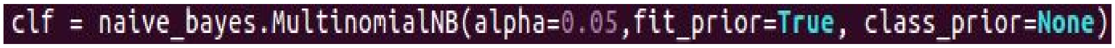
Data Mining HW4

Scikit-Learn

Name: 李友岐 Department: 電機碩二 Student ID: r06921048

1. News Dataset: Testing label is provided
   a. Implement Naive Bayes on News dataset
      i. What’s the parameters and performance of your best model ? (Baseline: Test accuracy 85%) [10%]

A:



```
clf = naive_bayes.MultinomialNB(alpha=0.05,fit_prior=True, class_prior=None)
```

Test accuracy= 89.44%

      ii. Compare different distribution assumption, which is the most suitable for News dataset ? List the testing accuracy. [5%]

A:

Test accuracy (all use default parameters, may not be the best model):

naive_bayes.GaussianNB(): 80.98%

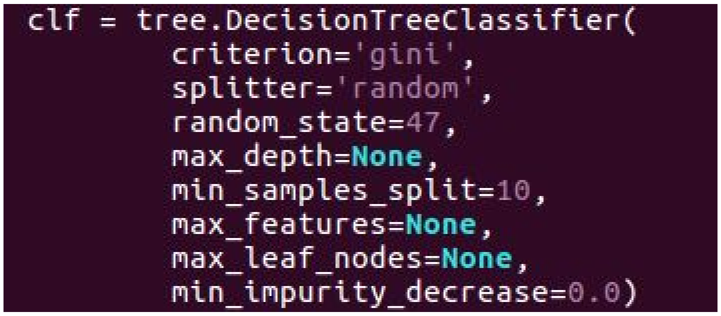
naive_bayes.BernoulliNB(): 76.78%

naive_bayes.MultinomialNB(): 82.45%

Obviously, Multinomial is the most suitable for News dataset among these three different distributions.

   b. Implement Decision Tree on News dataset
      i. What’s the parameters and performance of your best model ? (Baseline: Test accuracy 61%) [10%]

A:



```
 clf = tree.DecisionTreeClassifier(
           criterion='gini',
           splitter='random',
           random_state=47,
           max_depth=None,
           min_samples_split=10,
           max_features=None,
           max_leaf_nodes=None,
           min_impurity_decrease=0.0)
```

Test accuracy= 65.45%

   c. How do you choose the parameters to get the best model ? [5%]

A: First, I slightly modify each parameter to see whether there is a huge improvement. If modifying this parameter can highly improve the accuracy, then I would keep adjusting this parameter and make the other parameters remain unchanged. After a lot of tests, I can get the best model.

2. Mushroom Dataset: Testing label is provided

   a. How do you preprocess the mushroom dataset? [5%]

   A: First, I use ord() function in Python to transform all character values to their decimal value in ASCII table.

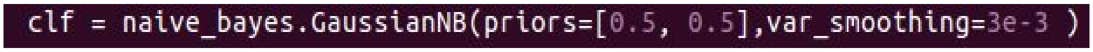
   Second, I delete the whole 11th column in order to handle the missing values in 11th attribute. After removing attribute #11, there is no missing value remain.

   Finally, I transform the dataset into one hot encoding by get_dummies from pandas. As a result, the dataset only contains 0 and 1.

   b. Implement Naive Bayes on mushroom dataset

      i. What's the parameters and performance of your best model ? (Baseline: Test accuracy 98%) [10%]

      A:



```
clf = naive_bayes.GaussianNB(priors=[0.5, 0.5],var_smoothing=3e-3 )
```

      Test accuracy= 99.20%

      ii. Compare different distribution assumption, which is the most suitable for mushroom dataset ? List the testing accuracy. [5%]

      A:

      Test accuracy (all use default parameters, may not be the best model):

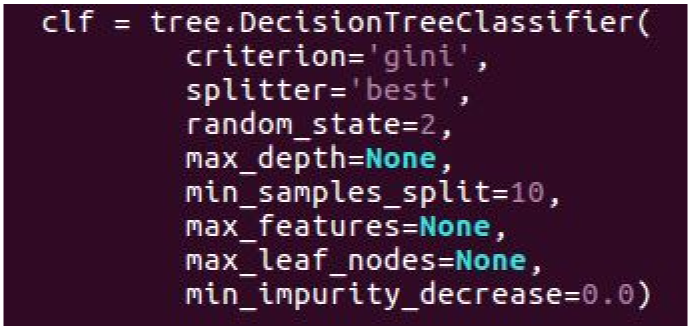
      naive_bayes.GaussianNB(): 98.95%
      naive_bayes.BernoulliNB(): 94.95%
      naive_bayes.MultinomialNB(): 95.93%

      Obviously, Gaussian is the most suitable for mushroom dataset among these three different distributions.

   c. Implement Decision Tree on mushroom dataset

      i. What's the performance of your best model ? (Baseline: Test accuracy 99%) [10%]

      A:

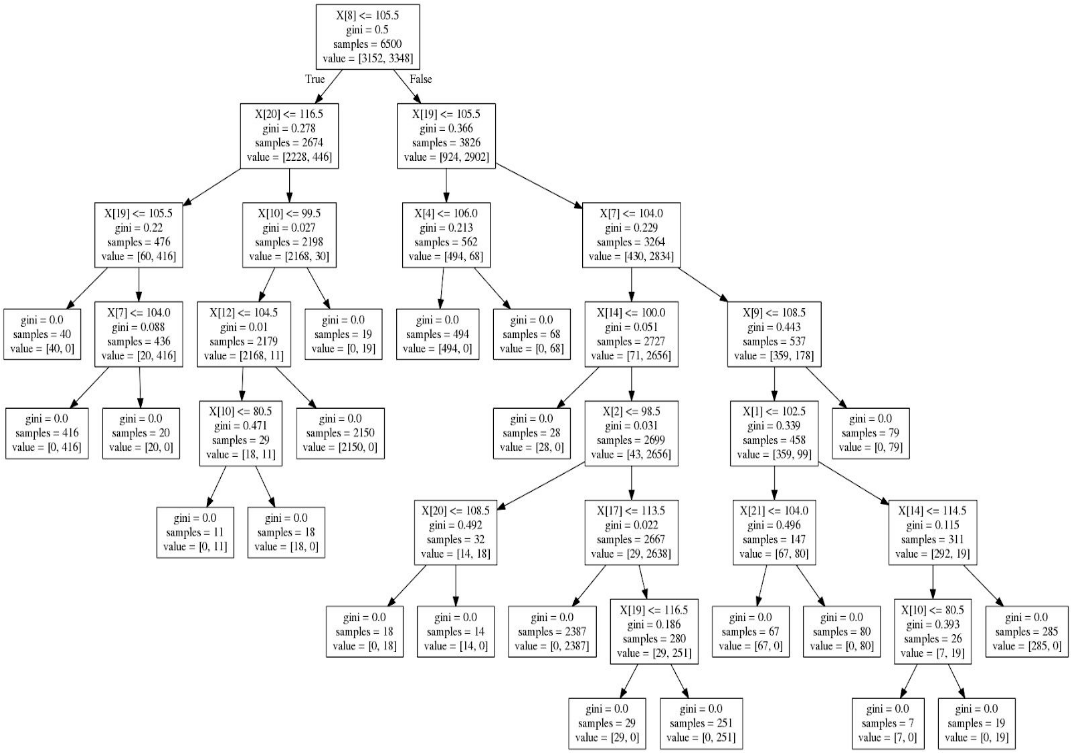


```
clf = tree.DecisionTreeClassifier(
        criterion='gini',
        splitter='best',
        random_state=2,
        max_depth=None,
        min_samples_split=10,
        max_features=None,
        max_leaf_nodes=None,
        min_impurity_decrease=0.0)
```

      Test accuracy: 100%

ii. Use graphviz tool to plot your decision tree [5%]

A:



d. Observe the data properties of News and mushroom dataset. According to the model performance, what kind of dataset is more suitable for naive bayes / decision tree ? [5%]

A: If the dataset has lots of attributes (column) but the amount of samples (row) is not large (# of columns >> # of rows), then naive bayes can result in better accuracy, which means such dataset is more suitable for naive bayes. Decision tree often lead to overfitting on such case.

By contrast, if the dataset has lots of samples (row) but only contains a few attributes (column), which means (# of rows >> # of columns), then it is more suitable for decision tree. Decision tree can lead to better accuracy.

3. Income Dataset: Testing label is **not** provided
   Implement Naive Bayes and Decision Tree on income dataset
   a. How do you preprocess the data ? Missing value ? [10%]

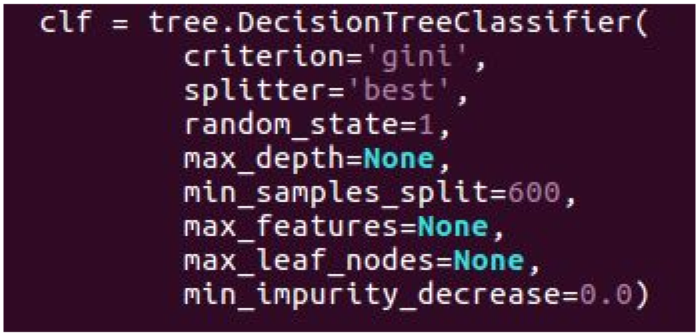
   A: First, I transform all string values to integer by summing the ASCII decimal value of each character of the string.
   (e.g. ‘Male’ -> 77+97+108+101 -> 383).

   Second, I change each missing value to be the mean of the column it belongs to by SimpleImputer from sklearn.

b. Which model gets better performance ? Show the parameters. (Surpass the weak baseline (Test accuracy: 80%) for 10%. Strong baseline (Test accuracy: 85%) for 10%)

A:

Decision Tree gets better performance.



```
  clf = tree.DecisionTreeClassifier(
          criterion='gini',
          splitter='best',
          random_state=1,
          max_depth=None,
          min_samples_split=600,
          max_features=None,
          max_leaf_nodes=None,
          min_impurity_decrease=0.0)
```