Danmarks
Tekniske
Universitet

DTU

02450 Introduction to Machine Learning and Data Mining

Project 1

Authors

Group 116

Cheng-Liang Lu - s220034
Jennifer Fortuny I Zhan - s230705
Emma Louise Blair - s214680

September 27, 2023

Responsibility table

Overview of each student's responsibility for the report.

	Cheng-Liang	Jennifer	Emma
Introduction & Description	90%	5%	5%
Attribute explanation	5%	5%	90%
Data visualisation	5%	90%	5%
Discussion	5%	90%	5%
Question 1	5%	5%	90%
Question 2	5%	90%	5%
Question 3	5%	90%	5%
Question 4	5%	5%	90%
Question 5	90%	5%	5%
Question 6	90%	5%	5%

The **GitHub link** to find our work.

Contents

1 Introduction

We are interested in the different income ranges in our society. The reasons effect our incomes, causing us to have a level difference community. Hence, the data records the number of people whose yearly incomes surpass $50K and the opposite. Their different backgrounds' data may lead us to discover the grounds for this phenomenon. In order to find the pattern, we would like to utilize these attributes from the data set to try to find a model that will be able to predict if a person's income could exceed $50K/yr by the same type of attributes we used for training.

2 Description of our data set

We will analyze the data set "Adult" [1] to determine whether our primary machine learning aim appears to be feasible. Our primary machine learning aim is to be able to predict whether income exceeds $50K/yr.

The previous analysis visualized the data attributes first to see if the attributes had potential biases or were unbalanced inside. After that, they selected the right attributes to be the data set to predict. They used five different models to see which one fit the data the best. The models also include Bayes machine learning models, which compare the traditional and Bayes ones. In a nutshell, they found Decision Tree model performed best both in training and testing.

The attributes we use for regression are the continuous data we possess. *age*, *edu-num* and *hours-per-week* will be the critical values for the regression model since continuous attributes are well-suited for capturing and modeling relationships between input variables and the continuous target variable. They can represent a wide range of numeric values, allowing the model to account for subtle variations and patterns in the data. They allow for flexible modeling. Many regression algorithms, such as linear regression and decision trees, work naturally with continuous features. They can adapt to different types of relationships, including linear, nonlinear, and complex patterns.

On the other hand, all the features we own will be used for the classification task as it is the main goal for our interest, predicting whether a person's income will surpass $50K/yr or not. Moreover, the properties of our continuous attributes are pretty suitable for the classification model. Divide these attributes by different ranges and make them become classified attributes.

3 Detailed explanation of the attributes of the data

Originally, our data set had 14 attributes. We chose to only focus on the following five attributes: *age*, *education-number*, *hours-per-week*, *work class* and *occupation*. Additionally we narrowed down our data set to only include the data of the countries with the 10 highest GPAs. Out of the five attributes we have three continuous attributes; *age*, *education-number* and *hours-per-week*, and two categorical attributes; *work class* and *occupation*. Attributes age and *hours-per-week* are both ratio attributes whereas *education-number* is an ordinal attribute. Attributes *work class* and *occupation* are both nominal attributes.

None of the selected continuous attributes have missing values. Both *work class* and *occupation* have missing values. The missing values will make it harder to predict how *work class* influences the probability of whether an income exceeds $50K/yr.

To get a better overview of our selected attributes we calculate the mean, standard deviation, median and the range of our continuous attributes, see Table 1.

	mean	***standard deviation***	***median***	***range***
age	37.70	12.94	36.50	73.00
edu-num	10.37	3.043	10.00	15.00
hours-per-week	39.57	11.23	40.00	98.00

Table 1: Basic summary statistic of the continuous attributes; *age*, *education-number* and *hours-per-week*.

From Table 1 we see that both *age* and *hours-per-week* has a very large range. We investigate maximum and minimum value of *hours-per-week*. The large range is a result of the minimum hours-per-week being 1 hour and the maximum being 99 hours. From the basic summary statistics we don't have sufficient information to determine whether 99 hours is an outlier, but it certainly is an extreme observation considering there are a total of 168 hours in a week. The large range in the *age* attribute is a result of the maximum age being 90 years and the minimum age being 17 years. This is not an unrealistic age range in a data set that wishes to have enough data to predict if an income exceeds $50K/yr.

In the next section we visualise our data with the help of histograms and box plots.

4 Data visualisation

We will know investigate whether there are outliers in the selected attributes. To assess if the attributes are normally distributed, we turned our attention to the histograms. Insights from the histograms reveal:

Continuous attributes:

- *age*: Demonstrates a right-skewed distribution as evident from its tail direction (see Figure 1).
- *edu-num*: Appears to represent a bimodal distribution (see Figure 2).
- *hours-per-week*: Indicates an outlier with a highly frequent value in the range of 35-40 (see Figure 3).

Categorical attributes:

- *work class*: Predominantly, "Private" emerges as the highest frequency category, with a frequency near 250. All other work classes have frequencies below 12 individuals. (see Figure 4).

- *occupation*: The distribution appears to be relatively uniform with certain roles emerging as outliers (see Figure 5). Transport-moving, Priv-house-serv, and Protective-serv are the three occupations with much lower frequencies compared to other roles. It is notable that the occupation titled "?" (unspecified) includes about 15 individuals. This count surpasses the occupations with the three lowest frequencies, placing the "?" category on part with Tech-support and other mid-frequency occupations.
- *income*: The histogram for the income attribute uncovers a pronounced disparity between two income brackets. Over 250 individuals report incomes $\leq 50K$, in contrast to slightly fewer than 100 individuals who claim incomes exceeding $50K$. This skewness leans heavily towards the lower income group (see Figure 6).

To discern if the attributes adhere to a formal normal distribution, we utilize Q-Q plots. The subsequent observations were made:

- *age*: Exemplifies a U-shape pattern (see Figure 7). The points' distribution below and above the $y = x$ line at varied ends suggests fewer extreme values relative to an ideal normal distribution. This could be attributed to the data set's focus on working-age individuals, thereby including fewer individuals who are very young or very old.
- *edu-num*: Demonstrates a W-shaped structure indicative of a bimodal distribution (see Figure 8). This implies the presence of two major distributions within education levels.
- *hours-per-week*: The Q-Q plot showcases a sharp incline, an elongated flat segment, and another sharp rise, indicating a non-linear plot, hence not a normal distribution (see Figure 9).

The correlation heat-map, as displayed in Figure 10, provides insights into the relationships between variables. We can see that there are no significantly strong correlations between any of the continuous variables.

4.1 PCA and MCA

The PCA is executed post the standardization of continuous data based on their standard deviation. Key aspects of the PCA analysis include:

1. Variation explained concerning the number of PCA components.
2. Principal directions of the included PCA components.
3. Data projection onto the selected principal components.

In initiating the PCA analysis, we did not predetermine the number of components. This approach allowed for a thorough examination of the explained variance inherent in each components. The emergence of 21 PCs - a notably high count- can be traced back to the application of one-hot encoding on our categorical data. While it is vital to encapsulate as much data detail as possible, it is equally important to manage its complexity. Hence, for the analysis and visualization, we used three principle components instead (see Figure

11). The 3D scatter plot for PCA did not show any discernible clustering, suggesting a lack of correlation between values. This is in-line in what we discovered through the correlation heat-map.

The MCA on the categorical data showcases two PCs explaining the entirety of the variation. Noting the result from this MCA analysis on the categorical data, where 2 PCs explain all the variation. The MCA scatter plot interestingly shows three clear outliers, a cluster at values where D1 is between -0.5 and 1.0 with D2 around 0, as well as a denser cluster where D1 is between -0.5 and 0.25 and D2 with D2 around 0. This suggests there is some clustering in our categorical data: work-class, occupations, and distribution of income. It might be work further investigating.

4.2 Figures

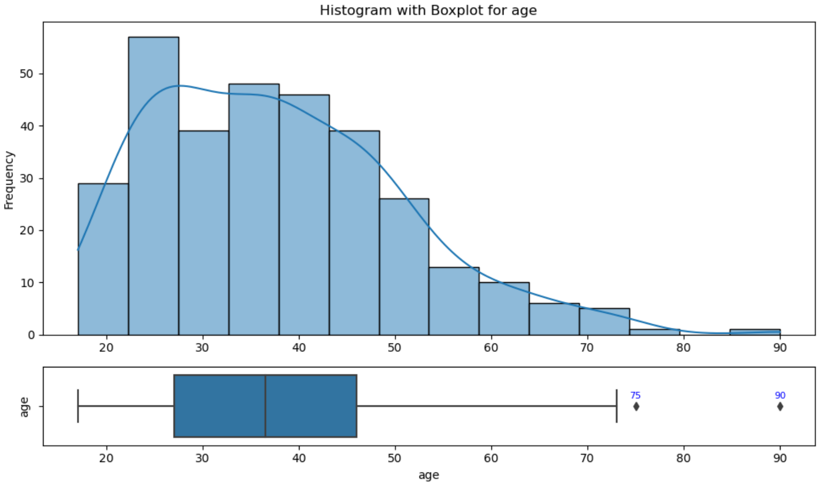

Figure 1: Distribution of *age*.

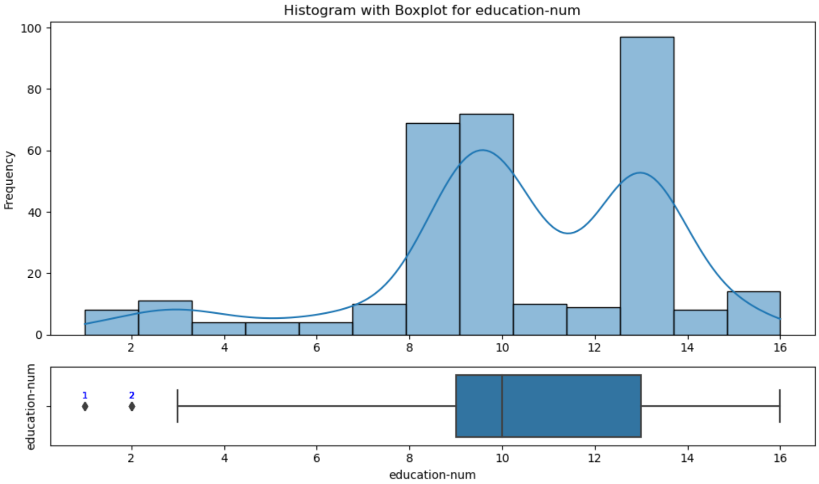

Figure 2: Distribution of *edu-num*.

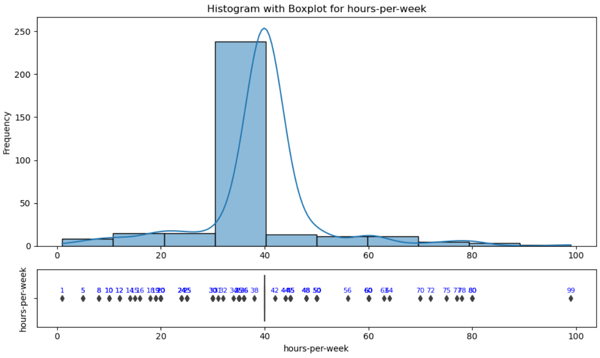

Figure 3: Distribution of *hours-per-week*.

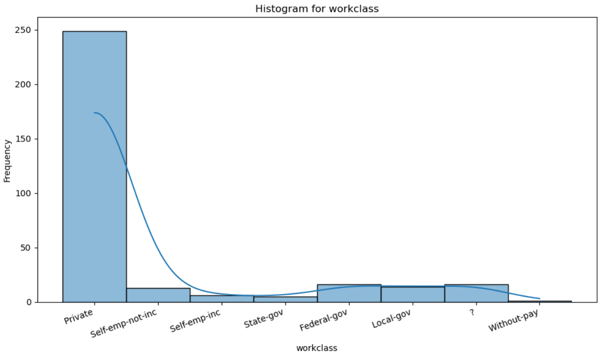

Figure 4: Distribution of *work-class*.

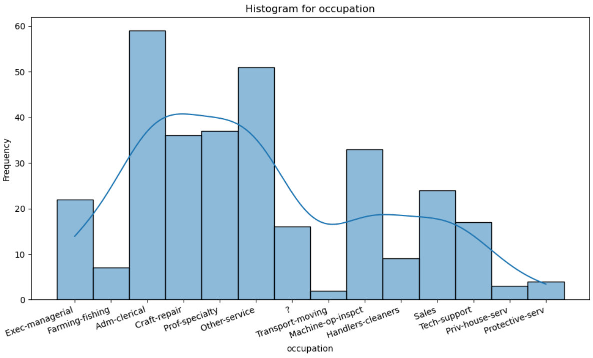

Figure 5: Distribution of *occupation*.

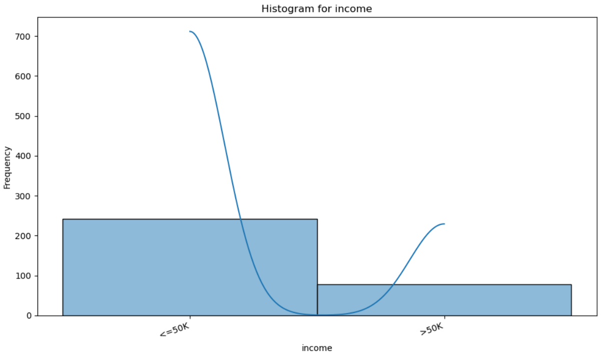

Figure 6: Distribution of *income*.

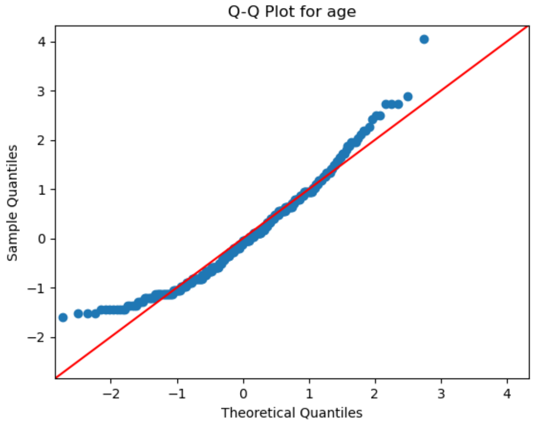

Figure 7: *age.*

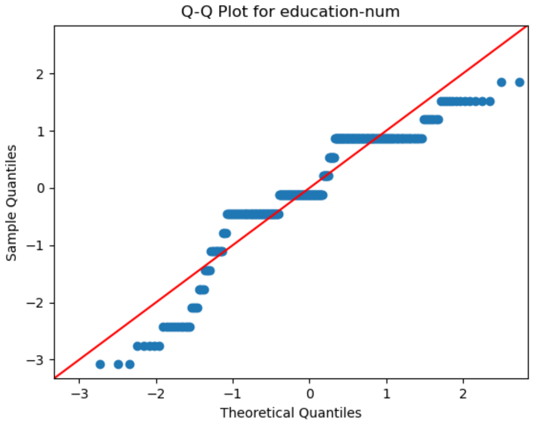

Figure 8: *edu-num.*

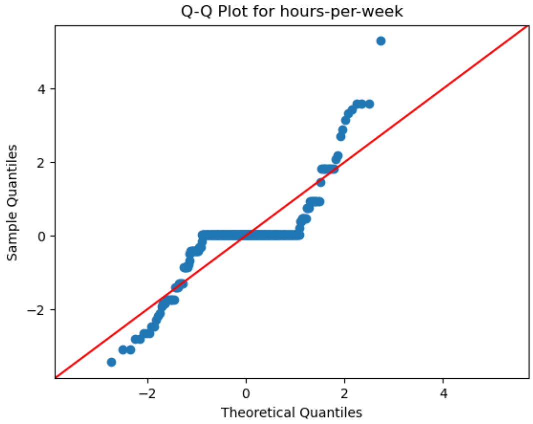

Figure 9: *hours-per-week.*

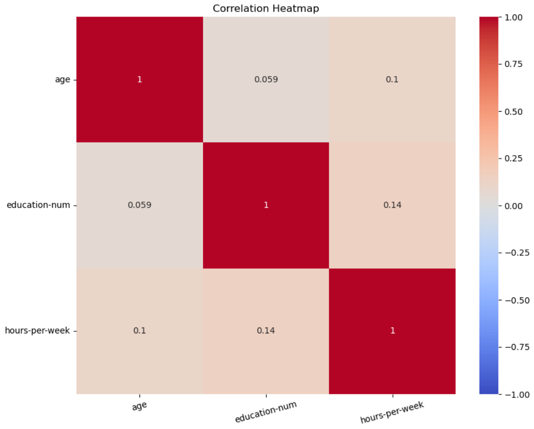

Figure 10: Correlation Heat-map.

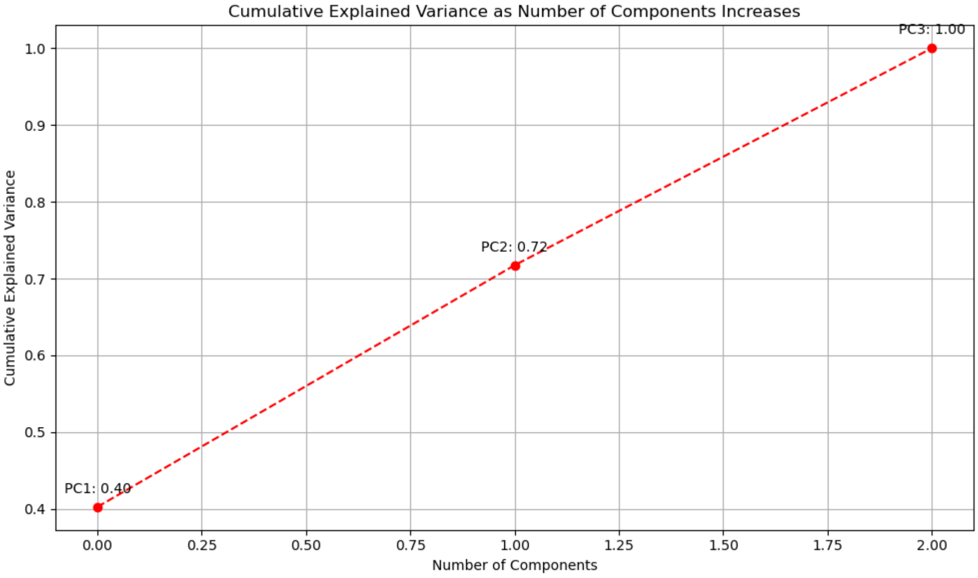

Figure 11: PCA cumulative of 3 principal components (PC)s.

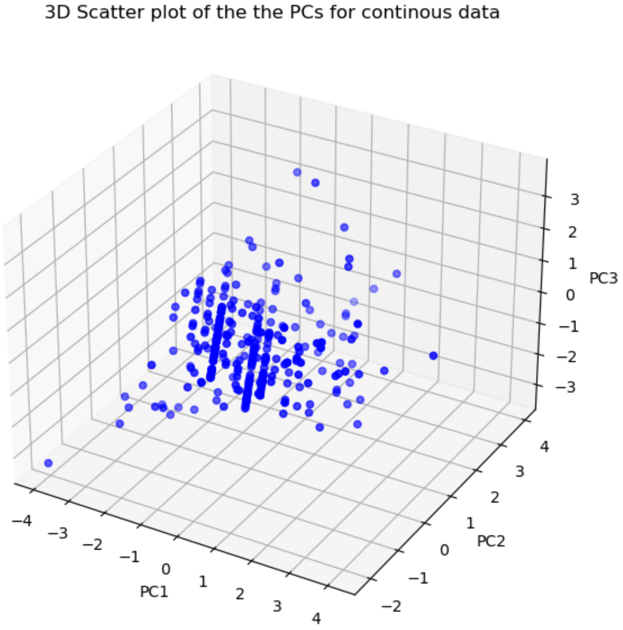

Figure 12: PCA 3D scatter plot of 3 PCs.

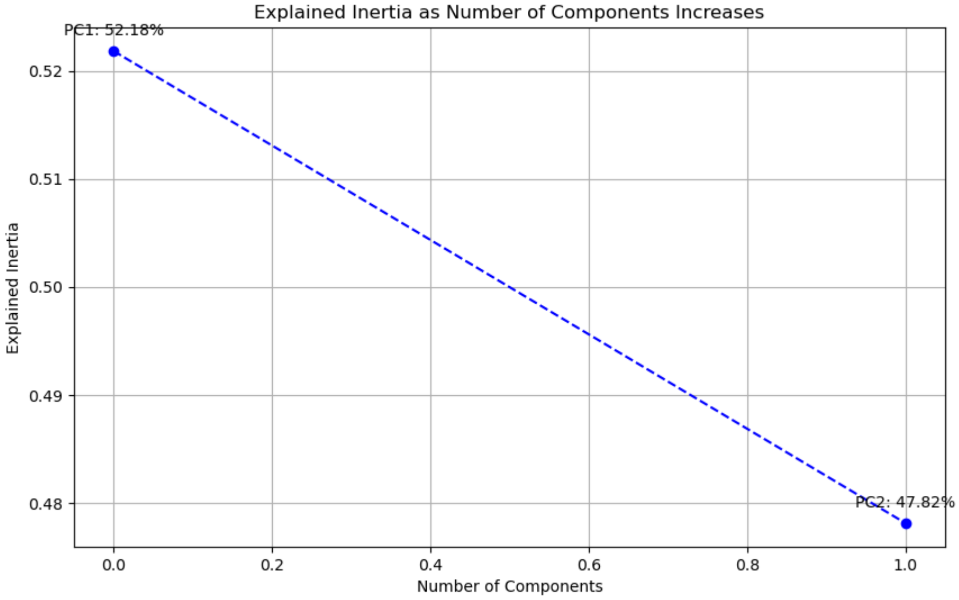

Figure 13: MCA explained interia.

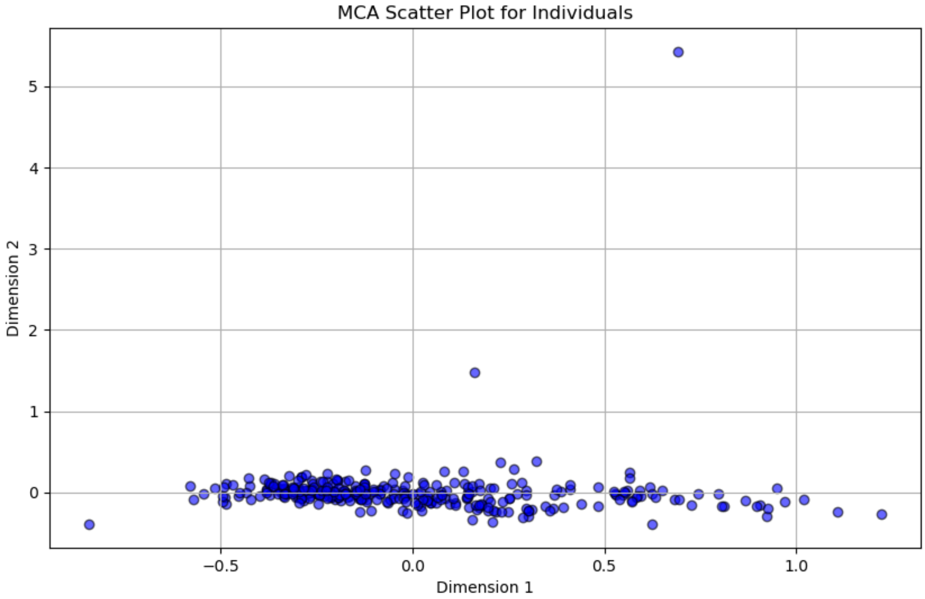

Figure 14: MCA scatter plot.

5 Discussion

From our correlation analysis of the *age*, *edu-num* and *hours-per-week* we see that there is a weak positive correlation between the continuous attributes. In the Heat-map, the highest correlation between *edu-num* and *hours-per-week* is only 0.14, which means it is a very low relation between them. Hence, we could easily assert that there is a weak correlation among the continuous attributes.

About PCA analysis. It's high component count (21 PCs) results from the one-hot encoding of our categorical data. While our aim is to capture as much variance as possible, managing complexity is crucial. Thus, we've resorted to using the first three principal components for further analysis and visualization. Notably, the PCA scatter plot did not show any distinct clustering, aligning with our correlation findings.

The detailed visualization offers valuable insights into the nature of our data. While some attributes, like *age*, exhibit predictable patterns that align with real-world knowledge (e.g., dominant working age), others, like *edu-num*, present intriguing bimodal patterns that warrant further investigation. The lack of pronounced correlation between the continuous variables is both a challenge and an opportunity: while it might complicate predictive accuracy, it ensures that our model won't suffer from multicollinearity.

The PCA results underscore the significance of cautious feature selection, especially post one-hot encoding. The absence of distinct clustering in the PCA scatter plot reaffirms our earlier correlation findings, suggesting that simple linear models might not be the best fit. More sophisticated, non-linear models could be more suitable.

6 Selected exam questions

1. **Option A**: To see this we look at the given attributes. We see that attribute x_1 consists of 30-minute interval (coded) meaning the values in the x_1 attribute is $x_1 \in [1..27]$. This means x_1 can not be an interval attribute instead we have x_1 is categorised. This means x_1 is a nominal attribute which only leaves us with option A. The three other postulates in option A are also true.

2. **Option A**: The p-norm distance between vectors $\mathbf{a}$ and $\mathbf{b}$ is given by:

$$d_p(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}}$$

To address that question - the distance between x_{14} and x_{18} is:

$$d_p(x_{14}, x_{18}) = (7^p + 2^p)^{\frac{1}{p}}$$

Since no p-value was given up-front, I decided to calculate out each of the options.

$$\begin{aligned}
\text{A.}\quad d_\infty(x_{14}, x_{18}) &= \max(|26-19|, |2-0|) \\
&= \max(7, 2) \\
&= 7 \\
\text{B.}\quad d_3(x_{14}, x_{18}) &= \left(7^3 + 2^3\right)^{\frac{1}{3}} \\
&= (343 + 8)^{\frac{1}{3}} \\
&\approx 7.039 \\
\text{C.}\quad d_1(x_{14}, x_{18}) &= |26-19| + |2-0| \\
&= 7 + 2 \\
&= 9 \\
\text{D.}\quad d_4(x_{14}, x_{18}) &= \left(7^4 + 2^4\right)^{\frac{1}{4}} \\
&= (2401 + 16)^{\frac{1}{4}} \\
&\approx 6.346
\end{aligned}$$

The only result which matched with the options provided in the question was A.

3. **Option A**: From the singular values matrix S, I can see that the squared singular values are:

$$\begin{aligned}
\lambda_1 &= 13.9^2 \\
\lambda_2 &= 12.47^2 \\
\lambda_3 &= 11.48^2 \\
\lambda_4 &= 10.03^2 \\
\lambda_5 &= 9.45^2
\end{aligned}$$

Using this we can calculate for the total variance:

$$\lambda_{\text{total}} = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5$$

To decide on the correct answer, I will write out the formula that would lead to it, then pug in the values.

Option A:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{\lambda_{\text{total}}} = \frac{193.21 + 155.5809 + 131.9504 + 100.6009}{670.6447} = 0.8735$$

Option B:

$$\frac{\lambda_3 + \lambda_4 + \lambda_5}{\lambda_{\text{total}}} = \frac{131.9504 + 100.6009 + 89.3025}{670.6447} = 0.4875$$

Option C:

$$\frac{\lambda_1 + \lambda_2}{\lambda_{\text{total}}} = \frac{193.21 + 155.5809}{670.6447} = 0.5200$$

Option D:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_{\text{total}}} = \frac{193.21 + 155.5809 + 131.9504}{670.6447} = 0.7179$$

I compare these answers to the options provided. The only correct option would be Option A ($0.8735 > 0.8$).

4. **Option D**: To see this we start by looking at the $\mathbf{V}$ matrix. Here we know the columns represent the principal components and the rows represent the coefficients assigned to each of the five attributes when creating the second PC.

   We can also see that for option A the given observation values ($\tilde{\mathbf{X}}$) do not correspond to the values in the PC column 5 of $\mathbf{V}$. Same goes for option B. For option D we see for PC 2 (column 2 of $\mathbf{V}$) we have $\mathbf{V}_2^{\mathrm{T}} = [-0.5\ 0.23\ 0.23\ 0.09\ 0.8]$, which corresponds to the given attribute information: x_1 has a low value and attributes x_2, x_3 and x_5 have high values. The same goes for option C.

   We can find the projection onto a PC by calculating the dot product between the $\tilde{\mathbf{X}}$ and the column of $\mathbf{V}$ corresponding to the PC in question. Both the dot product in option C and D give us a positive value. Hence, option D is the only true statement.

5. **Option A**: As we know the **Jaccard similarity** is:

$$J(x,y) = \frac{f_{11}}{K - f_{00}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

   - $f_{11}(s1, s2) = \{the, words\} = 2$
   - $f_{10} + f_{01} + f11 = 6 + 5 + 2 = 13$

   Hence, we could get the Jaccard similarity's value is:

$$J(s1, s2) = \frac{2}{13} = 0.153846$$

6. **Option B**: As we have already know $y = 2$, so

$$\begin{aligned} p(\hat{x_2} = 0 | y = 2) &= \sum_{\hat{x_7}} p(\hat{x_2} = 0, \hat{x_7} = i | y = 2) \\ &= p(\hat{x_2} = 0, \hat{x_7} = 0 | y = 2) + p(\hat{x_2} = 0, \hat{x_7} = 1 | y = 2) \\ &= 0.81 + 0.03 = 0.84 \end{aligned}$$

List of Figures

List of Tables

References

[1] B. Becker and R. Kohavi, “Adult.” UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.