Data Mining HW4

Scikit-Learn

Name: 張緣彩 Department: 資訊工程所 Student ID: R07922141

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%]
        By using Multinomial Naïve Bayes classifier with **alpha=0.05** and Test accuracy of **89.44%**

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

	Gaussian NB	**Multinomial NB**	Bernoulli NB	Complement NB
Acc.	80.98%	**89.44%**	76.78%	88.39%

        Multinomial distribution assumption is most suitable for News dataset.

    b. Implement Decision Tree on News dataset
        i. What's the parameters and performance of your best model? (Baseline: Test accuracy 61%) [10%]
        By using **max_depth=55** and Test accuracy of **62.87%**

    c. How do you choose the parameters to get the best model? [5%]
    I used grid search-based method to find the best parameters. Using a for loop and loop through all the possible parameters and find the best parameters which resulting best test accuracy. For example, in Naïve Bayes classifier, I only need to test every alpha value from 0 to 1 in every step of 0.01.

2. Mushroom Dataset: Testing label is provided
    a. How do you preprocess the mushroom dataset? [5%]
    By using one-hot encoding. Since there is a one-hot encoder built in sklearn.preprocessing library.

    b. Implement Naive Bayes on mushroom dataset
        i. What's the parameters and performance of your best model? (Baseline: Test accuracy 98%) [10%]
        By using Multinomial Naïve Bayes classifier with **alpha=1e-5** getting Test accuracy of **99.63%**

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

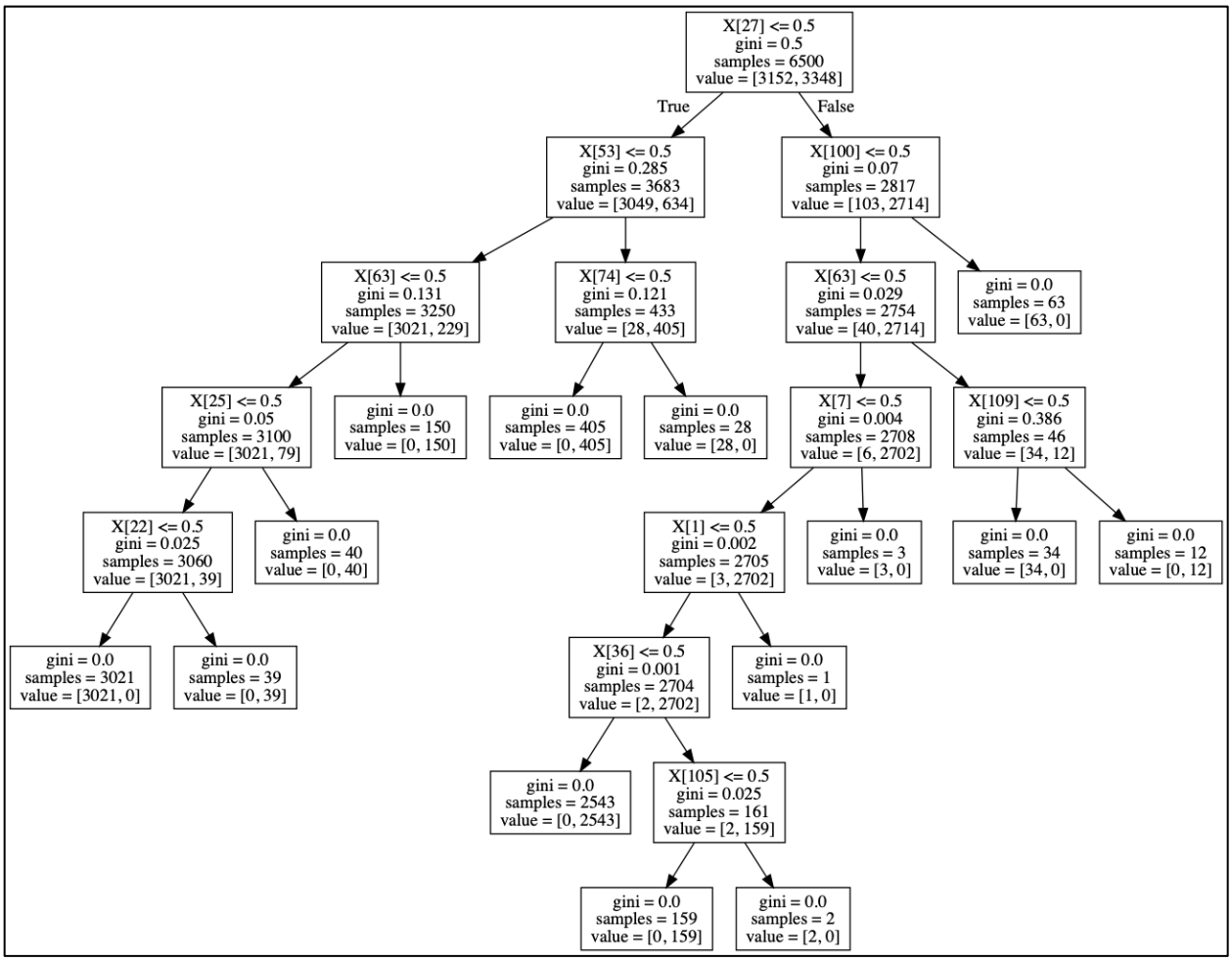
	Gaussian NB	**Multinomial NB**	Bernoulli NB	Complement NB
Acc.	99.20%	**99.63%**	99.32%	99.57%

Multinomial distribution assumption is most suitable for mushroom dataset.

c. Implement Decision Tree on mushroom dataset

i. What's the performance of your best model? (Baseline: Test accuracy 99%) [10%]
By using **max_depth=7** getting Test accuracy of **100%**

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



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%]
Numerical data will be more suitable for Naïve Bayes, and categorical data is best suited for Decision Tree.

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%]
   Since there are numerical and categorical features in income dataset, so I decided to scale numerical feature using MinMaxScaler (sklearn.preprocessing library) and using OneHotEncoder (sklearn.preprocessing library) to encode categorical features.
   For the missing value, I have tried 3 kinds of method.
      1. First method is removing those columns which have missing value.
      2. Second method is encoding the missing value as a new category of the columns.
      3. Third method is replacing the missing value with the most frequent category of the columns.

   It turns out that 3 methods have the almost the same accuracy. I used the second method for the homework.

   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%)
   Decision Tree classifier with **max_depth=7** gets better performance in Income Dataset.
   I randomly split training set into 2:1 (train:test) and uses train set to train a classifier while using test set to evaluate model accuracy, and repeating the process 500 times.

	Gaussian Naïve Bayes (var_smoothing=0.05)	**Decision Tree (max_depth=7)**
Mean of Acc.	80.10%	**85.53%**
Standard Deviation of Acc.	0.003456	**0.002946**