Machine Learning Assignment Group 71

(18CS10021) Hardik Aggarwal and (18CS30040) Sriyash Poddar

Assignment - 2 - Naive Bayes

Procedure :

1. Preprocess Data - Drop the id column as it is not a feature.

2. Fill the missing value in the data, using the mode *viz.* max frequency element of the corresponding feature, and encode the discrete features using an inbuilt function from sklearn.

3. Train the Naive Bayes classifier across 5-folds and report the average accuracy across different folds.

4. Finding out the class probability using the formula :
   $$P(C_i) = (\text{No. of samples classified as } C_i)/(\text{Total no. of samples})$$

5. Find out the probability matrix using the formula :
   $$P(X_j \mid C_i) = (\text{No. of samples classified as } C_i\text{, having feature } j = X_j)/(\text{No. of samples classified as } C_i)$$
   For continuous features, however, we use :
   $$P(X_j \mid C_i) \sim N(m, var);\ m, var = \text{Mean and Variance of feature } j\text{, across all samples classified as } C_i$$

6. For test data, find out the probability of sample $X = (X_{j1}, X_{j2}, X_{j3} \ldots.)$ by the formula :
   $$P(C_i \mid X) = \{[P(X_{j1} \mid C_i)*P(X_{j2} \mid C_i)\ldots]*P(C_i)\}/(\Sigma_i [P(X_{j1} \mid C_i)*P(X_{j2} \mid C_i)\ldots]*P(C_i))$$
   $$\text{Predicted class} = \max_i P(C_i \mid X)$$

7. For performing Principal Component Analysis on the given data, keeping 95% of the variance, we calculated the no. of components with explained variance ratio less just above or equal to 95%.

8. Transform the given data to the found components using an inbuilt function, plot the cumulative sum of explained variance ratios and eigenvalues of the components.

9. Trained the classifier on the transformed data and reported the test accuracy. (Repeat steps 3-6)

10. Remove outliers from the sample, using the following rule: Samples having max feature values beyond +-3*std of a feature are outliers.

11. Using the sequential backward selection process to remove features, i.e., given the accuracy of given features set F, on the test data(= acc_prev) calculate :
    $$acc_i = \text{Test accuracy of features } F\text{-}f_i$$
    Remove feature, where acc_i - acc_prev is maximum.

12. Run the process, starting from all features till the point where no improvement is possible i.e., acc_i - acc_prev is negative for all present features.

13. Trained the classifier on the transformed data and reported the test accuracy. (Repeat steps 3-6)

14. Observations :

Split Number	Accuracy
1	0.5019364833462432
2	0.4965143299767622
3	0.5243996901626646
4	0.5058094500387297
5	0.49147286821705427

Fig (a): Accuracy across different splits

Results : On kfold :-

Average train accuracy = 0.5041800419121046

On Test set :-

Test accuracy = 0.5012391573729864

Part (b) Principal Component Analysis:-

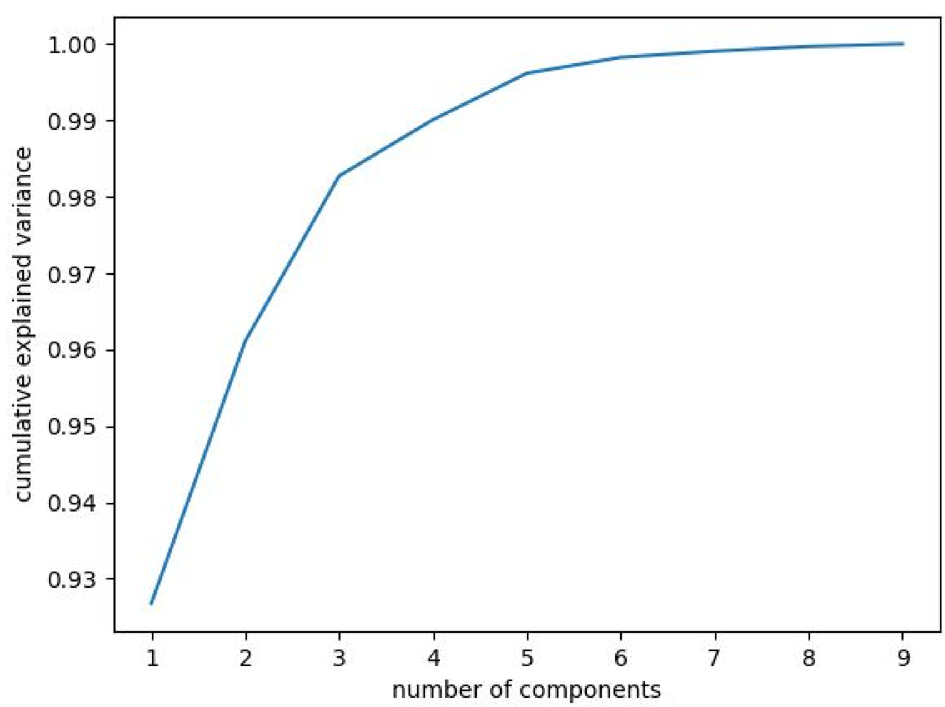

Fig (c) Explained variance plot during PCA.

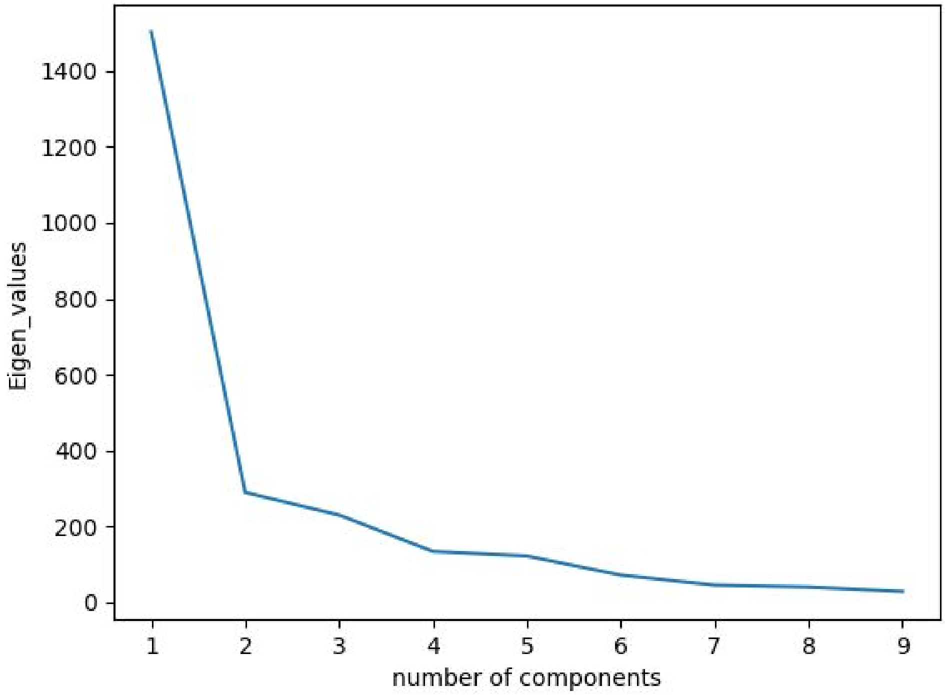

Fig (c) Eigenvalue plot during PCA

Observations after PCA

Split Number	Accuracy
1	0.3996901626646011
2	0.3756777691711851
3	0.418280402788536
4	0.37412858249419056
5	0.3875968992248062

Fig (a): Accuracy across different splits

Results : On kfold :-

Average train accuracy = 0.39107476326866386

On Test set :-

Test accuracy = 0.3990086741016109

Part (c)
(i) Removal of feature outliers :-
-> Samples before removal: 8068
-> Samples after removal: 7927

(ii) Sequential backward selection

Initial features: Gender, Ever_Married, Age, Graduated, Profession, Work_Experience, Spending_Score, Family_Size, Var_1

Remaining features: Ever_Married, Age, Graduated, Profession, Spending_Score, Family_Size, Var_1

Observations :

Split Number	**Accuracy**
1	0.46887312844759654
2	0.5157728706624606
3	0.48974763406940064
4	0.5299684542586751
5	0.47712933753943215

Fig (a): Accuracy across different splits

Results : On kfold :-

Average train accuracy = 0.496298284995513

On Test set :-

Test accuracy = 0.5044136191677175