MachineLearning_homework1_2

M10715812 張天佑

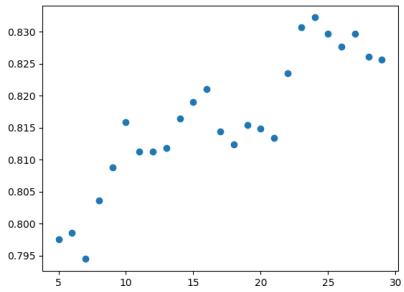
Bank Marketing :

When I get the data, I do a simple analysis of the data through basic data descriptions. We will get the values and strings from it. For the value we can use it directly.we needs to construct a dictionary and convert it to the value we need for strings. At the same time, there are unnecessary or negligible features in the data which we can handle in advance. Since the rate of missing data is not high, they are removed directly in the data preprocessing. In addition, for the unstable values of some features, they need to be normalized.

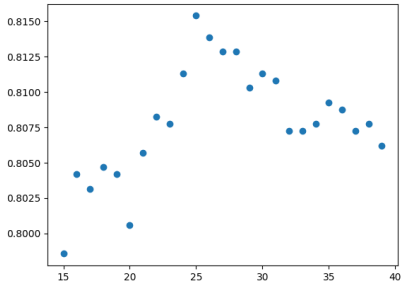
```
- ------------------------------
- 固定参数:
- name :      DT_max_depth            , score :      0.80418
- name :      DT_min_samples_split    , score :      0.79194
- name :      DT_min_samples_leaf     , score :      0.81540
- ------------------------------
- 调整参数:
- name :      DT_max_depth            , score :      0.84855
- name :      DT_min_samples_split    , score :      0.80061
- name :      DT_min_samples_leaf     , score :      0.83223
homework01 git:(master) ✗
```

The following three parameters are selected as the main adjustment targets：

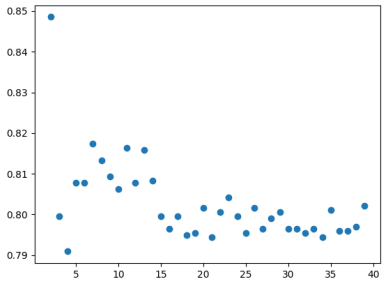
1.Min_samples_leaf



2.Samples_split



3.Max_Depth



In my implementation, I got it through a single decision tree. When my depth is about 6,7, I can get the best score. The best results are also shown when the leaf has a minimum sample size of 23 and the node minimum sample size is limited to 25.

Spooky Author Identification

For the second mission, to be honest, I didn't make it. I encountered some problems in the case of data processing. Then I refer to the article about the decision tree that is not suitable for applying analytical text. I probably get the following information. For the analysis of this data, the decision tree built will be very large.

Conclusion:

For the previous task, I implemented a random forest and the screenshot is as follows . Better results can be observed with random forests.

```
- 固定参数：
- name :       RF                    , score :      0.84192
- name :       DT_max_depth          , score :      0.82968
- name :       DT_min_samples_split  , score :      0.80775
- name :       DT_min_samples_leaf   , score :      0.80520
- ----------------------------
- 调整参数：
- name :       RF                    , score :      0.85773
- name :       DT_max_depth          , score :      0.82866
- name :       DT_min_samples_split  , score :      0.81336
- name :       DT_min_samples_leaf   , score :      0.81693
➜ homework01 git:(master) ✗
```

1. The single decision tree function is very limited and the random forest algorithm also has its use scene, which has its limitations. If the data is of short duration and the data quality is normal, the decision tree (random forest algorithm) is preferred, and other models are preferentially used for structured data such as voice, picture, and text.

2. From the perspective of data volume, the tree model has obvious advantages for small data.
3. The tree model does not require strict data preparation.
4. The tree model adjustment parameters are simple.
5. The integrated tree model has a certain predictive power, but the ability will decrease as the amount of data increasing.