LAB Manual

PART A

(PART A : TO BE REFFERED BY STUDENTS)

**Experiment No.06**

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical slot. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)***

|  |  |
| --- | --- |
| Roll No.: C026 | Name: Anirbaan Ghatak |
| Class : B | Batch : EB1 |
| Date of Experiment: 13/09/2023 | Date of Submission 14/09/2023 |
| Grade : | Time of Submission: |
| Date of Grading: | |

**B.1 Software Code written by student:**

***(Paste your c/c++/java code completed during the 2 hours of practical in the lab here)***

# Name: Anirbaan Ghatak

# Roll no: C026

# Aim: : Implementation of Naïve Bayes Classifier

import pandas as pd

import numpy as np

from sklearn.preprocessing import LabelEncoder

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import train\_test\_split

from sklearn.naive\_bayes import GaussianNB

data = pd.read\_excel("nvb.xlsx")

le = LabelEncoder()

for i in data.columns:

    data[i] = le.fit\_transform(data[i])

X = data.drop(columns=['buys\_computer'])

y = data['buys\_computer']

X\_train,X\_test,y\_train,y\_test = train\_test\_split(X,y)

classifier = GaussianNB()

classifier.fit(X\_train,y\_train)

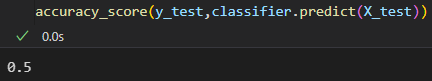
accuracy\_score(y\_test,classifier.predict(X\_test))

**B.2 Input and Output:**

**Input Data:**

****

**Output Data:**

****

**B.3 Observations and learning:**

***Naïve Bayes Classifier is an algorithm that relies mainly on likelihood of a certain attribute. It is mainly used in text classification that includes a high-dimensional training dataset. It simply assumes that the occurrence of a certain feature is independent of the occurrence of other features.***

**B.4 Conclusion:**

***Implemented Naïve Bayes Classifier on the selected dataset understood different evaluation metrics and hyperparameters importance.***

**B.5 Questions of Curiosity**

Q1.What are the issues in classification? Explain each with the help of an example.

Issues in Classification are

Imbalanced Classes: When one class dominates, leading to a biased model. Example: Detecting rare diseases with a dataset mostly consisting of healthy individuals.

Overfitting: Model memorizes training data noise, failing on new data. Example: Spam email filter overly reliant on specific words from the training set.

Underfitting: Model is too simple to capture data patterns. Example: Linear model for complex handwriting recognition.

Multi-Class Classification: Handling multiple classes, especially when some are rare. Example: Categorizing animals into various species.

Curse of Dimensionality: High-dimensional data complicates analysis. Example: Text data with thousands of word features.

Noise in Data: Errors in the data that mislead the model. Example: Medical diagnosis with measurement errors.

Lack of Interpretability: Complex models make it hard to explain decisions. Example: Loan approval system using deep learning.

Mitigating these issues requires data preprocessing, selecting suitable algorithms, tuning parameters, and using relevant evaluation metrics. Techniques like data augmentation, feature engineering, and ensembles can help.

Q2.Summarize all approaches used for classification with their advantages and limitations.

|  |  |  |
| --- | --- | --- |
| Classification Model | Advantages | Disadvantages |
| Logistic Regression | Probabilistic Approach, gives information about statistical significance of features. | The assumptions of logistic regression. |
| K – Nearest Neighbours | Simple to understand, fast and efficient. | Need to manually choose the number of neighbours ‘k.’ |
| Support Vector Machine (SVM) | Performant, not biased by outliers, not sensitive to overfitting. | Not appropriate for non-linear problems, not the best choice for large number of features. |
| Kernel SVM | High performance on non – linear problems, not biased by outliers, not sensitive to overfitting. | Not the best choice for large number of features, more complex. |
| Naive Bayes | Efficient, not biased by outliers, works on non – linear problems, probabilistic approach. | Based in the assumption that the features have same statistical relevance. |
| Decision Tree Classification | Interpretability, no need for feature scaling, works on both linear / non – linear problems. | Poor results on very small datasets, overfitting can easily occur. |
| Random Forest Classification | Powerful and accurate, good performance on many problems, including non – linear. | No interpretability, overfitting can easily occur, need to choose the number of trees manually. |