Homework 3

Due by: April 16, 2023, 11:59 PM

---

*All assignments must be submitted through **eLearning**. Alternative submission methods are not acceptable. **Submissions after the deadline will not be accepted**, and accordingly, a grade of zero will be automatically applied for a missing submission after the deadline.*

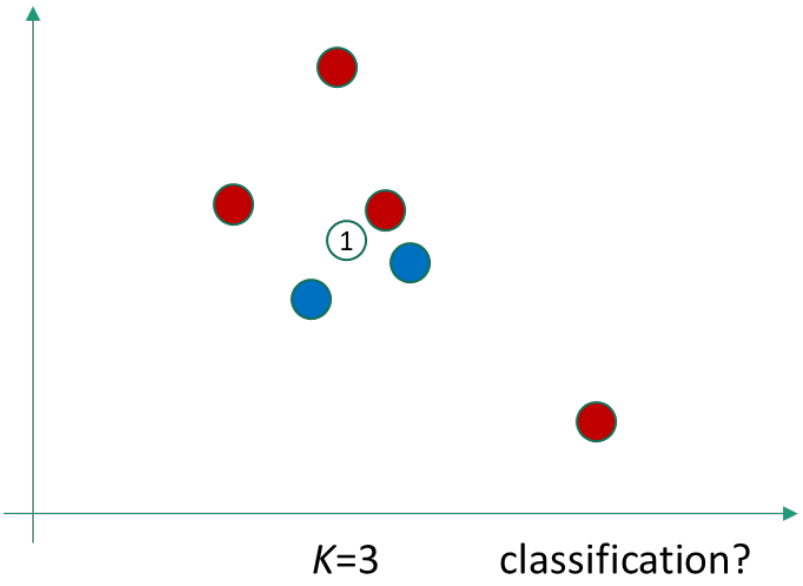
*When you submit, please make your file a **docx version** and specify the **name of the file** as follows. "**[Homework 3] Your first name**" (e.g., [Homework 3] Andrew)*

Q1. (1 point) [multiple choice] Which analytical methods belong to the 'supervised algorithm', not the 'unsupervised algorithm? (***Multiple choices possible***)

(1) K-means clustering
(2) K-nearest neighbor
(3) Hierarchical clustering
(4) Decision tree
(5) Logistic regression

Answer)

Q2. (1 point) [short answer] Assume that you are following the **K-nearest neighbor algorithm**. As you can see from the graph below, there are two classes, 'blue' and 'red'. If we set the number of nearby neighbors (i.e., K) as 3, which class the node number 1 will be classified in? 'blue' or 'red'?



Answer)

Q3. (1 point) [short answer] Fill in the blanks. The following sentence is directly related to the reason why we should learn about dimension reduction.

"( ) describes the explosive nature of **increasing data dimensions** and its resulting **exponential increase in computational efforts** required for its processing and/or analysis."

Answer)

Q4. (1 point) [multiple choice] Which analytical method is most required for the following task?

"To do a dimension reduction, we want a construct a new set of derivative features that are fewer in number but capture most of the variation of existing variables' information."

(1) Domain knowledge-based dimension reduction
(2) Decision tree
(3) Principal Component Analysis (PCA)
(4) Contingency table
(5) Correlation matrix

Answer)

You can download 'USArrests.csv' file from eLearning. Please download the file and locate it in your R working directory folder. After that, type '`df <- read.csv("USArrests.csv", header=TRUE, stringsAsFactors=TRUE)`' at the console display to load the data at your R Studio. Install and load the '**factoextra**' package. Load the '**cluster**' package.

Q5. (1 point) [R practice] Type the following code first: `df_scaled <- df`. Make a standardization for each independent variables (i.e., Murder, Assault, UrbanPop, and Rape) and assign those variables to '**df_scaled**'. You should specify the **R code** here. (Tip: refer to the Week 11 ppt)

Answer)

Q6. (1 point) [R practice] Make an elbow plot by using '**fviz_nbclust**' function to choose the optimized number of K in the K-means clustering. You should specify (1) **R code**, (2) **Graph**, and (3) **the optimized value of K** by interpreting the graph. (Tip: refer to the Week 11 ppt)

Answer)

Q7. (1 point) [R practice] Make a silhouette plot by using '**fviz_nbclust**' function to choose the optimized number of K in the K-means clustering. You should specify (1) **R code**, (2) **Graph**, and (3) **the optimized value of K** by interpreting the graph. (Tip: refer to the Week 11 ppt)

Answer)

Q8. (1 point) [R practice] Type the following code first: `set.seed(1004)`. Conduct a K-means clustering by using '**kmeans**' function. Assume that the optimized number of K is 4. You should specify the **R code** here. (Tip: refer to the Week 11 ppt)

Answer)