[Project Code : HSNB]

Patient's Hospital Stay Prediction using Naive Bayes Classifier Learning Model

Project Duration : 22-Jan-2023 ~~ 11-Feb-2023
Submission Information : (via) CSE-Moodle

Objective:
AC Roy hospital is one of the busiest hospitals in town. Hundreds of patients are admitted to and discharged from the hospital every week. A common question that every patient and his/her family members want to get answered is how many days would the treatment last and when can the patient be healthy again. Doctors generally estimate this duration based on the illness the patient is suffering from. However, this estimate is not very accurate as there are other factors like severity of illness, age of the patient, quality of hospital resources, etc. which should also be taken into account.

You have been given the data of around 285,000 patients and 12 categories indicating the duration of a patients' stay. Your task is to build a Naive Bayes classifier which classifies a patient into any one of the given categories. In particular, you shall be doing the following tasks:

1. Randomly divide the data into 80% for training and 20% for testing. Apply the following:
    a. Handle the missing values in both train and test set.
    b. Encode categorical variables using appropriate encoding method (in-built function allowed).
    c. After completing step (a) and (b), compute 5-fold cross validation on the training set (train using Naive Bayes). Normalization of data is allowed, if required. Print the final test accuracy.

2. Apply PCA (select number of components by preserving 95% of total variance) on the processed data from step (1).
    a. Plot the graph for PCA (in-built function allowed for PCA and visualization).
    b. Use the features extracted from PCA to train your model. Compute 5-fold cross validation on the training set. Normalization of data is allowed, if required. Print the final test accuracy.

3. Using the processed data from step (1), apply the following:
    a. A feature value is considered as an outlier if its value is greater than mean + 3 x standard deviation. A sample having maximum such outlier features must be dropped.
    b. Using the sequential backward selection method, remove features.
    c. Print the final set of features formed.
    d. Compute 5-fold cross validation on the training set (normalization of data is allowed if required). Print the final test accuracy.

4. Create a detailed report including the results obtained from Steps 1,2 and 3.

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.*

DataSets:
Filename: `hospital.csv`

Brief Description about the data (You are expected to play with the data to figure out more details):

The **attributes** are as follows.

- *Hospital_code*
- *Hospital_type_code*
- *City_Code_Hospital*
- *Hospital_region_code*
- *Available Extra Rooms in Hospital*
- *Department*
- *Ward_Type*
- *Ward_Facility_Code*
- *Bed Grade*
- *City_Code_Patient*
- *Type of Admission*
- *Severity of Illness*
- *Visitors with Patient*
- *Age*
- *Admission_Deposit*

Output Classes: 0-10, 11-20, More than 100 days, etc.

Tasks to be done:

1. The dataset is not divided into train and validation sets. The first task is to randomly partition the complete dataset into 5 parts: assign the first part as validation set and the rest for training the classifier. Repeat the process 5 times, assigning the validation sets in a round robin manner. *(5 fold cross-validation)*
2. Naive Bayes Classifier Model:
    a. Implement naive bayes algorithm in your code with appropriate feature extraction and normalization (as per the preprocessing mentioned earlier in the problem statement) and discuss the same in the report. Do NOT use scikit-learn for your written naive-bayes algorithm part.
    b. Test the implementation of the Classifier from scikit-learn package (code snippet provided).
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.
    c. Report the average score for the 5 folds.

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

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 # linear algebra
import csv # data processing, CSV file I/O
import pandas # data processing, CSV file I/O
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.model_selection import KFold
from sklearn import naive_bayes # sklearn Naive Bayes
import operator
from math import log
from collections import Counter
```

   Your program should be standalone and should **not** use any *special purpose* library for Machine Learning for the Naive Bayes classifier algorithm. 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., *Group99_CANB.zip* for code-distribution and *Group99_CANB.pdf* for report)
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. Submit through CSE-MOODLE only.
   Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=508

You should not use any code available on the Web. Submissions found to be plagiarized 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:
Suryansh Kumar (Email: suryanshkumar3@gmail.com)