Lab 4- Ensemble Learners

Using the Sklearn Package in Python

Please utilize the table of contents. The data sections are very long.

Table of Contents

Task 1 – Train and Tune two ensemble Models

Adaboost

Tuning Process

I tuned the model based the following parameter set:

{Algorithm, maximum depth, number of estimators, learning rate}

I choose these parameters because they seemed to have the biggest effect on the accuracy of the model after some experimentation. In the end all my tests ranged from about 79% accurate on test data, up to 83% accurate on test data. My tuning had significant improvement (roughly 4%) over the baseline model which was around 79% accurate. Unfortunately this ensemble had difficulty gaining incredible (>95%) accuracy on such a small dataset. Also for this particular learning package the number of base classifiers was not tune-able for the Adaboost model.

In the table on the next page, you can see a grid search for the parameters with the accuracy being evaluated against the test data set. After a new "best" combination has been achieved, a star is printed to distinguish it from other rows. The final best accuracy is highlighted in yellow and copied below:

Algorithm	max_depth	n_estimators	learning_rate	accuracy
SAMME.R	1	9	1	0.830565

Discussion

Adaboost Confusion Matrix:

```
[[228  10]
 [ 41  22]]
```

The experiment showed me that ensemble model could be tuned to achieve better results. This dataset has roughly 79% of it's rows in class zero. This I think is the reason why the model is much better as classifying the zero cases. In fact before tuning most of the models were only marginally better than the trivial "guess zero" approach. Due to the small dataset size it was hard to squeeze out better performance.

Besides that one of the key assumptions for training the ensemble model is that the classifiers should be diverse. With so few data points I ponder that perhaps that in later experiments we may see that our ensembles may have trouble being particularly strong because our models will likely make the same mistakes on the same data points. In that case it is likely that our ensemble in part 3 will just be roughly as accurate as the other models.

Data and Accuracy

Algorithm	max_depth	n_estimators	learning_rate	accuracy
SAMME	1	1	0.001	0.790698
*				
SAMME	1	2	0.001	0.790698
SAMME	1	3	0.001	0.790698
SAMME	1	4	0.001	0.790698
SAMME	1	5	0.001	0.790698
SAMME	1	6	0.001	0.790698
SAMME	1	7	0.001	0.790698
SAMME	1	8	0.001	0.790698
SAMME	1	9	0.001	0.790698
SAMME	1	10	0.001	0.790698
SAMME	1	11	0.001	0.790698
SAMME	1	12	0.001	0.790698
SAMME	1	13	0.001	0.790698
SAMME	1	14	0.001	0.790698
SAMME	1	15	0.001	0.790698
SAMME	1	16	0.001	0.790698
SAMME	1	17	0.001	0.790698
SAMME	1	18	0.001	0.790698
SAMME	1	19	0.001	0.790698
SAMME	1	20	0.001	0.790698
SAMME	1	21	0.001	0.790698
SAMME	1	22	0.001	0.790698
SAMME	1	23	0.001	0.790698
SAMME	1	24	0.001	0.790698
SAMME	1	1	0.002	0.790698
SAMME	1	2	0.002	0.790698
SAMME	1	3	0.002	0.790698
SAMME	1	4	0.002	0.790698
SAMME	1	5	0.002	0.790698
SAMME	1	6	0.002	0.790698
SAMME	1	7	0.002	0.790698
SAMME	1	8	0.002	0.790698
SAMME	1	9	0.002	0.790698
SAMME	1	10	0.002	0.790698
SAMME	1	11	0.002	0.790698
SAMME	1	12	0.002	0.790698
SAMME	1	13	0.002	0.790698

SAMME	1	14	0.002	0.790698
SAMME	1	15	0.002	0.790698
SAMME	1	16	0.002	0.790698
SAMME	1	17	0.002	0.790698
SAMME	1	18	0.002	0.790698
SAMME	1	19	0.002	0.790698
SAMME	1	20	0.002	0.790698
SAMME	1	21	0.002	0.790698
SAMME	1	22	0.002	0.790698
SAMME	1	23	0.002	0.790698
SAMME	1	24	0.002	0.790698
SAMME	1	1	0.005	0.790698
SAMME	1	2	0.005	0.790698
SAMME	1	3	0.005	0.790698
SAMME	1	4	0.005	0.790698
SAMME	1	5	0.005	0.790698
SAMME	1	6	0.005	0.790698
SAMME	1	7	0.005	0.790698
SAMME	1	8	0.005	0.790698
SAMME	1	9	0.005	0.790698
SAMME	1	10	0.005	0.790698
SAMME	1	11	0.005	0.790698
SAMME	1	12	0.005	0.790698
SAMME	1	13	0.005	0.790698
SAMME	1	14	0.005	0.790698
SAMME	1	15	0.005	0.790698
SAMME	1	16	0.005	0.790698
SAMME	1	17	0.005	0.790698
SAMME	1	18	0.005	0.790698
SAMME	1	19	0.005	0.790698
SAMME	1	20	0.005	0.790698
SAMME	1	21	0.005	0.790698
SAMME	1	22	0.005	0.790698
SAMME	1	23	0.005	0.790698
SAMME	1	24	0.005	0.790698
SAMME	1	1	0.1	0.790698
SAMME	1	2	0.1	0.790698
SAMME	1	3	0.1	0.790698
SAMME	1	4	0.1	0.790698
SAMME	1	5	0.1	0.790698

SAMME	1	6	0.1	0.790698
SAMME	1	7	0.1	0.790698
SAMME	1	8	0.1	0.790698
SAMME	1	9	0.1	0.790698
SAMME	1	10	0.1	0.790698
SAMME	1	11	0.1	0.790698
SAMME	1	12	0.1	0.790698
SAMME	1	13	0.1	0.790698
SAMME	1	14	0.1	0.790698
SAMME	1	15	0.1	0.790698
SAMME	1	16	0.1	0.790698
SAMME	1	17	0.1	0.790698
SAMME	1	18	0.1	0.790698
SAMME	1	19	0.1	0.790698
SAMME	1	20	0.1	0.790698
SAMME	1	21	0.1	0.790698
SAMME	1	22	0.1	0.790698
SAMME	1	23	0.1	0.790698
SAMME	1	24	0.1	0.790698
SAMME	1	1	0.25	0.790698
SAMME	1	2	0.25	0.790698
SAMME	1	3	0.25	0.790698
SAMME	1	4	0.25	0.790698
SAMME	1	5	0.25	0.790698
SAMME	1	6	0.25	0.790698
SAMME	1	7	0.25	0.790698
SAMME	1	8	0.25	0.790698
SAMME	1	9	0.25	0.790698
SAMME	1	10	0.25	0.790698
SAMME	1	11	0.25	0.790698
SAMME	1	12	0.25	0.790698
SAMME	1	13	0.25	0.790698
SAMME	1	14	0.25	0.790698
SAMME	1	15	0.25	0.790698
SAMME	1	16	0.25	0.790698
SAMME	1	17	0.25	0.790698
SAMME	1	18	0.25	0.790698
SAMME	1	19	0.25	0.790698
SAMME	1	20	0.25	0.790698
SAMME	1	21	0.25	0.790698

SAMME	1	22	0.25	0.790698
SAMME	1	23	0.25	0.790698
SAMME	1	24	0.25	0.790698
SAMME	1	1	0.5	0.790698
SAMME	1	2	0.5	0.790698
SAMME	1	3	0.5	0.790698
SAMME	1	4	0.5	0.790698
SAMME	1	5	0.5	0.790698
SAMME	1	6	0.5	0.790698
SAMME	1	7	0.5	0.790698
SAMME	1	8	0.5	0.790698
SAMME	1	9	0.5	0.790698
SAMME	1	10	0.5	0.790698
SAMME	1	11	0.5	0.790698
SAMME	1	12	0.5	0.790698
SAMME	1	13	0.5	0.790698
SAMME	1	14	0.5	0.790698
SAMME	1	15	0.5	0.790698
SAMME	1	16	0.5	0.790698
SAMME	1	17	0.5	0.790698
SAMME	1	18	0.5	0.790698
SAMME	1	19	0.5	0.790698
SAMME	1	20	0.5	0.790698
SAMME	1	21	0.5	0.790698
SAMME	1	22	0.5	0.790698
SAMME	1	23	0.5	0.790698
SAMME	1	24	0.5	0.790698
SAMME	1	1	1	0.790698
SAMME	1	2	1	0.790698
SAMME	1	3	1	0.790698
SAMME	1	4	1	0.790698
SAMME	1	5	1	0.790698
SAMME	1	6	1	0.813953
*				
SAMME	1	7	1	0.803987
SAMME	1	8	1	0.813953
SAMME	1	9	1	0.803987
SAMME	1	10	1	0.803987
SAMME	1	11	1	0.803987
SAMME	1	12	1	0.803987

SAMME	1	13	1	0.803987
SAMME	1	14	1	0.803987
SAMME	1	15	1	0.803987
SAMME	1	16	1	0.803987
SAMME	1	17	1	0.803987
SAMME	1	18	1	0.803987
SAMME	1	19	1	0.803987
SAMME	1	20	1	0.803987
SAMME	1	21	1	0.803987
SAMME	1	22	1	0.803987
SAMME	1	23	1	0.803987
SAMME	1	24	1	0.803987
SAMME	1	1	1.25	0.790698
SAMME	1	2	1.25	0.790698
SAMME	1	3	1.25	0.770764
SAMME	1	4	1.25	0.790698
SAMME	1	5	1.25	0.790698
SAMME	1	6	1.25	0.797342
SAMME	1	7	1.25	0.790698
SAMME	1	8	1.25	0.787375
SAMME	1	9	1.25	0.817276
*				
SAMME	1	10	1.25	0.787375
SAMME	1	11	1.25	0.817276
SAMME	1	12	1.25	0.800664
SAMME	1	13	1.25	0.820598
*				
SAMME	1	14	1.25	0.820598
SAMME	1	15	1.25	0.820598
SAMME	1	16	1.25	0.817276
SAMME	1	17	1.25	0.817276
SAMME	1	18	1.25	0.817276
SAMME	1	19	1.25	0.817276
SAMME	1	20	1.25	0.817276
SAMME	1	21	1.25	0.817276
SAMME	1	22	1.25	0.817276
SAMME	1	23	1.25	0.817276
SAMME	1	24	1.25	0.817276
SAMME	1	1	1.5	0.790698
SAMME	1	2	1.5	0.790698

SAMME	1	3	1.5	0.770764
SAMME	1	4	1.5	0.790698
SAMME	1	5	1.5	0.784053
SAMME	1	6	1.5	0.780731
SAMME	1	7	1.5	0.780731
SAMME	1	8	1.5	0.790698
SAMME	1	9	1.5	0.784053
SAMME	1	10	1.5	0.790698
SAMME	1	11	1.5	0.784053
SAMME	1	12	1.5	0.790698
SAMME	1	13	1.5	0.784053
SAMME	1	14	1.5	0.790698
SAMME	1	15	1.5	0.784053
SAMME	1	16	1.5	0.790698
SAMME	1	17	1.5	0.784053
SAMME	1	18	1.5	0.79402
SAMME	1	19	1.5	0.787375
SAMME	1	20	1.5	0.79402
SAMME	1	21	1.5	0.787375
SAMME	1	22	1.5	0.803987
SAMME	1	23	1.5	0.797342
SAMME	1	24	1.5	0.807309
SAMME	1	1	2	0.790698
SAMME	1	2	2	0.790698
SAMME	1	3	2	0.790698
SAMME	1	4	2	0.790698
SAMME	1	5	2	0.790698
SAMME	1	6	2	0.790698
SAMME	1	7	2	0.790698
SAMME	1	8	2	0.790698
SAMME	1	9	2	0.790698
SAMME	1	10	2	0.790698
SAMME	1	11	2	0.790698
SAMME	1	12	2	0.790698
SAMME	1	13	2	0.790698
SAMME	1	14	2	0.790698
SAMME	1	15	2	0.790698
SAMME	1	16	2	0.790698
SAMME	1	17	2	0.790698
SAMME	1	18	2	0.790698

SAMME	1	19	2	0.790698
SAMME	1	20	2	0.790698
SAMME	1	21	2	0.790698
SAMME	1	22	2	0.790698
SAMME	1	23	2	0.790698
SAMME	1	24	2	0.790698
SAMME.R	1	1	0.001	0.790698
SAMME.R	1	2	0.001	0.790698
SAMME.R	1	3	0.001	0.790698
SAMME.R	1	4	0.001	0.790698
SAMME.R	1	5	0.001	0.790698
SAMME.R	1	6	0.001	0.790698
SAMME.R	1	7	0.001	0.790698
SAMME.R	1	8	0.001	0.790698
SAMME.R	1	9	0.001	0.790698
SAMME.R	1	10	0.001	0.790698
SAMME.R	1	11	0.001	0.790698
SAMME.R	1	12	0.001	0.790698
SAMME.R	1	13	0.001	0.790698
SAMME.R	1	14	0.001	0.790698
SAMME.R	1	15	0.001	0.790698
SAMME.R	1	16	0.001	0.790698
SAMME.R	1	17	0.001	0.790698
SAMME.R	1	18	0.001	0.790698
SAMME.R	1	19	0.001	0.790698
SAMME.R	1	20	0.001	0.790698
SAMME.R	1	21	0.001	0.790698
SAMME.R	1	22	0.001	0.790698
SAMME.R	1	23	0.001	0.790698
SAMME.R	1	24	0.001	0.790698
SAMME.R	1	1	0.002	0.790698
SAMME.R	1	2	0.002	0.790698
SAMME.R	1	3	0.002	0.790698
SAMME.R	1	4	0.002	0.790698
SAMME.R	1	5	0.002	0.790698
SAMME.R	1	6	0.002	0.790698
SAMME.R	1	7	0.002	0.790698
SAMME.R	1	8	0.002	0.790698
SAMME.R	1	9	0.002	0.790698
SAMME.R	1	10	0.002	0.790698

SAMME.R	1	11	0.002	0.790698
SAMME.R	1	12	0.002	0.790698
SAMME.R	1	13	0.002	0.790698
SAMME.R	1	14	0.002	0.790698
SAMME.R	1	15	0.002	0.790698
SAMME.R	1	16	0.002	0.790698
SAMME.R	1	17	0.002	0.790698
SAMME.R	1	18	0.002	0.790698
SAMME.R	1	19	0.002	0.790698
SAMME.R	1	20	0.002	0.790698
SAMME.R	1	21	0.002	0.790698
SAMME.R	1	22	0.002	0.790698
SAMME.R	1	23	0.002	0.790698
SAMME.R	1	24	0.002	0.790698
SAMME.R	1	1	0.005	0.790698
SAMME.R	1	2	0.005	0.790698
SAMME.R	1	3	0.005	0.790698
SAMME.R	1	4	0.005	0.790698
SAMME.R	1	5	0.005	0.790698
SAMME.R	1	6	0.005	0.790698
SAMME.R	1	7	0.005	0.790698
SAMME.R	1	8	0.005	0.790698
SAMME.R	1	9	0.005	0.790698
SAMME.R	1	10	0.005	0.790698
SAMME.R	1	11	0.005	0.790698
SAMME.R	1	12	0.005	0.790698
SAMME.R	1	13	0.005	0.790698
SAMME.R	1	14	0.005	0.790698
SAMME.R	1	15	0.005	0.790698
SAMME.R	1	16	0.005	0.790698
SAMME.R	1	17	0.005	0.790698
SAMME.R	1	18	0.005	0.790698
SAMME.R	1	19	0.005	0.790698
SAMME.R	1	20	0.005	0.790698
SAMME.R	1	21	0.005	0.790698
SAMME.R	1	22	0.005	0.790698
SAMME.R	1	23	0.005	0.790698
SAMME.R	1	24	0.005	0.790698
SAMME.R	1	1	0.1	0.790698
SAMME.R	1	2	0.1	0.790698

SAMME.R	1	3	0.1	0.790698
SAMME.R	1	4	0.1	0.790698
SAMME.R	1	5	0.1	0.790698
SAMME.R	1	6	0.1	0.790698
SAMME.R	1	7	0.1	0.790698
SAMME.R	1	8	0.1	0.790698
SAMME.R	1	9	0.1	0.790698
SAMME.R	1	10	0.1	0.790698
SAMME.R	1	11	0.1	0.790698
SAMME.R	1	12	0.1	0.790698
SAMME.R	1	13	0.1	0.790698
SAMME.R	1	14	0.1	0.790698
SAMME.R	1	15	0.1	0.790698
SAMME.R	1	16	0.1	0.790698
SAMME.R	1	17	0.1	0.790698
SAMME.R	1	18	0.1	0.790698
SAMME.R	1	19	0.1	0.790698
SAMME.R	1	20	0.1	0.790698
SAMME.R	1	21	0.1	0.790698
SAMME.R	1	22	0.1	0.790698
SAMME.R	1	23	0.1	0.790698
SAMME.R	1	24	0.1	0.790698
SAMME.R	1	1	0.25	0.790698
SAMME.R	1	2	0.25	0.790698
SAMME.R	1	3	0.25	0.790698
SAMME.R	1	4	0.25	0.790698
SAMME.R	1	5	0.25	0.790698
SAMME.R	1	6	0.25	0.790698
SAMME.R	1	7	0.25	0.790698
SAMME.R	1	8	0.25	0.790698
SAMME.R	1	9	0.25	0.790698
SAMME.R	1	10	0.25	0.790698
SAMME.R	1	11	0.25	0.790698
SAMME.R	1	12	0.25	0.790698
SAMME.R	1	13	0.25	0.790698
SAMME.R	1	14	0.25	0.790698
SAMME.R	1	15	0.25	0.790698
SAMME.R	1	16	0.25	0.797342
SAMME.R	1	17	0.25	0.797342
SAMME.R	1	18	0.25	0.797342

SAMME.R	1	19	0.25	0.797342
SAMME.R	1	20	0.25	0.797342
SAMME.R	1	21	0.25	0.797342
SAMME.R	1	22	0.25	0.797342
SAMME.R	1	23	0.25	0.797342
SAMME.R	1	24	0.25	0.797342
SAMME.R	1	1	0.5	0.790698
SAMME.R	1	2	0.5	0.790698
SAMME.R	1	3	0.5	0.790698
SAMME.R	1	4	0.5	0.790698
SAMME.R	1	5	0.5	0.790698
SAMME.R	1	6	0.5	0.790698
SAMME.R	1	7	0.5	0.790698
SAMME.R	1	8	0.5	0.790698
SAMME.R	1	9	0.5	0.797342
SAMME.R	1	10	0.5	0.797342
SAMME.R	1	11	0.5	0.803987
SAMME.R	1	12	0.5	0.803987
SAMME.R	1	13	0.5	0.803987
SAMME.R	1	14	0.5	0.803987
SAMME.R	1	15	0.5	0.803987
SAMME.R	1	16	0.5	0.803987
SAMME.R	1	17	0.5	0.800664
SAMME.R	1	18	0.5	0.800664
SAMME.R	1	19	0.5	0.800664
SAMME.R	1	20	0.5	0.800664
SAMME.R	1	21	0.5	0.803987
SAMME.R	1	22	0.5	0.803987
SAMME.R	1	23	0.5	0.803987
SAMME.R	1	24	0.5	0.803987
SAMME.R	1	1	1	0.790698
SAMME.R	1	2	1	0.790698
SAMME.R	1	3	1	0.790698
SAMME.R	1	4	1	0.810631
SAMME.R	1	5	1	0.803987
SAMME.R	1	6	1	0.82392
*				
SAMME.R	1	7	1	0.810631
SAMME.R	1	8	1	0.810631
SAMME.R	1	9	1	0.830565

*				
SAMME.R	1	10	1	0.82392
SAMME.R	1	11	1	0.830565
SAMME.R	1	12	1	0.830565
SAMME.R	1	13	1	0.82392
SAMME.R	1	14	1	0.820598
SAMME.R	1	15	1	0.820598
SAMME.R	1	16	1	0.820598
SAMME.R	1	17	1	0.817276
SAMME.R	1	18	1	0.817276
SAMME.R	1	19	1	0.817276
SAMME.R	1	20	1	0.817276
SAMME.R	1	21	1	0.817276
SAMME.R	1	22	1	0.817276
SAMME.R	1	23	1	0.817276
SAMME.R	1	24	1	0.817276
SAMME.R	1	1	1.25	0.790698
SAMME.R	1	2	1.25	0.790698
SAMME.R	1	3	1.25	0.790698
SAMME.R	1	4	1.25	0.790698
SAMME.R	1	5	1.25	0.797342
SAMME.R	1	6	1.25	0.797342
SAMME.R	1	7	1.25	0.813953
SAMME.R	1	8	1.25	0.810631
SAMME.R	1	9	1.25	0.797342
SAMME.R	1	10	1.25	0.830565
SAMME.R	1	11	1.25	0.803987
SAMME.R	1	12	1.25	0.803987
SAMME.R	1	13	1.25	0.803987
SAMME.R	1	14	1.25	0.820598
SAMME.R	1	15	1.25	0.800664
SAMME.R	1	16	1.25	0.813953
SAMME.R	1	17	1.25	0.79402
SAMME.R	1	18	1.25	0.813953
SAMME.R	1	19	1.25	0.810631
SAMME.R	1	20	1.25	0.810631
SAMME.R	1	21	1.25	0.810631
SAMME.R	1	22	1.25	0.79402
SAMME.R	1	23	1.25	0.810631
SAMME.R	1	24	1.25	0.807309

SAMME.R	1	1	1.5	0.790698
SAMME.R	1	2	1.5	0.770764
SAMME.R	1	3	1.5	0.790698
SAMME.R	1	4	1.5	0.790698
SAMME.R	1	5	1.5	0.790698
SAMME.R	1	6	1.5	0.784053
SAMME.R	1	7	1.5	0.797342
SAMME.R	1	8	1.5	0.803987
SAMME.R	1	9	1.5	0.803987
SAMME.R	1	10	1.5	0.79402
SAMME.R	1	11	1.5	0.79402
SAMME.R	1	12	1.5	0.79402
SAMME.R	1	13	1.5	0.790698
SAMME.R	1	14	1.5	0.790698
SAMME.R	1	15	1.5	0.777409
SAMME.R	1	16	1.5	0.787375
SAMME.R	1	17	1.5	0.770764
SAMME.R	1	18	1.5	0.787375
SAMME.R	1	19	1.5	0.770764
SAMME.R	1	20	1.5	0.797342
SAMME.R	1	21	1.5	0.784053
SAMME.R	1	22	1.5	0.787375
SAMME.R	1	23	1.5	0.784053
SAMME.R	1	24	1.5	0.787375
SAMME.R	1	1	2	0.790698
SAMME.R	1	2	2	0.305648
SAMME.R	1	3	2	0.790698
SAMME.R	1	4	2	0.305648
SAMME.R	1	5	2	0.790698
SAMME.R	1	6	2	0.780731
SAMME.R	1	7	2	0.790698
SAMME.R	1	8	2	0.780731
SAMME.R	1	9	2	0.790698
SAMME.R	1	10	2	0.780731
SAMME.R	1	11	2	0.790698
SAMME.R	1	12	2	0.780731
SAMME.R	1	13	2	0.790698
SAMME.R	1	14	2	0.780731
SAMME.R	1	15	2	0.790698
SAMME.R	1	16	2	0.780731

SAMME.R	1	17	2	0.790698
SAMME.R	1	18	2	0.780731
SAMME.R	1	19	2	0.790698
SAMME.R	1	20	2	0.780731
SAMME.R	1	21	2	0.790698
SAMME.R	1	22	2	0.780731
SAMME.R	1	23	2	0.790698
SAMME.R	1	24	2	0.780731

Random Forest

Tuning Process

I tuned the following hyper-parameters:

{number of estimators, max features, minimum sample splits,

minimum sample leaf, maximum depth}

I choose these parameters because they seemed to have the biggest effect on the accuracy of the model after some experimentation. In the end all my tests ranged from about 79% accurate on test data, up to 84% accurate on test data. My tuning had significant improvement (roughly 5%) over the baseline model which was around 79% accurate.

Below is the best parameters for the model:

num_estimators	max_features	min_sample_splits	min_samples_leaf	max_depth	accuracy
4	2	2	4	5	0.840532

Discussion

Random Forest Confusion Matrix:

```
[[222  16]
 [ 32  31]]
```

Surprisingly this model was my most accurate one. As you will see the ensembles in this project that I create fail because of two main reasons that I can detect:

- The base models that it is built on are likely not very diverse. The most likely make the same mistakes on the same samples. (The minority of samples that classify as a 1)
- The models have very similar accuracies. So much so that weighting them with their accuracy instead of a constant doesn't change the prediction of a single data point.

Due to these two reasons, I surmise the following:

- The final ensemble models are not as powerful as the random forest model because for the majority of cases the models make the same mistakes on the same data points.
- In the case wherein some model, like say the random forest, may predict correctly on a particular data point, it is possible that the majority of the others will not. Normally this could be offset by using the accuracies as weights, however because the accuracies are so close to each other the models would need very good diversity to beat the best one it's built on.

The only thing that I think might push the accuracies a little father was if we weighted the ensemble voting system by class-accuracies and not my overall accuracies. By weighting by class accuracies and then exaggerating those weights I can tune the ensemble using my knowledge of how each model performs by class.

Data

num_estimators	max_features	min_sample_splits	min_samples_leaf	max_depth	accuracy
1	2	2	2	3	0.810631
*					
1	2	2	2	4	0.807309
1	2	2	2	5	0.82392
*					
1	2	2	3	3	0.810631
1	2	2	3	4	0.810631
1	2	2	3	5	0.827243
*					
1	2	2	4	3	0.810631
1	2	2	4	4	0.810631
1	2	2	4	5	0.827243
1	2	2	5	3	0.790698
1	2	2	5	4	0.820598
1	2	2	5	5	0.803987
1	3	2	2	3	0.810631
1	3	2	2	4	0.800664
1	3	2	2	5	0.750831
1	3	2	3	3	0.810631
1	3	2	3	4	0.803987
1	3	2	3	5	0.784053
1	3	2	4	3	0.810631
1	3	2	4	4	0.803987
1	3	2	4	5	0.784053
1	3	2	5	3	0.82392
1	3	2	5	4	0.803987
1	3	2	5	5	0.777409
2	2	2	2	3	0.817276
2	2	2	2	4	0.813953
2	2	2	2	5	0.827243
2	2	2	3	3	0.817276
2	2	2	3	4	0.817276
2	2	2	3	5	0.827243
2	2	2	4	3	0.817276
2	2	2	4	4	0.817276
2	2	2	4	5	0.827243
2	2	2	5	3	0.803987

2	2	2	5	4	0.820598
2	2	2	5	5	0.830565
*					
2	3	2	2	3	0.817276
2	3	2	2	4	0.813953
2	3	2	2	5	0.810631
2	3	2	3	3	0.817276
2	3	2	3	4	0.813953
2	3	2	3	5	0.807309
2	3	2	4	3	0.817276
2	3	2	4	4	0.813953
2	3	2	4	5	0.813953
2	3	2	5	3	0.803987
2	3	2	5	4	0.82392
2	3	2	5	5	0.777409
3	2	2	2	3	0.810631
3	2	2	2	4	0.82392
3	2	2	2	5	0.817276
3	2	2	3	3	0.810631
3	2	2	3	4	0.803987
3	2	2	3	5	0.827243
3	2	2	4	3	0.810631
3	2	2	4	4	0.803987
3	2	2	4	5	0.830565
3	2	2	5	3	0.797342
3	2	2	5	4	0.820598
3	2	2	5	5	0.810631
3	3	2	2	3	0.797342
3	3	2	2	4	0.817276
3	3	2	2	5	0.820598
3	3	2	3	3	0.797342
3	3	2	3	4	0.817276
3	3	2	3	5	0.820598
3	3	2	4	3	0.797342
3	3	2	4	4	0.817276
3	3	2	4	5	0.820598
3	3	2	5	3	0.807309
3	3	2	5	4	0.807309
3	3	2	5	5	0.807309
4	2	2	2	3	0.82392

4	2	2	2	4	0.837209
*					
4	2	2	2	5	0.833887
4	2	2	3	3	0.82392
4	2	2	3	4	0.827243
4	2	2	3	5	0.830565
4	2	2	4	3	0.82392
4	2	2	4	4	0.827243
4	2	2	4	5	0.840532
*					
4	2	2	5	3	0.810631
4	2	2	5	4	0.827243
4	2	2	5	5	0.810631
4	3	2	2	3	0.797342
4	3	2	2	4	0.813953
4	3	2	2	5	0.820598
4	3	2	3	3	0.797342
4	3	2	3	4	0.813953
4	3	2	3	5	0.813953
4	3	2	4	3	0.803987
4	3	2	4	4	0.813953
4	3	2	4	5	0.82392
4	3	2	5	3	0.807309
4	3	2	5	4	0.810631
4	3	2	5	5	0.817276

Task 2 – Training 5 models with light tuning

Model data

In this section I'm going to present the accuracies of these models, the tuned parameters I ended up choosing, and their confusion matrices.

neural network score : 0.8272425249169435

neural network matrix: [[222 16] [36 27]]

params: { activation='tanh', solver='lbfgs', alpha=1, learning_rate='constant'}

```
activation = ['identity', 'logistic', 'tanh', 'relu']
solver = ['lbfgs', 'sgd', 'adam']
alpha = [0.0001, 0.001, 0.01, 0.1, 1]
learning_rate = ['constant', 'invscaling', 'adaptive']
```

knn score : 0.8039867109634552

k nearest neighbor matrix: [[229 9] [50 13]]

params: { n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=4 }

```
neigh = range(1, 10)
weigh = ['uniform', 'distance']
algo = ['auto', 'ball_tree', 'kd_tree', 'brute']
leaf_s = range(1, 35)
```

logistic regression score : 0.813953488372093

logistic regression matrix: [[232 6] [50 13]]

params: { random_state=1, penalty='l1', tol=1e-06, C=1, class_weight=None

, solver='liblinear', fit_intercept=False}

```
penalty = ['l1', 'l2']
tol = [1e-6, 1e-5, 1e-4, 1e-3, 1e-2]
c = [1, .1, .01, 1.5, 2]
weights = [None, 'balanced']
solver = ['newton-cg', 'lbfgs', 'sag' ]
solverl1 = ['liblinear', 'saga']
fit = [True, False]
```

naive bayes score : 0.8006644518272426

naive bayes matrix: [[226 12] [48 15]]

decision trees : 0.8172757475083057

decision tree matrix: [[217 21] [34 29]]

params : { presort=True, random_state=1,

max_features=3,

min_samples_split=2, min_samples_leaf=2, max_depth=5 }

```
max_features = range(2, 4)
min_sample_splits = range(2, 3)
min_samples_leaf = range(2, 6)
max_depth = range(3, 6)
```

Discussion:

I tuned the models using a grid search but I exclude the data because it isn't asked for. Each model improved from its accuracy under default conditions. I am happy I was able to move every model into the 80s. Unfortunately every model is more or less as accurate as the others and it doesn't end up affecting the weighted versus unweighted model very much. (near constant weights perform identically to constant weights).

Task 3 – Training custom ensembles

Creating Process

I employ a weighted voting system against either 5 or 7 different models (marked with a _5 if using only the five models from part 2) and also either weighted by accuracies or unweighted (weights are held constant at 1) which is marked with a _u.

Accuracy

ensemble_5 unweighted accuracy : 0.8272425249169435

ensemble_5 accuracy : 0.8272425249169435

ensemble unweighted accuracy : 0.8372093023255814

ensemble accuracy : 0.8372093023255814

Confusion Matrix

Ensemble_u5 matrix: [[229 9] [43 20]]

Ensemble_5 matrix: [[229 9] [43 20]]

Ensemble_u matrix: [[229 9] [40 23]]

Ensemble matrix: [[229 9] [40 23]]

Discussion

I already gave some reasons for why I think the ensembles have a worse accuracy, and a suggestion for how to improve it. Overall I'd like to talk about what I learned from this lab. I learned how important it is to have adequate data for training, in terms of raw number of samples and number of features. I bet I could have gotten better accuracy. And for ensembles it's very important that the base models are diverse. If they aren't then the accuracy will most likely be less then or equal to the best.