Math for Ai

Assignment 1

By Joel Poah

Class DAAA/02

2112729

Q1

1. I would use a correlation matrix as the scale of the variables have different units of measurements. Vitamin A or ‘Vit\_A’ is in microgram while other variables are in grams. Coverting ‘Vit\_A’ to gram will make it too small and thus, insignificant as compared to the other variables in the Principal Component. In this case, we have no choice but to scale it and use correlation matrix to carry out PCA.

Graphical user interface, text

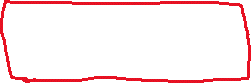
Description automatically generated



I arrived at the conclusion that we should extract 2 PCs through the following 3 reasons. The 1st reason is based off Kaiser’s rule, which states that we should retain PCs whose eigenvalues are greater than 1. PC 1 and PC 2 have eigenvalues that are significantly larger than 1. Meanwhile PC 3’s eigenvalue is not that close to 1 either.

Graphical user interface, text

Description automatically generatedSecondly, we can also look at the percentage total variance to understand how many PCs is needed to explain most of the variability of the dataset. Since we are carrying out PCA to further analyze the food\_nutrition dataset, we will need a cumulative percentage of at least 80-90%. The cumulative explained variance of the first 2 PCs is 82.6% but the cumulative explained variance for the first 3 PCs would be 89.5%. It isn’t clear how many PCs we should extract out so I would decide after looking at the 3rd reason.



Chart, line chart

Description automatically generated

The 3rd criteria I could look at would be the scree plot and to look out for the inflection point where the graph begins to ‘flatten’. I notice how the graph is significantly flatten on the 3rd Component number which suggests that we use only the first 2 PCs. Using the combination of the 3 methods, the majority suggests that we should only take the first 2 PCs.

(c) (i)

| Here are the columns of the food\_nutrition dataset^ |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |

PC 1

z6

For PC1, the loadings on Protein and Fat are opposite in sign as compared to other loadings. This PC seems to measure a weighted average of nutrients in Food products that contrasts against protein and fat

For PC 1 to be positive, the values of Protein and Fat must below average, and the values of Carbohydrate, Sugar, Fiber and Vitamin A have to be above average.

For PC 1 to be negative, the loading of Carb Sugar, Fiber and Vitamin A will have to be below average while having an above average of Protein and Fat.

PC 2



z6

The loadings for all the nutrients in PC 2 are all positive. This PC seem to measure a weighted average of nutrients in food which is high in protein and fat.

For PC 2 score to be positive, the values of Protein and Fat must be above average and the other values of Carbohydrates ,Sugar , Fiber and Vitamin A to be at least average or above average.

For PC 2’s score to be negative, the values of Protein and Fat must be below average and the other values of Carbohydrates , Sugar , Fiber and Vitamin A should be at least average or below average.

(ci) For low protein and low fat and high in other nutrients,

Chart, line chart

Description automatically generated



As shown on the loading plot above , for low fat and low protein it would be pointing at the 4th quadrant of the graph which is the negative of PC2. Furthermore, with high nutrients , it points towards the 2nd and 4th quadrant of the graph which is the positive side of the X-axis or PC1.



Chart

Description automatically generated

Furthermore, looking at the score plot,

if the meaning of low protein and low fat means that it is below average(assuming 25% quartile range for protein and fat) and assuming 75% quartile range for Carb,Sugar,Fiber and Vitamin A , then the PC2 value would be negative which results in the type of food being ‘Fruits and Fruit Juices’ and ‘Vegetables and Vegetable Products’ .

(cii) For high protein and high fat but low in other nutrients,

Based on PC1 it would be a negative value for x-axis

And based on PC2 it would be a positive value for y-axis

Thus, the type of food would be ‘Poultry Products’

Graphical user interface, text

Description automatically generated

d) Let Food we are predicting be Food\_X ,

A picture containing text, gauge

Description automatically generated

Since, there was no population Standard Deviation I will be taking the sample standard deviation instead.

Graphical user interface, text

Description automatically generated with medium confidence

Using data\_reduced.describe() I can see the standard deviation .

Using the Z – score formula , I calculated each of the Z – value for each of Food\_X attributes.

Text

Description automatically generated

I then calculated the PC1 for food X by taking the loading values multiplied by its z-score value and the same goes for PC2

Text

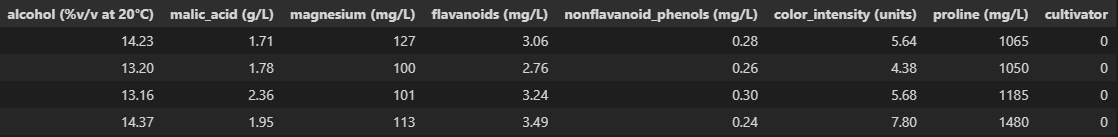
Description automatically generated

Chart, line chart

Description automatically generated

I then plotted the following PC1 and PC2 values into the score plot. The food type should be Fruits and Fruit juices.

Q2)



1. I would use a correlation matrix as the variables have different units of measurements. Alcohol is measured in volume percentage, color intensity is in units and the others is in g/L or mg/L. Furthermore, proline have too huge a difference in scale. Proline will be too big as compared to the other attributes. In this case, we have no choice but to scale it and use correlation matrix to carry out PCA.
2. Based on the Cumulative Explained Variance where we will retain at least 80% explained variance, we will need to use the top 3 PCs: PC 1, PC 2 and PC 3

A picture containing text, road, screenshot

Description automatically generated

Graphical user interface

Description automatically generated

PC 1

z6 + 0.5080z7

For PC1, the loadings on malic\_acid and nonflavanoid\_phenols are opposite in sign as compared to other loadings. This PC seems to measure a weighted average of constituents in wine of different cultivator.

PC1(positive), the values of malic\_acid and nonflavanoid\_phenols will have to be below average and above average for alcohol,magnesium,flavonoids,color\_intensity and proline.

PC1(negative) the values of malic\_acid and nonflavanoid\_phenols will have to be above average, while the other constituents like alcohol,magnesium,flavonoids,color intensity and proline will have to be below average as well

PC 2

z6 + 0.1209z7

For PC 2

The loadings on flavonoids are in contrast to the other loadings. This PC seems to measure a weighted average of constituents in wine of different cultivator.

For PC 2 (Positive), the values of flavonoids have to be below average which makes it positive and the other constituents will have to be above average

For PC 2 (Negative), the value of flavonoids will have to be above average and the other constituents will have to be below average.

PC 3

z6 - 0.2403z7

The loadings on alcohol , nonflavanoid\_phenols , color\_intensity and proline are in contrast to the other loadings. The PC seems to measure a weighted average of constituents in wine of different cultivator contrasting against alcohol,nonflavanoid\_phenols, color\_intensity and proline.

For PC 3(Positive) , the values of alcohol,nonflavanoid\_phenols, color\_intensity and proline will have to be below average and the other constituents will have to be above average.

For PC 3 (Negative) , the values of alcohol,nonflavanoid\_phenols, color\_intensity and proline will have to be above average while the other values will have to be below average.

2c) Let unknown wine be Alcohol\_X

Chart, scatter chart

Description automatically generated

From the graph I plot using PC1 and PC 2, I could predict that the cultivator of Alcohol\_X is of cultivator 2 as it is much nearer to the cluster of cultivator 2.

Chart, scatter chart

Description automatically generated

Looking at the plot PC2 against PC3 , It was nearly impossible to tell and predict the cultivator as it was in the middle between Cultivator 0 and Cultivator 2.



Chart, scatter chart

Description automatically generated

Looking at the plot PC1 against PC3, it was impossible to tell which cluster it belonged to, the cultivator 1 and cultivator 2 was clearly mixed up and there was no way to distinguish them. However, the plots individually may not help distinguish the cultivator, but all plots point to 1 common cultivator which is cultivator 2.



2d) I think overall PC3 is useful in helping me identify the cultivator in (2c) as the plot PC1 against PC2 may be able to show that the wine is closer to the cultivator 2 cluster. However, there is still a chance that the point may belong to cultivator 1.0. Looking at the combination of the plots PC2 against PC3 and PC3 against PC1 they both suggests that the unknown alcohol should belong to cultivator 2. Thus, this give me a greater assurance that the unknown alcohol is cultivator 2 and that PC3 was useful in providing more information on the unknown alcohol.