Fabian Meyer, 6816480
Theresa Naß, 7122946

Support Vector Machines

This exercise is about classifying images of Chagas Parasites, similar to exercise 3. Thus, we use the same features as before (i.e. min/max/mean of the color channels).
To find the best hyperparameters, we use Scikit Learn's GridSearchCV. As hyperparameter, we use all combinations of

$$\begin{aligned} C &\in [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10] \\ \gamma &\in [1e-09, 2.15443469e-08, 4.6415888e-07, 1e-05, \\ &\quad 0.000215443469, 0.004641588, 0.1, 2.15443469, 46.4158883, 1000.0] \\ Kernels &\in [sigmoid,\ radial\,basis\,kernels,\ linear,\ polynomial\ (d = 1, 2, 3, 4)] \end{aligned}$$

We utilize a 5-fold cross-validation, which is standard in Scikit's library which is used to to generalize the results of the training, as we use, in contrast to the "traditional" way using a 70/20/10 split, splits of the complete data set to validate the model. This results in a longer training time, which is no problem in the our case, as we only have 60 training instances (We test five folds for 1440 parameter combinations, so 7200 fits in total). A negative aspect of the k-fold is, that we can only use 4/5 of our data to fit the parameter to our data.

Results

In Table 1 you can see the best results with the respective mean test score (averaged from the 5 folds). Many different combinations averaged similar results. For the full results see `https://github.com/gitfabianmeyer/ml_ss20/blob/master/Exercise4/results.csv`. The SVM algorithm classifies the given problem with all kernel function very good except for the sigmoid kernel, so we are not able to determine a best set of hyperparameter or features.

C	Degree	Gamma	Kernel	Mean Test Score	Rank
0.005	2	2.15443e-08	linear	0.933333	1
0.01	3	0.000215443	poly	0.933333	1
0.005	2	46.4159	linear	0.933333	1
0.1	2	0.000215443	poly	0.933333	1
0.005	4	2.15443	linear	0.933333	1
0.005	3	4.64159e-07	linear	0.933333	1
0.005	2	1e-09	linear	0.933333	1
0.005	4	1e-09	linear	0.933333	1
0.005	4	46.4159	linear	0.933333	1
0.005	1	1000	linear	0.933333	1

Table 1: Top ten results with hyperparameter

`https://github.com/gitfabianmeyer/ml_ss20`

Due to the design of Scikit's GridSearchCV module, we are not able to report the number of support vectors for every set of parameters, but for the best set of parameter for each kernel. Table 2 shows the number of support vectors for each kernel. It is observable, that the number of support vectors increases when the score decreases.

Kernel	Score	γ	C	Support Vectors
Linear	0.933	$1 * 10^{-9}$	0.01	9 8
Poly	0.933	0.0046	0.001	9 9
RBF	0.916	0.0002	5	12 11
Sigmoid	0.816	$1 * 10^{-9}$	0.001	30 30

Table 2: Best results for each kernel

1 Conclusion

Even though we were able to classify the given problem better (i.e. with a higher accouracy) with the GDA in the last exercise, we would argue that this happened because of the design of the classification process: As we have only 60 training instances, we fit the whole data to the GDA model. Thus, it is very likely that the models suffered from severe overfitting. As we use a five fold cross-validation, we use just 12 examples for testing. In these few examples, outlier will have a great impact on the performance and thus on the mean average score. Thus we can not conclude, that the SVM is less suitable for the given problem than the GDA from exercise 3.