Introduction to Machine Learning

REPORT

Princy Gautam

(B20BB051)

- Import necessary libraries for performing calculations and reading csv files.
- Download the dataset and read it using pandas library.
- The dataset has nine columns:
    - Pregnancies
    - Glucose
    - Blood Pressure
    - Skin Thickness
    - Insulin
    - BMI
    - Diabetes Pedigree Function
    - Age
    - Outcome
- The first eight are features and the last one (Outcome) is the label. Outcome has two types of labels.
    - 0 – non-diabetic
    - 1—diabetic
- Drop the label column from the dataset and then apply standard scaler on the leftover data frame.
- Initialize variables as follows:
    - X = data
    - Y = Outcome
- Now we split these variables into train and test set. To split train and test set we will import *train_test_split* function. Split in 70:30 ratio.
- After splitting, generate a naïve bayes model on the training set and further make predictions on test set.
- Import metrics for accuracy calculations.
- Evaluate the model through accuracy using actual and predicted values. Check how precisely our classifier identifies a person is diabetic or not.

Implementing Naïve Bayes classifier from scratch.

- Import math, random libraries.
- Define a function to split the data giving the dataset and ratio as parameter. This function divides the dataset into 70:30 ratio
- Initialize a dictionary and group the data rows under each class as yes or no in dictionary.
- Define a function to calculate mean.
- Define a function to calculate standard deviation.
- Define another function which gives mean and standard deviation for the dataset attribute and delete summaries of last class.
- To find mean and standard deviation under each class define a function for it and collect the info in it.

- Since naïve bayes have independent classes thus we use gaussian density function to calculate gaussian probabilities.
- Further defining class probabilities gives us conditional probabilities.
- At last, create a function to make predictions which returns highest probability as prediction to us.
- Define function for obtaining predictions and append them in a list.
- Define function to check accuracy of our model.
- Give the driver code with path to file as filename and then dropping the header of the data frame so as to make our work easier.
- Prepare and test the model by calling predictions function.
- Plot individual histograms and inter-relation histogram.

Results & conclusions

- Naïve bayes is an efficient classifier.
- We got the approximately same accuracy from the inbuilt code and from the scratch code with a minute difference of 0.4

COLAB FILE LINK:

https://colab.research.google.com/drive/1ZR6zJim-vRIE1MAtu1d-IoPVMG3GCH2E#scrollTo=0SXd75jgOUwR