Data Mining HW4

Scikit-Learn

Name: 陳思愷 Department: 台大土木系營建管理組 Student ID: r06521705

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%]

         MultinomialNB(alpha=0.1, class_prior=None, fit_prior=True)

         Performance:
         Accuracy on training set: 0.973
         Accuracy on testing set: 0.891

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

         GaussianNB(default):
         Accuracy on training set: 0.970
         Accuracy on testing set: 0.810

         BernoulliNB(default):
         Accuracy on training set: 0.853
         Accuracy on testing set: 0.810

         MultinomialNB(alpha=0.1):
         Accuracy on training set: 0.973
         Accuracy on testing set: 0.891

         最好的是 MultinomialNB

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

      DecisionTreeClassifier(max_depth = 50, random_state = 42)

      Performance:
      Accuracy on training set: 0.968
      Accuracy on testing set: 0.618

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

目前都是了解參數內容後對參數值進行猜測，並依 performance 去做修正

2. Mushroom Dataset: Testing label is provided
   a. How do you preprocess the mushroom dataset? [5%]
      因為資料中有一些缺失的值，所以利用 Pandas 中 get_dummies 的方法將
      Atrribute 映射到更高為的空間，例如:資料中某一個 attribute 有男，女，?
      那 dummies 就會將原先的一個 attribute 映射到高維成為:
      是否為男 ， 是否為女， 是否為? ，是則為 1，否則為 0
      3 個維度去進行紀錄，保留所有 data 的訊息
   b. Implement Naive Bayes on mushroom dataset
      i. What's the parameters and performance of your best model ? (Baseline: Test accuracy 98%) [10%]
         MultinomialNB(alpha = 0.001):
         Accuracy on training set: 0.992
         Accuracy on testing set: 0.994

      ii. Compare different distribution assumption, which is the most suitable for mushroom dataset ? List the testing accuracy. [5%]
         GaussianNB(default):
         Accuracy on training set: 0.956
         Accuracy on testing set: 0.955

         BernoulliNB(default):
         Accuracy on training set: 0.938
         Accuracy on testing set: 0.945

         MultinomialNB(alpha = 0.001):
         Accuracy on training set: 0.992
         Accuracy on testing set: 0.994
         最好的還是 MultinomialNB

c. Implement Decision Tree on mushroom dataset

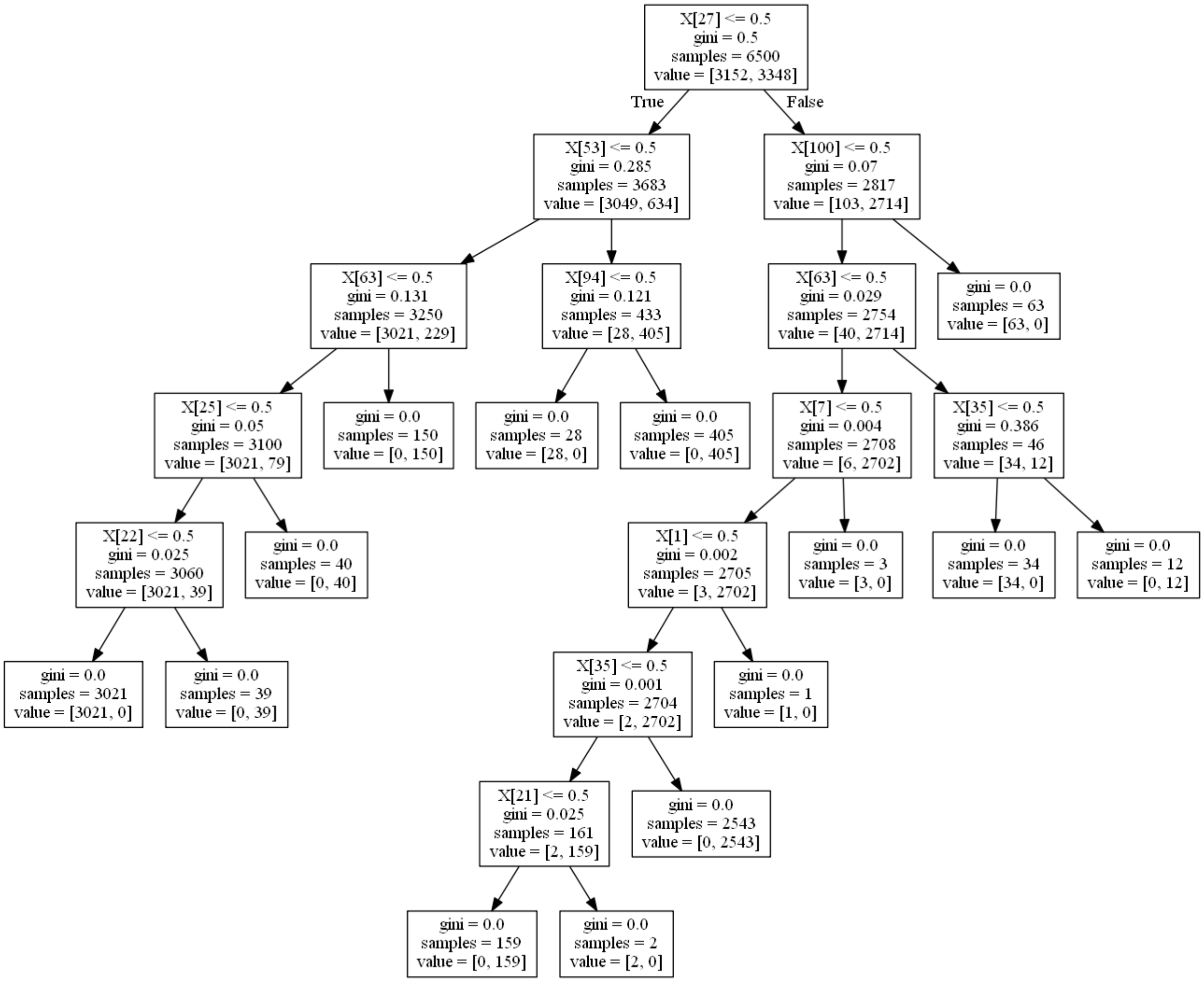
What’s the performance of your best model ? (Baseline: Test accuracy 99%) [10%]

DecisionTreeClassifier(): (全部是 default)

Accuracy on training set: 1.000

Accuracy on testing set: 1.000

i. 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%]

mushroom dataset 比較適合 decision tree
news dataset 比較適合 naive bayes
應該是與資料的稀疏程度有關
Mushroom 的資料經過 dummies 映射到較高為後的 attribute 為 117
News 的資料原本 attribute 就已經高達 23910 項

在過度稀疏的 attibute 且資料量充足的情況下會較適合 naïve bayes

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%]

1.在將資料讀入後將一些相較之下較為不重要的 attribute 給去除掉:
"age", "education-num", "relationship", "race", "native-country", "workclass"

2.爾後一樣在對資料做 dummies 的操作將 attribute 映射到較高為的空間

3.之後再對資料做 normalization (preprocessing.MinMaxScaler())

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%)

資料分割為原先 training set 的 0.8 作為 training 用 0.2 作為 test 用

DecisionTreeClassifier(max_depth = 9, random_state = 42) :
Accuracy on training set: 0.864
Accuracy on testing set: 0.854

MultinomialNB(alpha=1.0)
Accuracy on training set: 0.832
Accuracy on testing set: 0.828

DecisionTreeClassifier 的表現較好