PML Auto-SKLearn Fail

Milana M. Wolff

December 18, 2023

1 Auto SKLearn Fail

New answer: Auto-SKLearn fails because the data is continuous and encoded using OneHot encoding with the flag `handle_unknown = 'ignore'`, which is intended for discrete/categorical variables. Removing OneHot encoding entirely improves the performance of Auto-SKLearn from 0.5975 to [still running].

I ran the provided code on the Teton/Beartooth cluster (Python 3.7.16, GCC 11.2.0) and obtained a random forest accuracy of 0.6525 and an AutoML accuracy of 0.5975. I admit, I haven't figured out the source of the problem. I think it may be related to a warning I receive while running the code (UserWarning: Fitting transformer with a pandas series which has the dtype category. Inverse transform may not be able preserve dtype when converting to np.ndarray). I investigated the data and determined that the wine quality dataset was being used once again. I considered the OneHot encoding and whether or not this preprocessing step might influence performance. However, since the same preprocessing methods were applied to both the data used to train the random forest and the automated ensemble, the encoding seemed unlikely to cause the problem. I looked at the leaderboard of model candidates/contributors to the ensemble and found nothing obviously amiss.

model_id	rank	ensemble_weight	type	cost	duration
19	1	0.18	liblinear_svc	0.345960	4.164929
15	2	0.10	random_forest	0.348485	4.644477
17	3	0.16	passive_aggressive	0.351010	3.279581
2	4	0.04	random_forest	0.353535	3.428900
31	5	0.06	random_forest	0.363636	11.010535
8	6	0.04	random_forest	0.398990	3.812121
4	7	0.04	lda	0.439394	3.418703
22	8	0.06	random_forest	0.560606	3.855150
7	9	0.22	lda	0.568182	2.594877
26	10	0.06	random_forest	0.568182	8.400223
27	11	0.04	random_forest	0.651515	4.974203

At the present time, I have drawn no solid conclusions about the source of the Auto-SKLearn fail.