularization is pure Lasso.

Q3: Analyze the importance of the features for predicting presence or absence of protein using two different approaches. Explain the similarity/difference between outcomes.

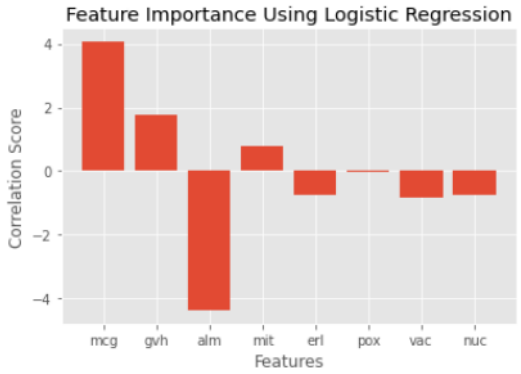
Answer: For analyzing the importance of the dataset features we used correlation and ranking based approach. In the correlation technique, we find the correlation score of features with each other and lastly target the correlation score of input features with the target variable. The range of the correlation score is from -1 to 1. The absolute value nearer to 1 show the strong relationship between variables while the absolute value nearer to the 0 show the weak correlation. The sign of the score shows the direction of relatedness: if the sign is positive then the both variables are directly proportional and they will increase or decrease together and if the coefficient of correlation score is negative then the both variable increase or decrease inversely.

The correlation plot of our independent features with target variable is shown in below figure. It showed the relatedness of only first three features because these features have the 0.54, 0.39 and -

0.48 correlation score respectively. While the rest of the features did not show the any strong relationship because their correlation scores are very nearer to 0.



In the second approach, we used the feature ranking algorithm by using the logistic regression machine learning model. Logistic regression model train on the features of the dataset and after that it return the importance scores of each feature for predicting class. The importance score of all independent features is also shown in below bar chat. The below figure showed the reasonable importance for all features except the 'prox' feature. Collectively, by analyzing the both technique of feature ranking, 'prox' is the less related feature and commonly predicted by the both techniques.



Q4: Create three supervised machine learning (ML) models except any ensemble approach for predicting presence or absence of protein.

a) Report performance score using a suitable metric. Is it possible that the presented result is an overfitted one? Justify.

Answer: We used the linear regression, logistic regression and SVM machine learning model with linear kernel for predicting the presence of protein. The proposed dataset for this task was

converted into class balance dataset and the accuracy is best measure for class balance dataset. Here we measure the performance of the trained model on the basis of accuracy score. We got the 0.90%, 0.88% and 0.89 test accuracy score for linear regression, KNN and decision tree respectively. The confusion metrices of the trained model are shown in below Figure for better understanding.
As all the models were trained with few features and there were the chances of model overfitting. The dataset was also small that tend the model towards under fitting. But we train the model on training samples and then calculate the accuracy for training and testing samples individually. We did not get the significant difference in training and testing accuracy that indicate that models are properly trained and did not tend towards the overfitting.

b) Justify different design decisions for each ML model used to answer this question.

Answer: For the prediction of protein presence, we used the linear regression, logistic regression, SVM, KNN and Naïve Bayes. We used all the methods by using the hit and trail technique. By following the training process of all model, we evaluate the performance of each model by calculating the accuracy score. We got the highest accuracies with liner regression, KNN, and decision tree. Resultantly, we selected the best performing design decision for this query.

c) Have you optimized any hyper-parameters for each ML model? What are they? Why have you done that? Explain.

Answer: We perform hyperparameter tuning for KNN and DT model. We tuner the n-neighbor hyperparameter of KNN and kernel parameter of DT. We got the best accuracy with linear kernel and 5 n-neighbors.

d) Finally, make a recommendation based on the reported results and justify it.

Answer: K Nearest Neighbor (KNN) is the most robust and efficient model for the prediction of protein presence. We got the highest accuracy score for KNN model compared to the other trained machine learning models.

Q5: Build three ensemble models for predicting presence or absence of protein

a) When do you want to use ensemble models over other ML models?

Answer: Machine learning models usually faces the following challenges:

- Machine Learning models are very sensitive to the variance in input features. They did not provide the good result in presence of high variance input data.
- Machine learning models are also relied on few features for making prediction.

We used the ensemble models in the following situations:

- When the input features contained the values with high variance, we used the ensemble methods to overcome the sensitivity of input data.

- When the objective to make prediction by using the all features rather then the very few features, ensemble method is best approach in this situation.

b) What are the similarities or differences between these models?

Answer: The similarities and differences of machine learning models and ensemble methods are following

- All the ensemble base methods are learning based algorithms.
- All models can be used for binary and multiclass classification.
- We used the Random Forest, AdaBoost and Gradient Boost ensemble learning based models for binary classification. Random Forest make multiple trees and extract the prediction from each tree. Lastly it makes the final prediction based on the majority votes. AdaBoost ensemble method used the boosting technique that mainly aim to combine the numerous weak learning model to form a strong model. Gradient tree also aims to form strong model by combining weak model. It uses the decision tree behind the boosting to form a strong method.
- All of these models are very helpful for complex dataset and used to increase the performance of the model.

c) Is there any preferable scenario for using any specific model among the set of ensemble models?

Answer: Random Forest model is very efficient model for high dimensional and sparse data. When the dataset based on very large number of features, Random Forest model perform well among the other ensemble learning based model.

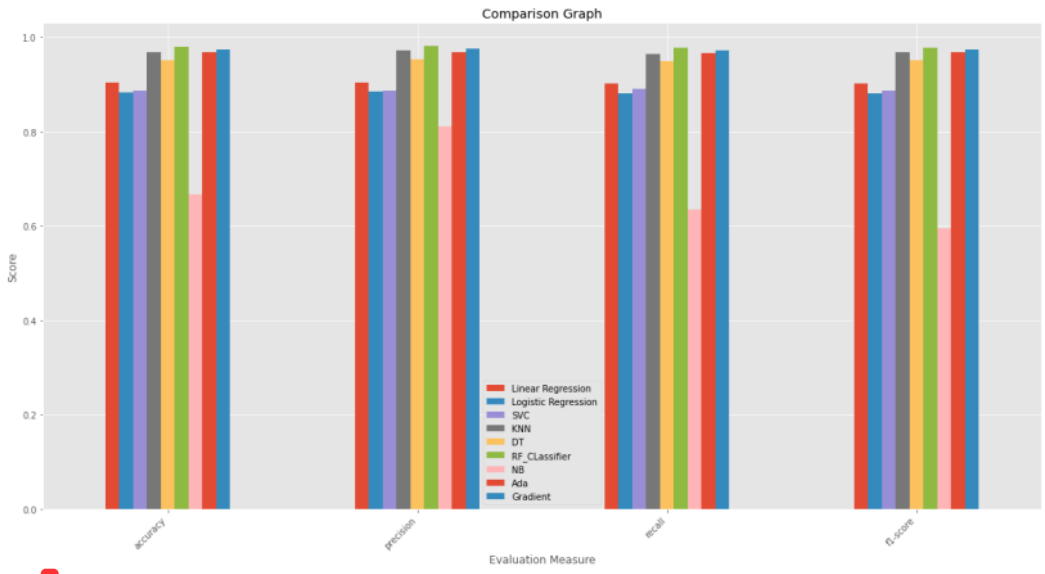
When the size of dataset is very small or dataset is base on few features, normally gradient boosting technique show the efficient results.

d) Write a report comparing performances of models built in question 5 and 6. Report the best method based on model complexity and performance.

Answer: We trained different machine learning models on the training set and evaluate the performance of each model using test samples. The compare the performance of the trained model base on accuracy score. We also measure the precision, recall and f1-score of the trained models. We got the 0.97% highest accuracy with gradient tree ensemble learning based model compared to the other model. The below table showed complete comparison of all trained models.

Metrics	Linear Regression	Logistic Regression	SVC	KNN	Decision Tree	Random Forest	Naïve Bayes	Ada Boost	Gradient Boosting
accuracy	0.9032	0.8817	0.8870	0.9677	0.9516	0.9784	0.6666	0.9677	0.9731
precision	0.9032	0.8836	0.8866	0.9719	0.9528	0.9809	0.8098	0.9685	0.9746

recall	0.9015	0.8809	0.8895	0.9647	0.9498	0.9764	0.6352	0.9665	0.9715
f1-score	0.9023	0.8813	0.8868	0.9672	0.9511	0.9782	0.5955	0.9674	0.9728



e) Is it possible to build ensemble model using ML classifiers other than decision tree? If yes, then explain with an example.

Answer: Yes, it is possible to build the ensemble model using the ML classifiers. Voting classifier model of ensemble learning is the example of building the ensemble model with machine learning models other than decision trees. Voting Classifier normally based on numerous different kinds of machine learning models. It takes the predictions from each model and make the final prediction based on majority votes. Voting classifier can combine any type model either machine learning or either ensemble learning. The below code snippet show the voting classifier

Protein Report

ORIGINALITY REPORT

12%	11%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

11%

★ myassignmenthelp.com

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Protein Report

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5