[Project Code : MCNB]

Mushroom Classification using Naive Bayes Algorithm

Project Duration : 21st Jan 2024 to 10th Feb 2024
Submission Information : (via) CSE-Moodle

Objective:
In this project, the objective is to classify whether a mushroom is edible or poisonous. The mushroom dataset contains 22 features (e.g. cap-shape, cap-size, odor) of a mushroom and the task is to predict if that mushroom is poisonous or edible.

You have to create a Naive Bayes classifier to predict if a mushroom is edible or poisonous.

In particular, you shall be doing the following tasks:

1. Based on the dataset (described later), you will write a program to learn a **Naive Bayes Classifier.**
2. Compare the results with the results generated by the Naive Bayes classifier learning algorithm from a pre-created package such as scikit-learn.

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used.*

Dataset:
Filename: mushrooms.csv

Data Description: The **attribute** Information is given as follows.

- cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
- cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
- cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
- bruises: bruises=t,no=f
- odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
- gill-attachment: attached=a,descending=d,free=f,notched=n
- gill-spacing: close=c,crowded=w,distant=d
- gill-size: broad=b,narrow=n
- gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
- stalk-shape: enlarging=e,tapering=t
- stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
- stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- veil-type: partial=p,universal=u

- veil-color: brown=n,orange=o,white=w,yellow=y
- ring-number: none=n,one=o,two=t
- ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
- spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
- population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
- habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

Output classes: edible=e, poisonous=p

Your Tasks:

1. The train dataset is not divided into train and validation sets. The first task is to randomly partition the train dataset into train and test sets using 80-20 split. Use the train split for training the tree and test split for testing.
2. Naive Bayes Classifier Model:
    a. Implement naive bayes algorithm in your code and mention the same in the report. ***DO NOT use scikit-learn for this part.***
    b. Compare the results of your implemented model with the Naive Bayes Classifier from scikit-learn package.
3. Classification Report:
    a. Create a classification report in tabular form.
    b. You need to calculate precision, recall, f1-score and accuracy of the model.

Submission Details: (to be submitted under the specified entry in CSE-Moodle)

1. ZIPPED Code Distribution in CSE-Moodle
2. A brief (2-3 page) report/manual of your work
   (with your hyperparameter tuning results also presented in that report)

Submission Guidelines:

1. You may use one of the following languages: C/C++/Java/Python.
2. Your Programs should run on a Linux Environment.
3. You are **not** allowed to use any library apart from these (Also explore all these libraries if doing in Python, or equivalent of these):
    ```
    import numpy as np # linear algebra
    import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import accuracy_score
    from sklearn.metrics import classification_report
    import operator
    from math import log
    from collections import Counter
    from statistics import mean
    ```
   Your program should be standalone and should **not** use any *special purpose* library for Machine Learning. Numpy and Pandas may be used. And, you can use libraries for other purposes, such as generation and formatting of data.
4. You should submit the program file and README file and not the output/input file.
5. You should name your file as <GroupNo_ProjectCode.extension>
   (e.g., Group1_MCNB.pdf or Group1_MCNB..zip).
6. The submitted program file *should* have the following header comments:

    # Group Number
    # Roll Numbers : Names of members (listed line wise)
    # Project Number
    # Project Title

7. Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=561

You should not use any code available on the Web. Submissions found to be plagiarised or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:
Ayan Maity (Email: ayanmaity201@gmail.com)