Cmput 466 Mini-Project Report

Yulong Zhou

Introduction

Parkinson's disease (PD), also known as tremor palsy, is a common neurodegenerative disease in middle-aged and elderly people. In this report, we are going to building machine learning models to can be used to differentiate healthy people from people who having Parkinson’s disease. We used 4 different machine learning models namely, Zero Rules, Logistic Regression, KNN and SVM. We selected hyperparameters for each model (except the zero-rules) by performing cross-validation, and found the best hyperparameter settings to achieve the highest cross-validation accuracy. Based on the best number we found, we refitted the model on the training set and get the accuracy on the test set.

Dataset Information

This dataset is about the Parkinson’s disease classification, which contains 195 records of people with 23 different attributes. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column. If the patient has the Parkinson’s disease, his or her “status” which indicates the health status of the people will be 1, and if the patient does not have the Parkinson’s disease, then the “status” will be 0, which means he or she is healthy.

Citation: Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear)

Models

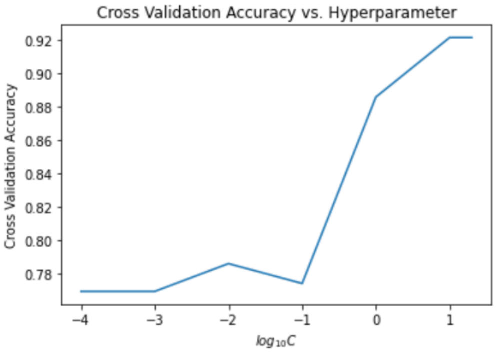
1. **Zero Rule:**

   Zero Rule is the benchmark procedure for classification algorithms whose output is simply the most frequently occurring classification in a set of data. In the dataset, the number of positive samples is larger than that of negatives, so we predict 1 for all instances. There is no hypeparameters in Zero Rule and the resulting test accuracy is 0.6307692307692307.

2. **Logistic Regression:**

   The Logistic regression uses the input feature vector x_i, weighted by the parameter w and a bias b to compute the linear formula: $z_i = w^T x_i + b$, then the value of z is used in the sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$, which gives an arbitrary value between 0 and 1. In the code, we tuned the regularization term C and evaluate the performance by cross validation accuracy.

We train the model using LogisticRegression class of sklearn.linear_model with possible values of C are shown with a list of numbers: [0.0001, 0.001, 0.01, 0.1, 1, 10, 20]. By iterations, we get the best hyperparameter in this list is 10, and we use the best hyperparameter to refit the model with the test set. At the end, it shows that the Logistic Regression test accuracy is 0.7076923076923077, which is higher than the Zero Rule.

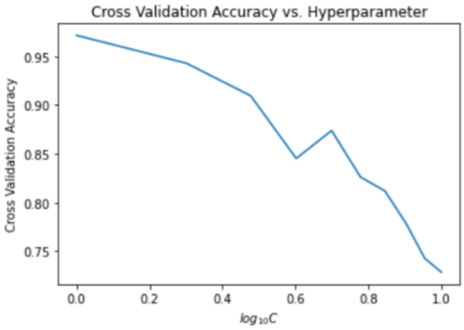


This plot shows the relation between the test accuracy and the $\log_{10}$(C).

3. **KNN：**

The K-NN model finds the similarities between the new data set (input) to the training data set. In our task, the model will put the new data point in either "status" = 1 or "status" = 0 category based on the most similar features. KNN chooses the nearest data points which is the value of K, and then calculate the distance of K number of neighbors.

In our experiment, we tuned the value of n_neighbors and evaluate the performance by cross validation accuracy, and we also train the model using KNeighborsClassifier class of sklearn.neighbors with the possible values of C (the n_neighbors) which is shown with a list of numbers: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. By iteration, we get the best hyperparameter (the best n_neighbors) in this list is 1, and we use this number to refit the model with the test set. At the end, it shows that the KNN test accuracy is 0.676923076923077, which is higher than the Zero Rule but lower than the Logistic regression.

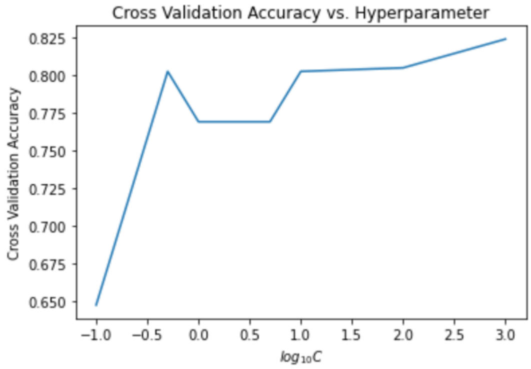


This plot shows the relation between the test accuracy and the $\log_{10}$(C).

4. **SVM：**

When given a training sample set, the SVM training algorithm builds a model that assigns new samples to one category or another, making it a non-probabilistic binary linear classifier. It maps the training samples to points to maximize the width of the gap between the two categories. The new samples are then mapped to the same space and predicted to belong to a certain category based on which side of the gap they fall on.

We tuned the regularization term C and evaluate the performance by cross validation accuracy, and we train the model using SVC class of sklearn.svm with the possible values of C which is shown with a list of numbers: [0.1, 0.5, 1, 5, 10, 100,1000]. By iteration, we get the best hyperparameter in this list is 1000, and we use this number to refit the model with the test set. At the end, it shows that the SVM test accuracy is 0.7846153846153846, which is the highest among all the model we consider in this task.



This plot shows the relation between the test accuracy and the $\log_{10}$(C). It is clear that the model is overfitting when the C becomes larger.

Conclusion

From the general results of 4 models, it is clear that the best model is the SVM, because it has the highest accuracy with 0. 7846153846153846, when training the test set, and the regularization term C = 1000. Meanwhile, all three machine learning models can achieve above 0.65 test accuracy, which is larger than baseline model Zero Rule that achieves 0.6307692307692307 as the test accuracy. However, the SVM is not the best machine learning model to deal with this task, since from the experiment, the C becomes larger and larger, which means the model will have overfitting. But if the C is closer to 0, we do not care whether the classification is correct or not, as long as the distance is larger, then we will not get the meaningful result (the model will have underfitting). Therefore, if we ignore the result of SVM model and only consider the rest 3 models, the best machine learning for determine whether a people get the Parkinson’s disease or not is the Logistic regression model, which has the accuracy of 0. 7076923076923077.