Bayes Classifier Report

GROUP: 0015
Members : Suchintan Pati (18CS10064) , D. Vamshidhar Reddy (18CS30017)

Procedure

Step 1

1.1 Preprocessing Dataset

The dataset provided had 1177 samples with 24 features and the label to be predicted was life expectancy. The missing values in the dataset were replaced with the mean value of that feature.

The feature ‘date’ was dropped from the dataset as it holds no relevance to the result. The features ‘’iso_code’, ‘continent’, ‘location’ were considered to be categorical and one-hot encoding was done to convert it to numerical data. Hence the number of features increased to 35. Label encoding was not preferred because there is no ordering in the above mentioned features.

The dataset was then split into training data and testing data in the ratio 80:20 and normalisation was done on the train data and the same scaling was used on testing data.

1.2 Implementing Bayes Classifier

The Bayes Classifier was implemented in the following way.

The training data was grouped based on the label (life expectancy) and for each group mean and standard deviation of all the features was computed.

$$\mathbf{P(class|data) = (P(data|class) * P(class)) / P(data)}$$

Each sample in the training data was classified by computing P(class|data). The denominator was neglected as it does not affect the result in any way. Each column was assumed to be independent from the others. Hence the P(data|class) can be computed by multiplying P(value at feature (i)|class) over all the features. P(class) was computed by iterating through the training data.

P(value at feature (i)|class) was computed by using the Gaussian Probability Density Function. The sample was classified into that class which had the highest P(class|data).

1.3 5 Fold Cross-Validation

The training data was further split into 5 parts. In each iteration 4 parts were used as training data and the remaining part was used as validation data. Validation accuracy was computed in each iteration. For computing the test accuracy the model which gave the highest validation accuracy was used.

Step 2

2.1 Principal Component Analysis

The dataset was split into training data and testing data in the ratio 80:20. Principal Component Analysis was performed on the training data and the same components were computed for testing data. The number of components was not specified initially, instead it was computed by observing the Variance Ratio of each component. The number of components to be selected is equal to the minimum number of components selected such that the sum of the variance ratio of each individual component ≥ 0.95.

The graph of Variance Ratio of each component was plotted against the component.

2.1 5 Fold Cross-Validation

The new dataset after performing PCA was used to perform 5 fold Cross Validation which was done as in **1.3**.

Step 3

3.1 Removing Outlier Data Points

A data point was considered to be an outlier if at least 2 of its feature values differ from the mean by more that 3 * standard deviation of that feature. All outlier points were removed from the dataset.

3.2 Sequential Backward Selection Algorithm

The following process was repeated till there exists no feature which when dropped caused an increase in Validation Accuracy. The feature to be dropped was selected by temporarily dropping it and finding the Validation Accuracy. That feature which caused the largest increase in Validation Accuracy was dropped permanently.

3.3 5 Fold Cross-Validation

The set of features to be selected was obtained from the Sequential Backward Selection Algorithm. Then the 5 fold Cross Validation was done as in **1.3.**

Results

Step 1

After segregating the training data (80%) into 5 sets of training and validation sets(5-fold cross validation), the validation accuracy was 100% for all the 5 sets as well the final test set. This indicates the presence of feature(s) that completely describe the class label.

Step 2

After performing PCA, it was determined that 9 components were to be used for the subsequent classification. The 5-fold cross-validation training yielded 100% accuracy for sets 1, 3 4 5 and 98.39% accuracy for set 2. Final accuracy test accuracy obtained also was 100%.

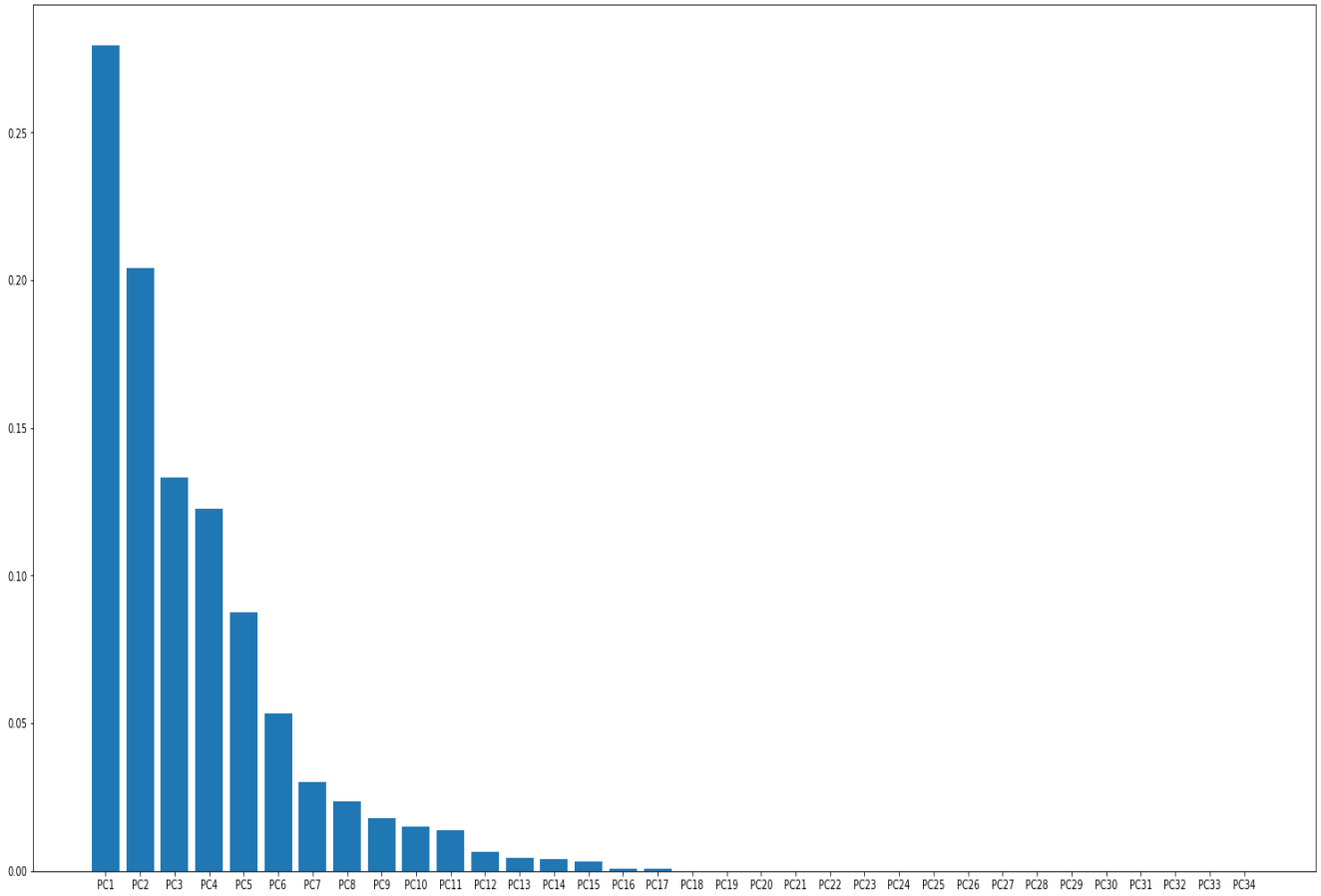

Fig 1: Explained Variance Ratio of each Component vs Component

Since the number of features after One hot encoding is 34 the maximum number of PCA components is 34. The variance ratio after PC17 is very small, of the order of 10^{-5}. The first 9 components variance ratio sum is greater than 0.95, hence the number of components to be used is 9 (PC1, PC2, … PC9).

Step 3

As per the procedure of outliers removal, 43 samples from train set and 12 samples from test set (in total 55 out of 1177) were removed.

On executing the Subsequent Backward Selection (SBS), all features except “diabetes-prevalence” were dropped. On performing 5-fold cross validation training, we obtain the same accuracies as obtained in step 1,i.e 100% accuracy on all sets and final test set.

Thus we have reduced (or shortened) our hypothesis without affecting the performance, which is in accordance with Occam’s razor principle.