Data Mining HW4

Scikit-Learn

Name: 吳睿哲 Department: 電機所碩二 Student ID: r06921095

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%]

         Ans: 我最好的 model 是使用 Multinomial distribution 的 naive bayes classifier，使用的參數為 alpha=0.1，fit_prior 為 false，accuracy 為 0.8839。

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

         Ans: 最適合的是 Multinomial distribution。

         GaussianNB：0.8552
         BernoulliNB：0.8182
         MultinomialNB：0.8839

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

         Ans: 我最好的 model 的參數是：
         crterion='gini, splitter='best',
         max_depth=24, min_samples_split=0.09,
         max_features=None, max_leaf_node=None,
         min_impurity_decrease=0.0, random_state=1。
         accuracy 為 0.6343。
   c. How do you choose the parameters to get the best model ? [5%]

      Ans: 我參數的找法是先把 training data 分成 8：2，用 8 成的 data 當作 training data，剩下的當成 validation data,將 decession tree classifer 當中可調的參數都取出來,
      使用 gridsearch 的方式,找出在 validation data 中準確率最高的參數當成所選的參數。

2. Mushroom Dataset: Testing label is provided
    a. How do you preprocess the mushroom dataset? [5%]

        Ans: 由於 mushroom dataset 中全部都是 categorial 的 data,所以我全部都直接使用 one hot encoder 的方式處理即可。
        至於 missing value 的部份則直接忽略。

    b. Implement Naive Bayes on mushroom dataset
        i. What's the parameters and performance of your best model ? (Baseline: Test accuracy 98%) [10%]

            Ans: 我最好的 model 是使用 Gaussian distribution 的 naive bayes classifier，使用的參數為 var_smoothing= 0.0002，accuracy 為 0.9908。

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

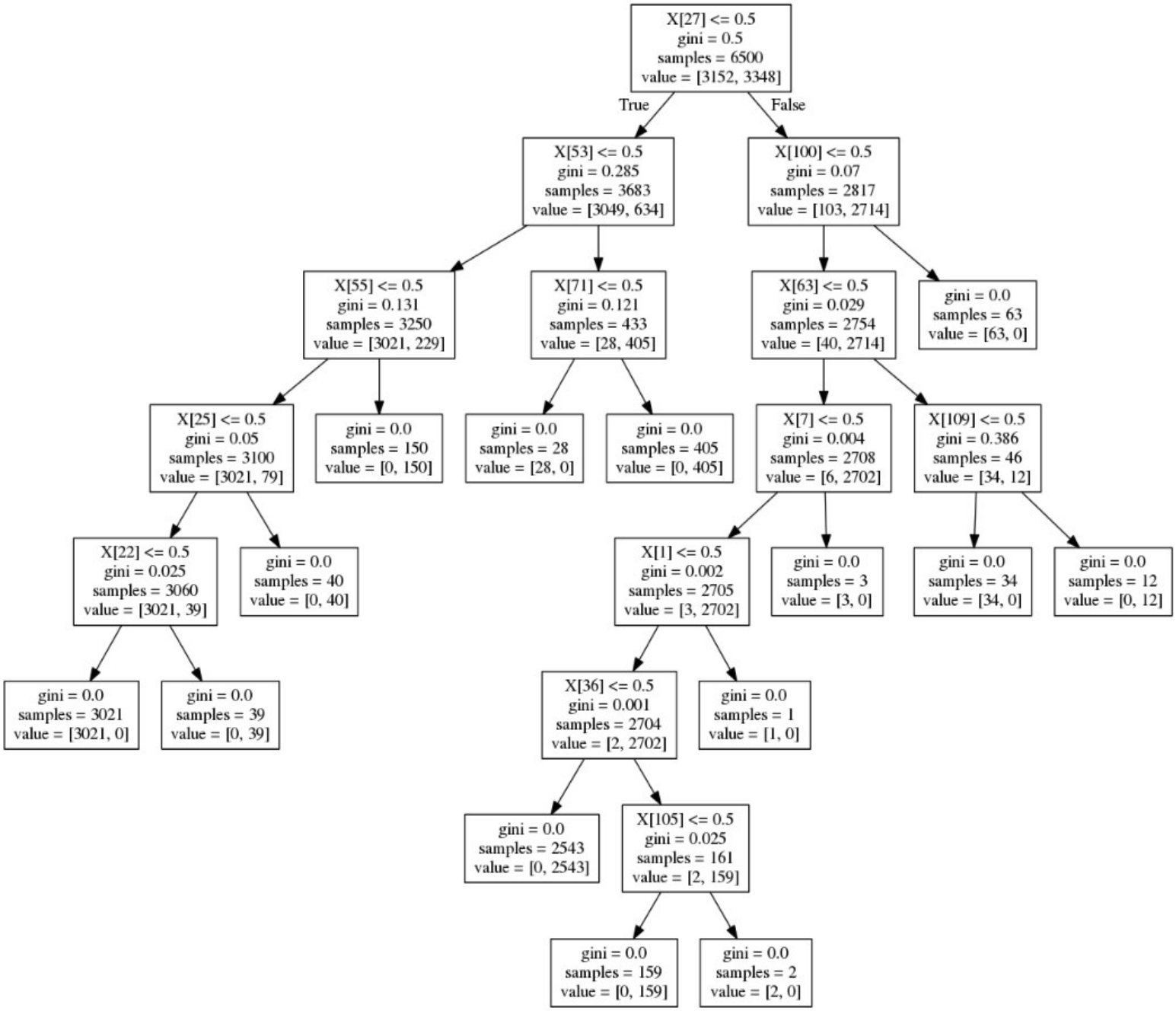
            Ans: 最適合的是 Gaussian distribution。

            GaussianNB：0.9908
            BernoulliNB：0.9446
            MultinomialNB：0.9563

    c. Implement Decision Tree on mushroom dataset
        i. What's the performance of your best model ? (Baseline: Test accuracy 99%) [10%]

            Ans: 我最好的 model 的參數是：
            crterion='gini, splitter='best',
            max_depth=None, min_samples_split=2,
            max_features=None, max_leaf_node=None,
            min_impurity_decrease=0.0, random_state=0。
            accuracy 為 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%]

Ans: 從 performance 來看的話 decision tree(accuracy：100%)比 naive bayes(accuracy：99%)適合。

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%]

      Ans:
      Data Preprocessing:
      income dataset 的 attribute 可以分為 continuous 以及 categorical 的 attribute，categorial 的 data 使用 one hot encoder 的方式處理，continuous 的 attribute 則維持原樣。

      Missing value:
      Missing value 不管是在 training data 或是 testing data 都是發生在第 1、7、14 個 attribute,而且比例都沒有很高，不到整體的 6%,所以我選擇直接忽略。

   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%)

      Ans: decission tree 的 performance 比較好，
      model 的參數是：
      crterion='gini, splitter='best',
      max_depth=17, min_samples_split=0.0177,
      max_features=None, max_leaf_node=None,
      min_impurity_decrease=1e-9, random_state=0。
      使用 10fold validation 的 mean accuracy 為 100%。