Midterm HW Report

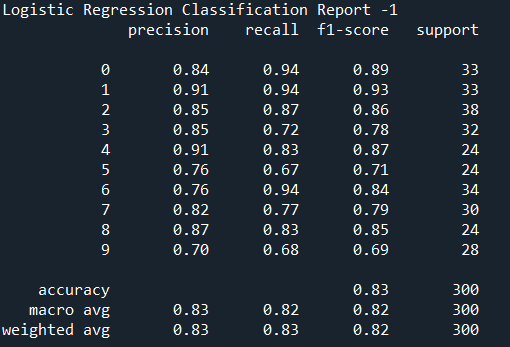
I have use the Logistic Regression, Decision Tree, and SVM classifier.

For Logistic Regression and Decision Tree, I split the dataset into training and testing.

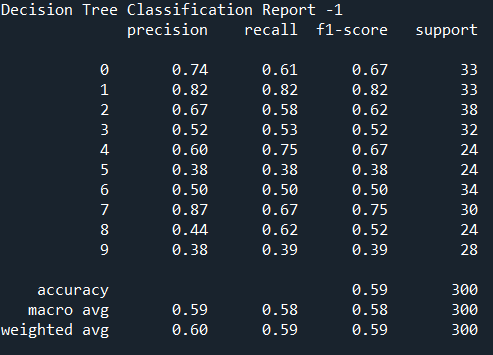
For SVM classifier, I split the dataset into training, validation and testing.

I only use several thousand samples from the whole sample. I’ve tried to use all, but it takes forever.

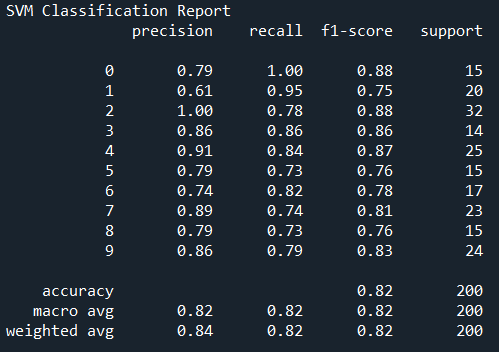
Logistic Regression Classification Report



Decision Tree Classification Report



SVM Classification Report



What is the best classifier and why?

By comparing the f-score with three different classifier, the performance of logistic regression and svm classifier are pretty similar, and the decision tree classifier is pretty low.

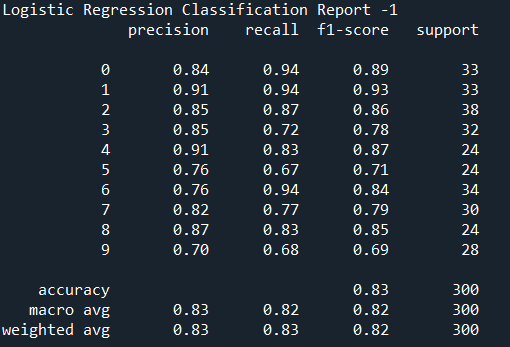
What affect the classification performance?

1. The first reason definitely is the different classifier. By comparing three classifier, we can see they have different performance.

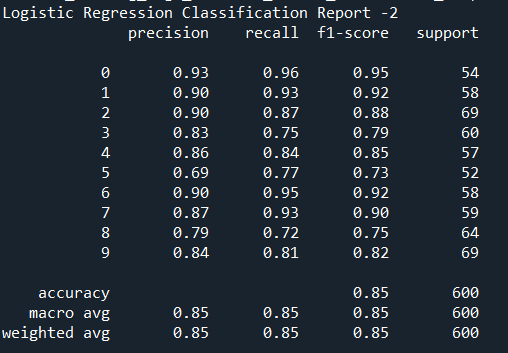
1. The second might be the size of training data size. I select different size training dataset and use the same classifier, and they performance differently.

Different size of training dataset in Logistic Regression

1000 sample

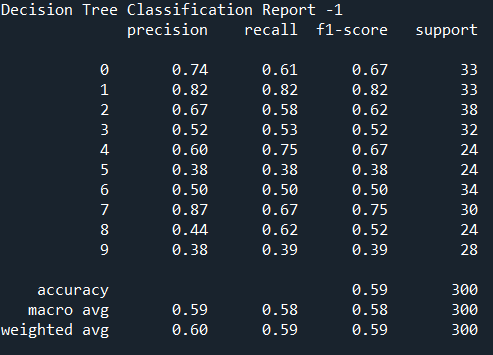


2000 sample

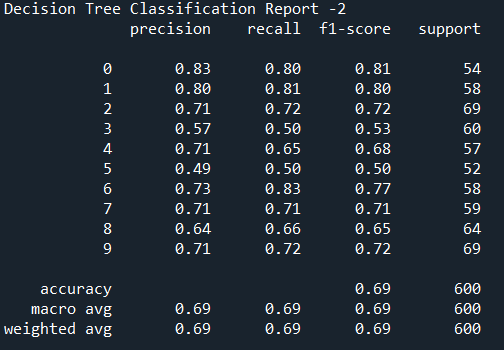


Different size of training dataset in Decision Tree Classifier

1000 samples



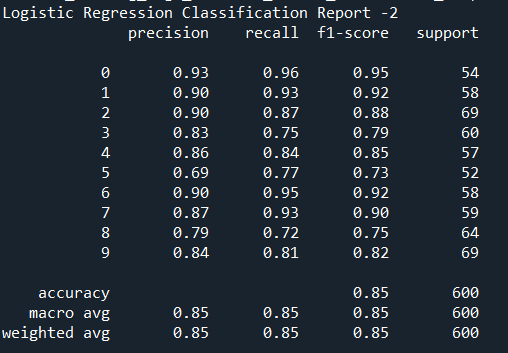
2000 samples



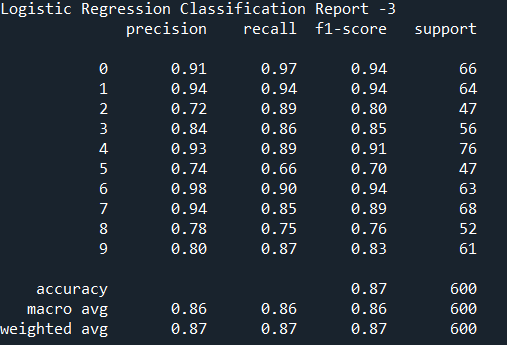
1. For the same size of dataset but using different training dataset will also influence the performance.

Different dataset for Logistic Regression

First 2000 data

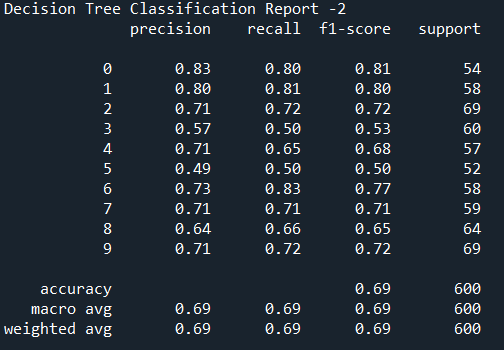


Last 2000 data

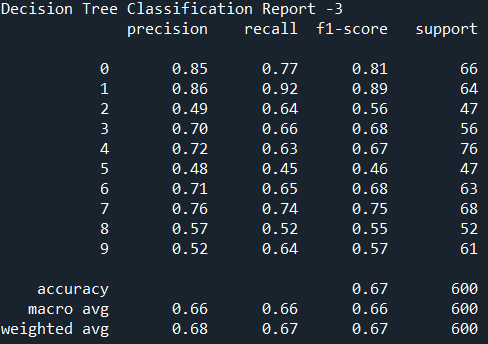


Different dataset for Decision Tree

First 2000 data



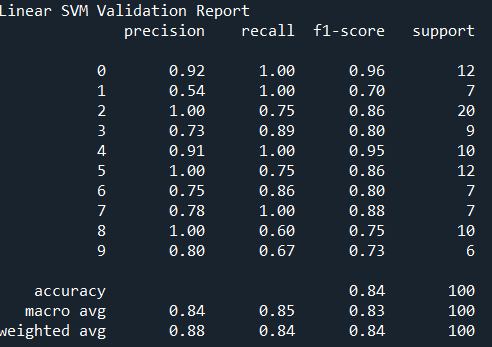
Last 2000 date



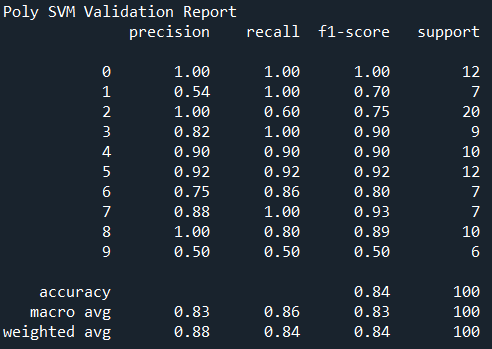
1. Different hyperparameter would influence the performance, too. For example, in SVM classifier, I use validation dataset to find the best kernel. The C and gamma are the same

Validation Report for SVM

Linear



Poly



Rbf

