Lab Report

Kaijie.Yang (2256526)

TAs name: TA: Xin Gao, Changwei Li, Yizhou Tan

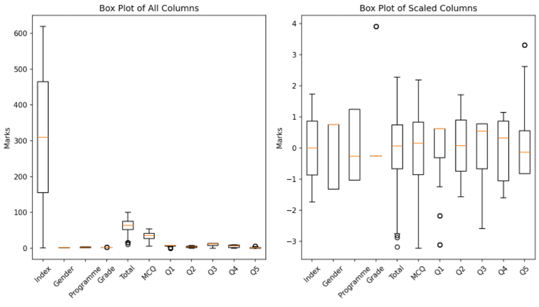
1 INTRODUCTION:

This experimental report studies the correlation between students' information and their final exam scores on programme.

The influence of different features in the survey data set on classification programme. The results of pca calculation are presented through the scatter plot, so as to judge whether there is a good classification result. Since no classifier is added in the experiment, the results of different programme will be presented in different colors. At the same time, in order to present the original data set and the standardized results, the box diagram is

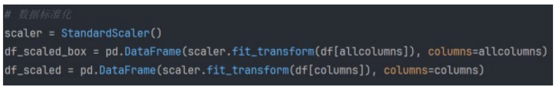
2 METHODOLOGIES:

A: Observe the distribution of raw data



（Figure 1.）

At the beginning of the experiment, we make some simple observations on the raw data. After some observations (Figure 1, left), we find that the numerical span of the Index is much larger than the other data. This makes data observation doubly difficult, so we standardized the raw data (Figure 1 right)。



```
# 数据标准化
scaler = StandardScaler()
df_scaled_box = pd.DataFrame(scaler.fit_transform(df[allcolumns]), columns=allcolumns)
df_scaled = pd.DataFrame(scaler.fit_transform(df[columns]), columns=columns)
```

Standardization process:

Centralization: The data is first centralized, that is, the value of each feature is subtracted from the mean of that feature, so that the mean of the data is 0.

Scaling: The centralized data is then scaled, that is, the value of each feature is divided by the standard deviation of that feature, so that the standard deviation of the data is 1. Scaling can eliminate scale differences in the data, ensuring that the values of different features are within the same scale range.

Through the standardization process, we can see that the mean value of the data changes to 0 and the standard deviation is converted to 1, eliminating the deviation of the data and making the data distributed symmetrically around the origin.

The standardized data is clearer and easier to observe. The box diagram represents its data.

There are the following elements for a box diagram

Box: The box represents the middle 50% range of the data, that is, the second quartile of the data (Q2, the median) is in the middle of the box.

Median Line: The horizontal line inside the box represents the median of the data, which is the middle value of the data. It shows the central location of the data.

Upper limit and lower limit (Whiskers): The upper limit and lower limit represent the range of the data.

Outliers: Outliers in a boxplot represent outliers in the data, i.e., data points that are significantly different from most data points.

Line segments at the top and bottom of the box: These line segments usually represent the maximum and minimum values of the data.

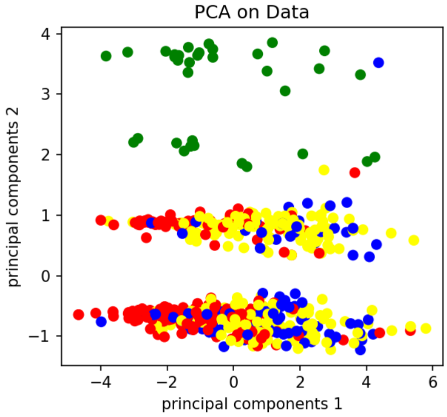
B: Principal Component Analysis (PCA) to the data

The key of this step is to use pca to extract principal components with greater weight from multidimensional data to disperse the data and make the differences between the data stand out. Principal Component Analysis (PCA) is one of the most widely used data dimensionality reduction algorithms. The main idea of PCA is to map N-dimensional features to K-dimensional features, which are new orthogonal features, also known as principal components, and reconstructed K-dimensional features on the basis of the original N-dimensional features.

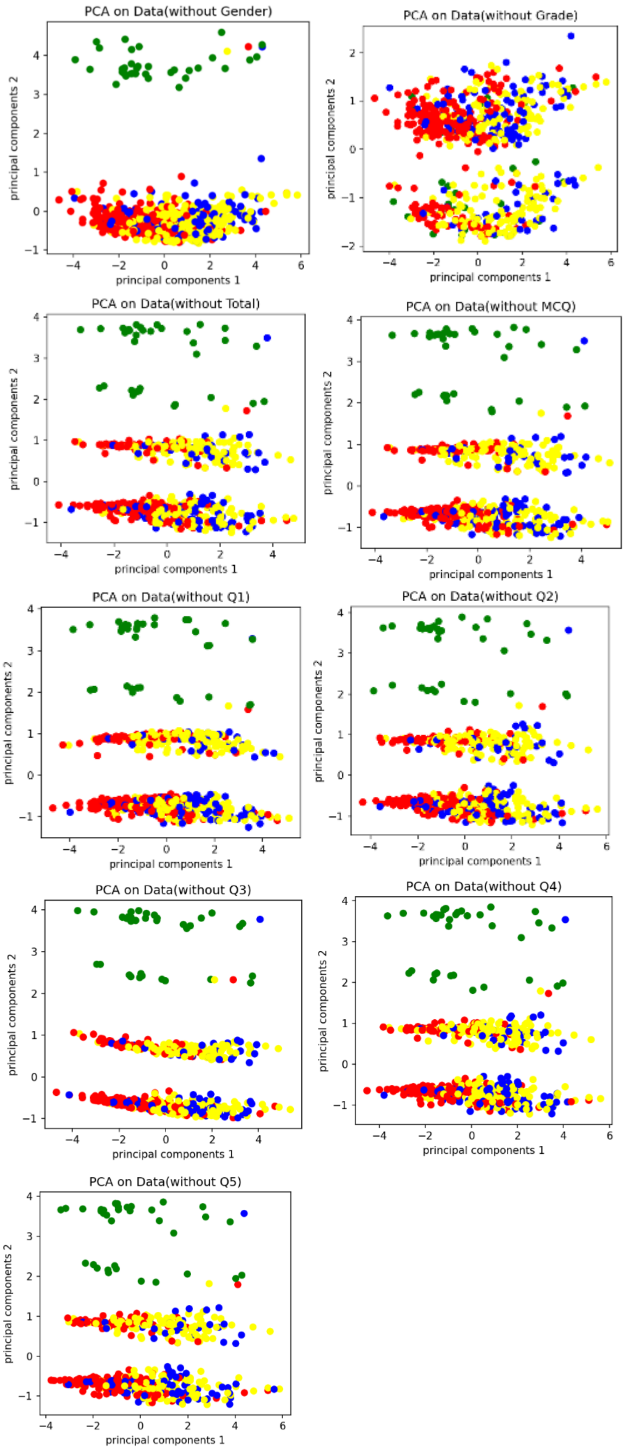
The exact mathematical process is as follows. Suppose we have a dataset X with m samples and n features, we can represent it as an m by n matrix. First, we compute the mean vector μ of the data, and then centralize the data set, that is, subtracting the mean of each feature to get the centralized data set X_c. Next, we compute the covariance matrix Σ of the centralized data set X_c. Then, we decompose Σ to get the eigenvalue λ and the corresponding eigenvector v. Finally, we choose the eigenvectors corresponding to the largest k eigenvalues as principal components, and project the data onto these principal components to achieve dimensionality reduction.

For the data set in the experiment, in order to highlight the features of each data, it becomes important to select the appropriate features for principal component analysis. Suitable feature selection can make the data points in the scatter plot after pca better dispersed.

According to the experimental requirements, we need to classify the data according to the programme. Label different programme in different colors. The programme from smallest to largest are red, blue, green and yellow. At the same time, it is common sense that index is a feature unrelated to data classification, so it is deleted by default. The following figure shows the results of pca analysis without index and programme.



As shown in the figure, we can observe that the data points are concentrated in 4 parallel lines. This is clearly the appropriate outcome. Then, we manually change the input data processed by pca. For features, we manually remove some features. Remove one feature at a time, keep a single variable, and keep the number of principal components at 2. Several experiments were carried out and the results were obtained as shown in the figure below. The deleted features are represented in the header of the scatter plot.



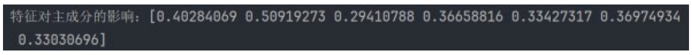
After many comparative experiments, we can conclude that gender and grade have a greater impact on image classification. They affect the distribution of the scatter plot, making the data points free and difficult to separate, so they are removed.

C: Feature weight judgment:

In order to verify whether the feature selection of our multiple experiments is correct, we introduce a method to output the proportion of the influence of each variable on the principal component generation during the PCA operation. Therefore, it is possible to know the importance or weight of each feature in the PCA transformation.

Therefore, better feature selection results can be obtained. The output is the weight value (absolute value) of each feature in each principal component. These weight values represent the importance of each feature in the corresponding principal component, and the greater the value, the greater the influence of the feature in the principal component. The output results list the weights for each feature in each principal component.

Figure 2 selection is characterized by [' MCQ ', 'Total', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5], in addition to the' Grade 'and' Gender '. Where the value is the effect of the feature on the principal component. It can be clearly found that the influence of the first two features on the principal component, one is 0.40284069 and the other is 0.50919273, is significantly greater than the other features. Therefore, it is thought that the two features of 'MCQ' and 'Total' make the classification of programs easier.

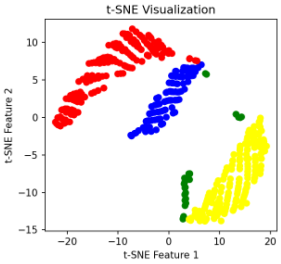


```
特征对主成分的影响: [0.40284069 0.50919273 0.29410788 0.36658816 0.33427317 0.36974934
 0.33030696]
```

(Figure 2.)

D: t-sne method verification:

The above experiments used the principle of pca to reach a conclusion, and then the experimental method was changed to t-sne. The same experimental steps were used to conduct a single variable test (no process picture was shown because of the space problem), and a good data classification was obtained in Figure 3.



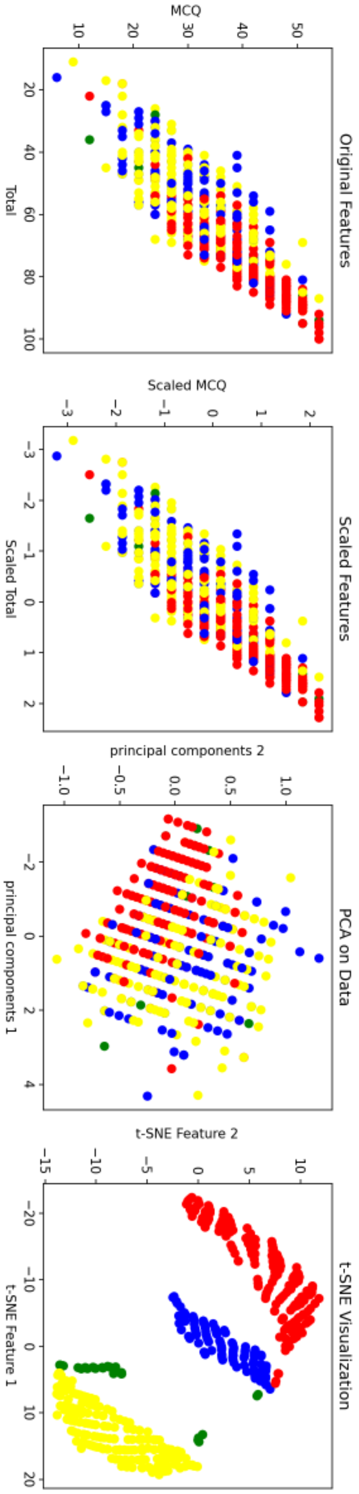
(Figure 3.)

You can clearly see that Figure 3 categorizes the data very well. The input features that produce this result are 'MCQ' and 'Total'. Therefore, it is also proved that 'MCQ' and 'Total' are good features for classifying data.

For t-sne objects, some parameter adjustments are made to make the results more obvious and clearer. First, the target dimension of dimensionality reduction is controlled to be 2-dimensional, and PCA initialization is used to speed up the optimization process. Control the influence range of local structure in t-SNE, reduce the influence of abnormal data on the distribution, and reduce the learning rate, and the control effect is good at 5.

3 Conclusion

Therefore, experiments from two models have gradually demonstrated our conjecture: MCQ and Total can distinguish the characteristics of program classes better. However, there are still some problems in the certainty of the classification, and the classification is only based on dimensionality reduction. In the final figure 4, visualize and compare raw features, scaled features, PCA features and your resulting features.



(Figure 4.)