Domande Machine Learning

1 Data

- Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234} the absolute average deviation is ...?

$$AAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_m| =$$

- Assume you are given a data base containing an attribute Food Quality which can take the following values excellent, very good, good, average, bad, terrible. Assume you want to apply binarization to the Food Quality attribute, How many new attributes do you need to create to achieve this goal, in case you can accept to induce correlation among them?

- Assume you are given a data base containing an attribute Payment Category which can take the following values excellent, good, average, bad, terrible. Assume that the first 10 records of the database are associated wirh the following values of the Payment Category attribute; excellent, good, average, average, average, bad, terrible, excellent, good, average. Which are the mode of the Payment Category attribute and the associated relative frequency value? **(average; 0.4) Bisogna prendere la moda e valutarne la frequenza relativa:**

$$fr = \frac{count}{N}$$

- Given the following set of values for the attribute Yearly Income (Euro) 2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234 the standard deviation is ...? **500.4 (prima cifra decimale arrotondata verso il basso)**

- Assume you are given a data base containing the following attribute, Credit Card owner. Assume that Credit Card owner can take values yes, no. Assume you have the following 10 records for the Credit Card owner attribute {yes, yes, no, yes, no, no, no, no, no, no}. Assume you want to form a sample contisting of 4 records. Then, you apply random sampling with stratification (stratified sampling) using the equal number option. Which of the following samples is a valid one?

- Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234} which are the values of the mean and trimmed mean? **(2,133; 2,090) La trimmed mean è la media calcolata togliendo il valore più alto e quello più basso.**

- Given the attribute named "car color", which of the following operations and/or quantities is/are meaningful? **(mode, entropy) poichè è un attributo categorico esse sono le due quantità interessanti.**

- Assume you are given a data base containing an attribute Payment Category which can take the following values excellent, good, average, bad, terrible. Assume that the first 10 records of the database are associated with the following values of the Payment Category attribute; excellent, good, average, average, average, bad, ?, excellent, good, average, where "?" stands for missing value. In case

you apply mode replacement which is the value used to replace "?"? **(average) poiché la moda è l'elemento con la frequenza più alta.**

- You have applied the procedure for supervised discretization of the Yearly Income attribute. In particular, you made the decision to use Entropy to find the optimal discretization for the Yearly Income attribute. Which of the following statements is/are true? **The optimal discretization is the one which achieves the minimum value of Entropy.**
  Achieving the optimal value of Entropy means you are implementing the purest possible discretization solution.
  Any valid discretization is associated with a non negative value of Entropy.

- Assume you are given a data base containing the customer data, in particular the following attributes; Name, Age, Purchase date, Yearly Income, Credit Card owner, and Payment Category. Assume that Credit Card owner can take values yes, no, Payment Category can take values excellent, good, average, bad, terrible. Which of the following statements is/are true? **"Payment Category" is a categorical attribute**
  The only binary attribute is "Credit Card owner"
  "Payment Category" is a nominal attribute

- Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234} the range is ...? **1424**

- Given the attribute named "patient weight", which of the following operations and/or quantities is/are meaningful? **percentiles**

- Assume you are given a data base containing an attribute Payment Category which can take the following values excellent, good, average, bad, terrible. Assume that the first 10 records of the database are associated wirh the following values of the Payment Category attribute; excellent, good, average, average, average, bad, terrible, excellent, good, average. In case you apply random sampling with replacement to form a sample consisting of 5 records, which of the following is/are a valid sample/s? **(good, average, excellent, average, terrible)**
  (good, good, bad, average, terrible)
  (good, good, bad, average, good)
  (average, average, excellent, average, average)
  (average, average, average, average, average)
  Vanno tutte bene percé c'è reinserimento.

- Given the attribute named "gender", which of the following operations and/or quantities is/are meaningful? **mode,contingency, entropy sono le misure utili per descrivere gli attributi binari.**

- The main advantages of aggregation are ... Choose one or more of the following statements: **Smaller data sets; we need less memory and processing time, thus more time consuming and possibly more effective algorithms can be used.**
  Reduced variance; attributes computed on aggregated records are more stable than those associated with the native records.

- Assume you are given a data base containing an attribute Food Quality which can take the following values excellent, very good, good, average, bad, terrible. Assume you want to apply binarization to the Food Quality attribute, How many new attributes do you need to create to achieve this goal, in case you can NOT accept the induced correlation among them? **6**

- Which properties of numbers are used to describe attributes? **type**

- Assume you have a dataset consisting of 100,000 records and 1,000 continuous attributes. You know that the probability of missing value for each attribute is equal to 0.005. Furthermore, assume that the missing mechanism is at random, i.e. that each attribute value can be missing or not and this does not influence in any way the missingness of other attributes. Which of the following statements is/are true?

  $$prob = (1 - 0.005)^{1000} = 0.006654$$

  The probability that a record contains no missing values is equal to 0.006654 (six digits representation)
  The expected number of records without any missing value is equal to 665 (rounded down to the nearest integer)

- Which of the following statements about standardization and normalization is/are true? **Standardization transform the processed attribute in such a way that the resulting attribute has zero mean and unit standard deviation.**
  They can be both applied to continuous attributes.

- Given the following set of values for the attribute Yearly Income (Euro) 2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234 which are the percentiles of order 1/3 and 2/3? **(1,667; 2,234)**

- Assume you are given a data base containing the following attribute, Credit Card owner. Assume that Credit Card owner can take values yes, no. Assume you have the following 10 records for the Credit Card owner attribute yes, yes, no, yes, no, no, no, no, no, no. Assume you want to form a sample contisting of 4 records. Then, you apply random sampling with stratification (stratified sampling) using the equal number option. Which of the following samples is a valid one? **(yes, yes, no, no)**
  (no, yes, yes, no)

- Assume you are given a data base containing an attribute Food Quality which can take the following values {excellent, very good, good, average, bad, terrible}. Assume you want to apply binarization to the Food Quality attribute, How many new attributes do you need to create to achieve this goal, in case you can accept to induce correlation among them? **3**

2 Introduction to classification

-
-
-

-
-
-
-
-
-
-
-
-
-
-
-

3 Classifier Evaluation

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

4 Class Imbalance

5 Introduction Clustering

-
-
-
-

6 Clustering Evaluation

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

7 Association Analysis

-
-
-
-
-
-
-
-

-
-
-
-
-
-
-
-
-