**Data Science and Machine Learning (MSc)**

DAMA51: Foundations in Computer Science

Academic Year: 2022–2023

|  |  |
| --- | --- |
| #2 Written Assignment | |
| Submission Deadline | Wednesday, 1 February 2023, 23:59:59 EET |
| Student Name |  |

# Guidelines

The deadline is definitive.

An indicative solution will be posted online along with the returning of the graded assignments.

The assignment is due via the STUDY submission system. **You are expected to deliver a document (.DOC, .ODT, .PDF) and a compressed (.ZIP, .RAR) file containing all your work:**

* 1 document file (this document) with the answers to all the topics, along with the R code and the results of the execution of the code
* 1 compressed file with 3 R scripts that correspond to topics 3, 4 and 5.

**You should not make any changes in the written assignment file other than providing your own answers.** You should also type all of your answers into Word and not attach any handwritten notes as pictures into your work otherwise a 5% reduction of your final grade will be applied. Make sure to name all the files (ZIP file, DOC file and R script files) with **your last name first followed by a dash and the names of each component at the end**. For example for the student with last name Aggelou the files should be named as follows: Aggelou-HW1.zip, Aggelou-HW1.doc, Aggelou-Topic3.R, Aggelou-Topic4.R and Aggelou-Topic5.R. The R script files should automatically run with the **source** command and generate the correct results. Also, please include comments before each command to explain the functionality of the command that follows. Unless otherwise stated in the question, all numerical answers should be given to **three decimal places**.

|  |  |  |
| --- | --- | --- |
| Topic | Points | Grades |
| 1. Online QUIZ | 40 |  |
| 1. **Article Review** | 10 |  |
| 1. **Principal Component Analysis** | 20 |  |
| 1. **Confusion Matrix** | 10 |  |
| 1. **Hypothesis Testing –**   **χ2 Test** | 20 |  |
| **TOTAL** | **100** | **/100** |

# Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

[https://study.eap.gr/mod/quiz/view.php?id=24566](about:blank)

You have one effort and unlimited time to complete the quiz, up to the submission deadline. **(40 points)**

# Topic 2: Article Review

The review article by Shamina Ahmed et al. entitled “Artificial intelligence and machine learning in finance: A bibliometric review” (available at <https://doi.org/10.1016/j.ribaf.2022.101646>) uses a bibliometric approach to review the artificial intelligence and machine learning literature in the finance field highlighting the main application areas. Summarize the main research streams reviewed as well as the methodologies used in each of these areas. Finally give your personal view on the main research challenges and how to tackle these based on AI and ML methodologies (you may use a personal business perspective, if available, beyond the finance field).

Note: You should write up your answer to a maximum of 300 words. Any text in excess of 300 words will not be taken into consideration.

**(10 points)**

|  |
| --- |
|  |

# Topic 3: Principal Components Analysis

For this topic, you’ll use the built-in data set **USArrests** of R to answer the following points. Within each answer frame below you should include the R code as well as the results **(20 points)**

1. Using R, first review each attribute included in the **USArrests** dataset (write the name and type) and calculate the mean and standard deviation. Include your answers in the Tables provided **(4 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(2 points) Name and type of variables***   |  |  | | --- | --- | | Variable Name | Variable Type | | Murder | number | | Assault | integer | | UrbanPop | integer | | Rape | number |   ***R-code and Results: data(USArrests),***  ***Names(USArrests)***  ***"Murder" "Assault" "UrbanPop" "Rape"***  ***str(USArrests)***  ***'data.frame': 50 obs. of 4 variables:***  ***$ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...***  ***$ Assault : int 236 263 294 190 276 204 110 238 335 211 ...***  ***$ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...***  ***$ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...***  ***(2 points) Mean and standard deviation for each variable***   |  |  |  | | --- | --- | --- | | Variable Name | Mean | Standard Deviation | | Murder | 7.788 | 4.35551 | | Assault | 170.8 | 83.33766 | | UrbanPop | 65.54 | 14.47476 | | Rape | 21.23 | 9.366385 |   ***R-code and Results: summary(USArrests)***  ***Murder Assault UrbanPop Rape***  ***Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30***  ***1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07***  ***Median : 7.250 Median :159.0 Median :66.00 Median :20.10***  ***Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23***  ***3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75 3rd Qu.:26.18***  ***Max. :17.400 Max. :337.0 Max. :91.00 Max. :46.00***  ***sd(USArrests$Murder)***  ***4.35551***  **sd(USArrests$Assault)**  **83.33766**  **sd(USArrests$UrbanPop)**  **14.47476**  **sd(USArrests$Rape)**  **9.366385** |

1. Using R, based on question (a), compute the principal components, ensuring that PCA is applied on scaled variables. **(4 points)**

|  |
| --- |
| result<-prcomp(USArrests, scale=TRUE)  result$rotation  result$rotation<--1\*result$rotation  result$rotation |

1. Using R and your results from question (b), write in the Table below, for each principal component, what is the percentage of the total variance that is explained. **(4 points)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Percentage of the total variance that is explained (1 point/PC)**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | PC1 | PC2 | PC3 | PC4 | | Percentage of the total variance explained | 62% | 24.7% | 8.9% | 4.3% |   ***R-code and Results:***  **summ<-summary(result)**  **> summ**  **Importance of components:**  **PC1 PC2 PC3 PC4**  **Standard deviation 1.5749 0.9949 0.59713 0.41645**  **Proportion of Variance 0.6201 0.2474 0.08914 0.04336**  **Cumulative Proportion 0.6201 0.8675 0.95664 1.00000**  **> summ$importance[2,]**  **PC1 PC2 PC3 PC4**  **0.62006 0.24744 0.08914 0.04336** |

1. Using R and based on the previous results, write the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed (PCA loadings) in the Table below. **(4 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***PCA loadings***   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | PC1 | PC2 | PC3 | PC4 | | Murder | -0.5358995 | 0.4181809 | -0.3412327 | 0.64922780 | | Assault | -0.5831836 | 0.1879856 | -0.2681484 | -0.74340748 | | UrbanPop | -0.2781909 | -0.8728062 | -0.3780158 | 0.13387773 | | Rape | -0.5434321 | -0.1673186 | 0.8177779 | 0.08902432 |   ***R-code:***  ***A <- as.matrix(data.frame(USArrests), byrow = TRUE)***  ***S<-scale(A)***  ***pca\_svd<-svd(S)***  ***pca\_svd$v*** |

1. Using R, and the previous results, create a scree plot. Based on the scree plot, how many principal components would you retain? **(4 points)**

|  |
| --- |
| ***(3 point) Create a scree plot***  ***R-code and Plot:***  ***results <- prcomp(USArrests, scale = TRUE)***  ***var\_explained = results$sdev^2 / sum(results$sdev^2)***  ***library(ggplot2)***  ***qplot(c(1:4), var\_explained) +***  ***geom\_line() +***  ***xlab("Principal Component") +***  ***ylab("Variance Explained") +***  ***ggtitle("Scree Plot") +***  ***ylim(0, 1)***  ***(1 point) How many principal components would you retain?***  ***I would retain the first 3.*** |

# Topic 4: Confusion Matrix

You are given the following vectors a = [1,0,1,0,1,1,1,0,0,1] and b = [1,0,0,0,1,1,0,0,0,1], which represent the True and predicted values, respectively, of a diagnostic test being positive (0) or negative (1) for a specific disease. **(10 points)**

1. Using “pen and paper”, fill in the confusion matrix below and then calculate the following statistics: sensitivity, specificity and accuracy (type the formulas, calculations and results in the answer frame below). **(5 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(2 points) Confusion Matrix***   |  |  |  | | --- | --- | --- | | **True Class** | **Predicted Class** | | | **Positive (0)** | **Negative (1)** | | **Positive (0)** | 4 | 0 | | **Negative (1)** | 2 | 4 |   ***(1 point) Sensitivity = TP/(TP+FN)= 1***  ***(1 point) Specificity = TN/(TN+FP)=0.6***  ***(1 point) Accuracy = (TP+TN)/n=0.8*** |

1. Using R, create the confusion matrix (fill in the one provided below) and calculate the following statistics: sensitivity, specificity and accuracy to verify your results from question (a). **(5 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(2 point) Confusion Matrix***   |  |  |  | | --- | --- | --- | | **True Class** | **Predicted Class** | | | **Positive (0)** | **Negative (1)** | | **Positive (0)** | 4 | 0 | | **Negative (1)** | 2 | 4 |   ***(1 point) Sensitivity = 1.0000***  ***(1 point) Specificity =0.6667***  ***(1 point) Accuracy = 0.8***  ***R-code and Results:***  ***ex\_values<-factor(c(1,0,1,0,1,1,1,0,0,1))***  ***> pr\_values<-factor(c(1,0,0,0,1,1,0,0,0,1))***  ***> matrix<-confusionMatrix(data=pr\_values, reference=ex\_values)***  ***> matrix***  ***Confusion Matrix and Statistics***  ***Reference***  ***Prediction 0 1***  ***0 4 2***  ***1 0 4***    ***Accuracy : 0.8***  ***95% CI : (0.4439, 0.9748)***  ***No Information Rate : 0.6***  ***P-Value [Acc > NIR] : 0.1673***    ***Kappa : 0.6154***    ***Mcnemar's Test P-Value : 0.4795***    ***Sensitivity : 1.0000***  ***Specificity : 0.6667***  ***Pos Pred Value : 0.6667***  ***Neg Pred Value : 1.0000***  ***Prevalence : 0.4000***  ***Detection Rate : 0.4000***  ***Detection Prevalence : 0.6000***  ***Balanced Accuracy : 0.8333***    ***'Positive' Class : 0*** |

# Topic 5: χ2 Test

For this topic, you will use the built-in data set **iris** of R in order to answer the following points. **(20 points)**

1. Using R, first find the variables included in the dataset **iris** and write their names and types in the Table below. Then add a new variable, Size\_sepal, which is “small”, if the length of the sepal is smaller than the median sepal length of all flowers, or “big” otherwise. **(5 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(2 points) Variable names and type***   |  |  | | --- | --- | | Variable Name | Variable Type | | Sepal.Length | number | | Sepal.Width | number | | Petal.Length | number | | Petal.Width | number | | Species | factor |   ***R-code and Results:***  ***(3 points) add variable “Size\_sepal”***  ***R-code:***  ***summary(iris$Sepal.Length)***  ***iris$Size\_sepal<-with(iris, ifelse(Sepal.Length <= 5.800,'Small','Big'))***  ***head(iris)*** |

1. Using R, create a contingency table, including sums, for the variables Species and Size\_sepal. Create a stacked barplot in R to visualize Size\_sepal (y-axis) for each of the Species (x-axis). **(5 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(3 points) contingency table for the variables Species and Size\_sepal***   |  |  |  |  | | --- | --- | --- | --- | |  | *Size\_sepal* | |  | | *Species* | ***Big*** | ***Small*** | ***Sum*** | | *Setosa* | 0 | 50 | 50 | | *versicolor* | 26 | 24 | 50 | | *virginica* | 44 | 6 | 50 | | *Sum* | 70 | 80 | 150 |   ***R-code and Results:***  ***con1<-table(iris$Species,iris$Size\_sepal)***  ***> con1***    ***Big Small***  ***setosa 0 50***  ***versicolor 26 24***  ***virginica 44 6***  ***> addmargins(con1)***    ***Big Small Sum***  ***setosa 0 50 50***  ***versicolor 26 24 50***  ***virginica 44 6 50***  ***Sum 70 80 150***  ***(2 points) Create a stacked barplot***  ***R-code and Plot:***  ***library(ggplot2)***  ***>ggplot(irisplus) +***  ***aes(x = Species, fill = Size\_sepal) +***  ***geom\_bar()*** |

1. Define two hypotheses (null and alternative) in the Table below, to test whether the Size\_sepal variable is independent of the Species variable. Using R, perform the Chi-Squared Test of Independence (using default values) and write if you can reject or not the null hypothesis at a significance level of a=0.05. **(5 points)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***(2 points) Define the two hypotheses below:***   |  |  | | --- | --- | | ***For Size\_sepal and Species*** | | | ***Ho (null Hypothesis):*** | ***Size\_sepal is independent of the Species*** | | ***H1 (alternative Hypothesis):*** | ***Size\_sepal is dependent of the Species*** |   ***(3 points) Perform the Chi-Squared Test of Independence and write if you can reject or not the null hypothesis for the variables Size\_sepal and Species***  ***R-code: chisq.test(con1)***  ***p-value: 8.373761e-18***  ***Decision: Reject the null Hypothesis. Size\_sepal is dependent of the Species*** |

1. Using R, based on the results from question (c), calculate the degrees of freedom (df) and fill in the Tables below with the observed and expected values for variables Species and Size\_sepal. **(5 points)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***(3 points) Observed and expected values for the variables Species and Size\_sepal***  ***Observed values for the variables Species and Size\_sepal***   |  |  |  | | --- | --- | --- | | ***Observed Values*** | ***Size\_sepal*** | | | ***Species*** | ***big*** | ***Small*** | | ***setosa*** | 1 | 49 | | ***versicolor*** | 29 | 21 | | ***virginica*** | 47 | 3 |   ***Expected values for the variables Species and Size\_sepal***   |  |  |  | | --- | --- | --- | | ***Expected Values*** | ***Size\_sepal*** | | | ***Species*** | ***big*** | ***small*** | | ***Setosa*** | 25.667 | 24.333 | | ***Versicolor*** | 25.667 | 24.333 | | ***Virginica*** | 25.667 | 24.333 |   ***R-code and Results:***  ***test1$observed***  ***big small***  ***setosa 1 49***  ***versicolor 29 21***  ***virginica 47 3***  ***> round(test1$expected,3)***  ***big small***  ***setosa 25.667 24.333***  ***versicolor 25.667 24.333***  ***virginica 25.667 24.333***  ***(2 points) Calculate the degrees of freedom (df) and write the result below:***  ***df(Species - Size\_sepal) = (3-1)x(2-1)=2*** |