INT104 CW1-Lab report

Ruiyang.Wu sID:2257475 E-Mail:Ruiyang.Wu22@student.xjtlu.edu.cn
TA: Yu Kang E-Mail:yu.kang23@student.xjtlu.edu.cn

I. Introduction

In this experiment, I first obtained the distribution state and correlation of the data through exploratory data analysis (EDA) of the raw data, and also used this as the basis for proposing hypotheses and proofs of the relationship between feature patterns and labels of some data. The raw data were then subjected to noise reduction and normalization to reduce the impact of noise and differences in data scale on subsequent operations. Then PCA dimensionality reduction was performed on the data as required, and by analyzing the distribution of the data on different feature vectors, the combination of vectors that best distinguished the original dataset was found, and the hypotheses proposed in the data observation were also verified. Finally I used my own way of extracting features from the raw dataset and comparing them with each of the previous features to each other. **To ensure reproducibility of the experiments**, the code for this paper has been placed in a GitHub repository: https://github.com/MushihimePepsi/XJTLU_ICS_Y2S2_Course-notes_23-24/tree/main

II. Task1-Data Observation, Cleaning & Normalization

There were a total of 619 samples in the raw dataset with no missing or incorrectly recorded data. Each tuple included the student's index, gender, grade, major (as label), and scores. To eliminate the effects of data scales and outliers on the analysis, the raw data were first standardized using Z-Score before plotting box plots based on different majors [Fig. 1]. I found that all Programme3 students belonged to grade3, while only a few students in other majors belonged to grade3. **Conjecture that Programme3 could be separated first based on the features of grades**. Furthermore, **in 'MCQ', 'Q1 ', 'Q4' features, the distribution of different majors varies greatly**, which may serve as potential separation features.

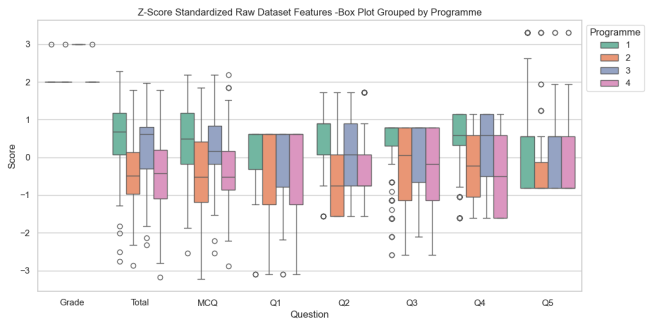

Fig. 1. standardized box plots grouped by 'Programme'(analysis distribution)

Since many statistical methods are not applicable to non-normally distributed data, normality test is required for continuous type variables. Due to small sample size(<2000), Shapiro-Wilk test was used here and all the variables were found to have p-value far less than 0.01 [Fig. 2] and **were not Gaussian distributed**.

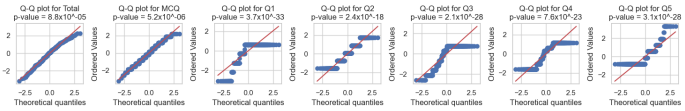

Fig. 2. Q-Q plot and p-value (Shapiro-Wilk test) of each feature

Due to the non-normal distribution of the data, I explored the monotonic correlation between features using the Spearman correlation coefficient instead of the Pearson correlation coefficient. It can be found that in the correlation between features, **'gender' and 'grade' have very little monotonic correlation with the score items ('Total' 'Q5')**; while the **high monotonic correlation of 'Total' with other score features** is due to the fact that it comes from a **linear combination** of the scores. Therefore, 'Total' can be considered to remove from the feature set.

Fig. 3. Spearman correlation between features

Due to the nonlinear relationship between features and labels, I used the maximum mutual information coefficient (MIC), which is based on information entropy, to measure the degree of correlation between each feature and Programme. It can be found that Grade has the greatest correlation with major, which is caused by the fact that **all Programme3 students are grade3**. Whereas, the MIC of 'Gender' is extremely low, so it **can be considered to be removed from**

the feature set. Also, the larger MICs for 'MCQ', 'Q1', and 'Q4' **validate the earlier conjecture that they may serve as separating features**.

TABLE I
MIC BETWEEN LABEL AND EACH FEATURE

Grade	Total	MCQ	Q2	Q4	Q5	Q3	Q1	Gender
0.280	0.252	0.147	0.128	0.106	0.100	0.084	0.068	0.047

III. TASK2-PCA DIMENSIONAL REDUCTION

Based on the above conclusions, I select seven features ['Grade','MCQ','Q1','Q2','Q3','Q4','Q5'] as inputs for Principal Components Analysis (PCA) dimensionality reduction, and the two feature vectors explaining the largest variance ratios are selected as the direction of dimensionality reduction. With 54.4%[Fig. 5] of the total variance explained in these two directions which can **separate Programme3 well** in principle component 2, and it can **also be found that the samples of Programme1 tend to be in the negative half-axis of principle component 1 while the samples of Programme4 tend to be in the positive half-axis of it** [Fig. 4].

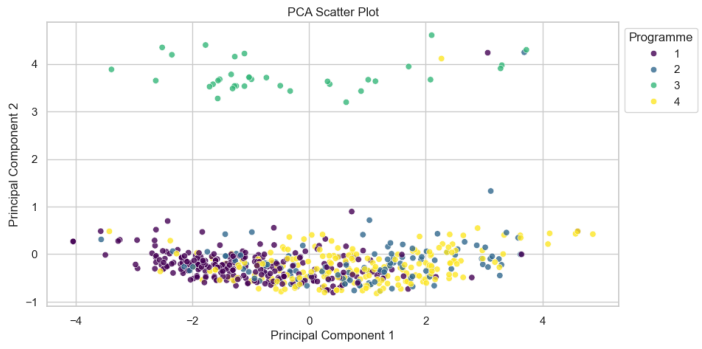

Fig. 4. PCA input=[Grade,MCQ,Q1,Q2,Q3,Q4,Q5] largest 2 eigenvectors

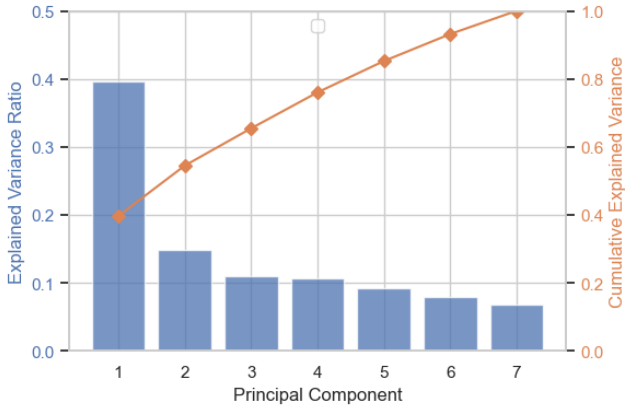

Fig. 5. Explained Variance Ratio & Cumulative Explained Variance

IV. TASK3-IT'S MYGO TO FIND RESULTING FEATURES

The correspondence between Grade=3 and Programme=3 can be statistically derived. In the dataset, Grade=3 was used as a predictor for predicting Programme=3:

TABLE II
PREDICTING 'PROGRAMME'=3 WITH FEATURE 'GRADE'=3

Confusion Matrix	**True Programme=3**	**True Programme≠3**
Pred. Programme=3	TP=35	FP=3
Pred. Programme≠3	FN=0	TN=581

TABLE III
PERFORMANCE OF THE CLASSIFICATION BY 'GRADE'

Precision	Recall	Accuracy	F1 Score
92.11%	100%	99.52%	95.89%

The performance of this classification is satisfactory. So we can **remove the feature 'Grade'** (the rest of the samples all belong to 'Grade' = 2). My subsequent feature extraction will focus on finding the feature space that separates Programme=1,2,3. I remove the Grade features and also remove the samples with Grade=3 from the dataset and **delete the outliers in the dataset by Isolation Forests**, then the **input features are ['MCQ', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5']** and there exist 400 samples (as visualized by t-SNE [Fig. 6]) It can be found that **the Programme1 and 4 have tendency of different distributions**.

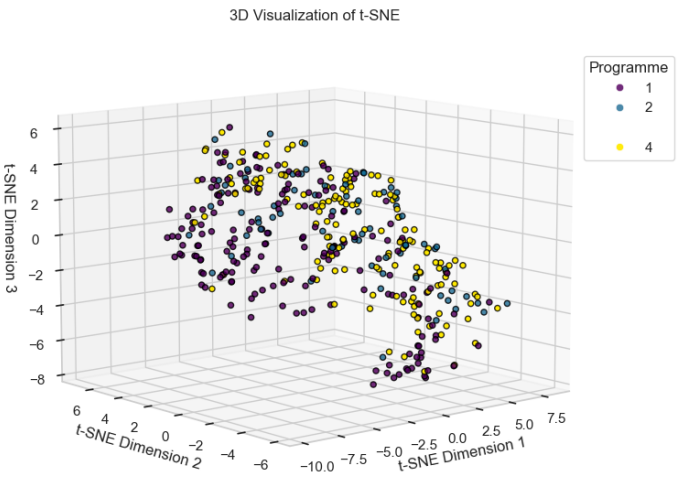

Fig. 6. My chosen feature space visualized via t-SNE

After comparing various dimensionality reduction methods, I finally used Kernel Principal Component Analysis (KPCA) in the case of t-distributed Stochastic Neighbor Embedding (t-SNE) (which reduced the original feature space to three dimensions) in an attempt to find the high-dimensional feature vector which maximizes the variance of the sample points to separate them. I used polynomial function [Fig.7] or Radial Basis Function (RBF) [Fig.8] as kernel function and performed parameter optimization to get the resulting features, it can be found that the two-dimensional feature distinguishes Programme1 and 4 more clearly, but classifies Programme2 poorly [Fig.7].This is the result of purely unsupervised learning, which already gives a two-dimensional feature space that tends to be approximately linearly divisible, with the features of new feature space coming from the nonlinear combination of the

raw features.

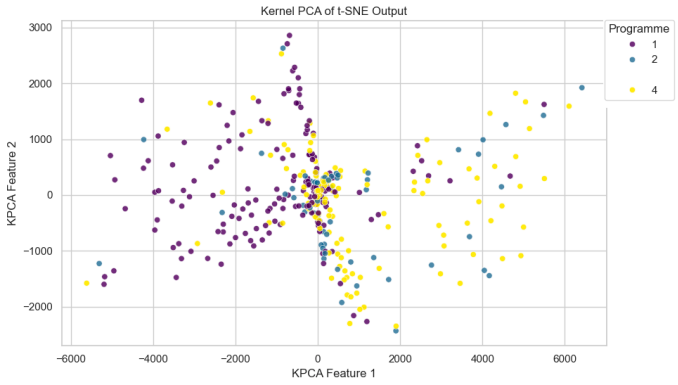

Fig. 7. KPCA using polynomials as kernel function

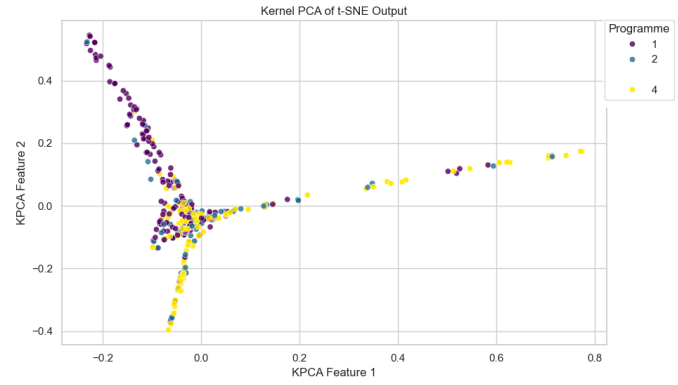

Fig. 8. KPCA using RBF as kernel function (optimized)

V. Task4-Conclusion & Critical thinking

A. Summary of the proposal of scientific hypotheses and proofs

In this experiment, **hypotheses are proposed and proved** to characterize the distribution patterns of several features and labels through exploratory data analysis of the original data. The correspondence between Programme3 and Grade3 is proved and used to reduce the dimensionality; the 'MCQ', 'Q1', and 'Q4' features are proved to be highly correlated to the 'Programme' and they are chosen as potential separation features; demonstrating the irrelevance of 'Gender' and using this to reduce dimensionality; and demonstrating that the high correlation between 'Total' and other scores is due to the fact that 'Total' comes from a linear combination of other scores and using this to reduce dimensionality.

B. Summary of comparison of raw /scaled /PCA /my resulting features

Relative to the original features, the scaled feature removes the effect of scale and appropriately removes low correlation or redundant features (Gender, Total) through statistical manipulation.The PCA feature finds a way to categorize Programme3 by transforming the coordinate system. My features went a step further by noise-reducing the data, attempting to reduce the feature space of the data to two dimensions through a combination of manifold learning and kernel functions.

C. Critical Thinking and Current Shortcomings

In my resulting features, the classification of 'Programme'=2 was poor. To solve this problem, I conjecture that nonlinear dimensionality reduction methods based on supervised learning should be used, such as combining Linear Discriminant Analysis (LDA) with kernel functions to find features of the sample points that differentiate between different LABELS, rather than features that maximize the variance of the sample. I also conjecture that the feature space that maximizes the ability to distinguish students' majors will come from a nonlinear combination of the original features.

D. Statement of my final feature selection

Overall, I will choose ['Grade', 'MCQ', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5'] as the features for subsequent classification. I discarded 'Gender' due to the low relevance of this feature for categorization (MIC); removing 'Total' due to the fact that this feature can be obtained from a linear combination of other features. I kept 'Grade', due to the strong correlation demonstrated above, which can be used to categorize Programme3 by this feature; and keeping the rest of the scores, as they still show a strong correlation to labels.

VI. Acknowledgment

I would like to thank the two TAs of SC375: Yu Kang and Yiqiang Cai for their detailed comments and guidance.

I would like to thank Dr. Shengchen Li for his careful design of the experiments in this course and his discussions with me in the classroom, which helped me to gain a deeper understanding of traditional machine learning, as well as an interest in the study of artificial intelligence. Thanks to Dr. Liu and Dr. Kang for their lectures and accepting me to participate in Dr. Liu's research.