DASHBOARD / I MIEI CORSI / APPELLI DI CLAUDIO SARTORI / SEZIONI / DATA MINING M / MACHINE LEARNING / MACHINE LEARNING EXAM - MODULE OF 91249 - MACHINE LEARNING AND DEEP LEARNING I.C.

Iniziato	Monday, 9 January 2023, 09:56
Stato	Completato
Terminato	Monday, 9 January 2023, 11:36
Tempo impiegato	1 ora 39 min.
Valutazione	Non ancora valutato

Domanda **1**

Completo

Punteggio max.: 20,00

The task is described in this document.

The data file is here. The file with the feature names is here

Upload only your notebook, not the data. Please name your notebook according to the directions given in the document linked above.

andrea_zecca3.ipynb

Domanda **2**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which is the main reason for the *standardization* of numeric attributes?

Scegli un'alternativa:

- ● a. Map all the numeric attributes to a new range such that the mean is zero and the variance is one. ✓
- ○ b. Change the distribution of the numeric attributes, in order to obtain gaussian distributions
- ○ c. Remove non-standard values
- ○ d. Map all the nominal attributes to the same range, in order to prevent the values with higher frequency from having prevailing influence

Your answer is correct.

La risposta corretta è: Map all the numeric attributes to a new range such that the mean is zero and the variance is one.

Domanda **3**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Given the two binary vectors below, which is their similarity according to the Simple Matching Coefficient?

a	b	c	d	e	f	g	h	i	j
1	0	0	0	1	0	1	1	0	1
1	0	1	1	1	0	1	0	1	0

Scegli un'alternativa:

- ● a. 0.5 ✓
- ○ b. 0.3
- ○ c. 0.2
- ○ d. 0.1

Risposta corretta.

La risposta corretta è: 0.5

Domanda **4**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What is the *single linkage?*

Scegli un'alternativa:

- ● a. A method to compute the distance between two sets of items, it can be used in hierarchical clustering ✓
- ○ b. A method to compute the distance between two objects, it can be used in hierarchical clustering
- ○ c. A method to compute the distance between two classes, it can be used in decision trees
- ○ d. A method to compute the separation of the objects inside a cluster

Your answer is correct.

La risposta corretta è: A method to compute the distance between two sets of items, it can be used in hierarchical clustering

Domanda **5**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Given the definitions below:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

which of the formulas below computes the accuracy of a binary classifier?

Scegli un'alternativa:

● a. (TP + TN) / (TP + FP + TN + FN) ✓

○ b. TP / (TP + FN)

○ c. TN / (TN + FP)

○ d. TP / (TP + FP)

Risposta corretta.

La risposta corretta è: (TP + TN) / (TP + FP + TN + FN)

Domanda **6**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What is the *Gini Index?*

Scegli un'alternativa:

● a. An impurity measure of a dataset alternative to the *Information Gain* and to the *Misclassification Index* ✓

○ b. An accuracy measure of a dataset alternative to the *Information Gain* and to the *Misclassification Index*

○ c. An impurity measure of a dataset alternative to *overfitting* and *underfitting*

○ d. A measure of the *entropy* of a dataset

Your answer is correct.

La risposta corretta è: An impurity measure of a dataset alternative to the *Information Gain* and to the *Misclassification Index*

Domanda **7**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

In a decision tree, the number of objects in a node...

Scegli un'alternativa:

● a. ...is smaller than the number of objects in its ancestor ✓

○ b. ...is smaller than or equal to the number of objects in its ancestor

○ c. ...is bigger than the number of objects in its ancestor

○ d. ...is not related to the number of objects in its ancestor

Risposta corretta.

La risposta corretta è: ...is smaller than the number of objects in its ancestor

Domanda **8**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the following is a base hypothesis for a bayesian classifier?

Scegli un'alternativa:

● a. The attributes must be statistically independent inside each class ✓

○ b. The attributes must be statistically independent

○ c. The attributes must have zero correlation

○ d. The attributes must have negative correlation

Risposta corretta.

La risposta corretta è: The attributes must be statistically independent inside each class

Domanda **9**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

With reference to the total *sum of squared errors* and *separation* of a clustering scheme, which of the statements below is true?

● a. They are strictly correlated, if, changing the clustering scheme, one increases, then the other decreases ✓

○ b. It is possible to optimise them (i.e. minimise SSE and maximise SSB) separately

○ c. They are strictly correlated, if, changing the clustering scheme, one increases, then the other does the same

○ d. They are two ways to measure the same thing

Your answer is correct.

La risposta corretta è:
They are strictly correlated, if, changing the clustering scheme, one increases, then the other decreases

Domanda **10**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the statements below is true? (One or more)

Scegli una o più alternative:

☑ a. Sometimes k-means stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters. ✓

□ b. K-means always stops to a configuration which gives the minimum distortion for the chosen value of the number of clusters.

☑ c. K-means is quite efficient even for large datasets ✓ No, k-means finds a local minimum of the distortion for an assigned number of clusters

☑ d. K-means is very sensitive to the initial assignment of the centers ✓ No, being based on distances, if the number of attributes is very large k-means is prone to the *curse of dimensionality*

Your answer is correct.

Le risposte corrette sono: Sometimes k-means stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters., K-means is quite efficient even for large datasets, K-means is very sensitive to the initial assignment of the centers

Domanda **11**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the statements below is true? (One or more)

Scegli una o più alternative:

- ☑ a. Sometimes DBSCAN stops to a configuration which does not include any cluster ✓
- ☐ b. DBSCAN always stops to a configuration which gives the optimal number of clusters
- ☑ c. DBSCAN can give good performance when clusters have concavities ✓
- ☑ d. Increasing the radius of the neighbourhood can decrease the number of noise points ✓

Your answer is correct.

Le risposte corrette sono: Sometimes DBSCAN stops to a configuration which does not include any cluster, DBSCAN can give good performance when clusters have concavities, Increasing the radius of the neighbourhood can decrease the number of noise points

Domanda **12**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What is the meaning of the statement: "*the support is anti-monotone*"?

Scegli un'alternativa:

- ● a. The support of an itemset never exceeds the support if its subsets ✓
- ○ b. The support of an itemset never exceeds the support if its supersets
- ○ c. The support of an itemset is always smaller than the support of its subsets
- ○ d. The support of an itemset is always smaller than the support of its supersets

Risposta corretta.

La risposta corretta è: The support of an itemset never exceeds the support if its subsets

Domanda **13**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Consider the transactional dataset below

ID	Items
1	A,B,C
2	A,B,D
3	B,D,E
4	C,D
5	A,C,D,E

Which is the *confidence* of the rule A,C ⇒ B?

Scegli un'alternativa:

● a. 50% ✓ 1 / 2

○ b. 100%

○ c. 40%

○ d. 20%

Risposta corretta.

La risposta corretta è: 50%

Domanda **14**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What is the **coefficient of determination R^2**?

● a. Provide an index of goodness for a linear regression model ✓

○ b. Measure the amount of error in a linear regression model

○ c. Measure the amount of error in a regression model

○ d. An index of goodness for a classification model

Your answer is correct.

La risposta corretta è: Provide an index of goodness for a linear regression model

Domanda **15**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What does K-means try to minimise?

Scegli un'alternativa:

● a. The *distortion*, that is the sum of the squared distances of each point with respect to its centroid ✓

○ b. The *separation*, that is the sum of the squared distances of each cluster centroid with respect tho the global centroid of the dataset

○ c. The *distortion*, that is the sum of the squared distances of each point with respect to the points of the other clusters

○ d. The *separation*, that is the sum of the squared distances of each point with respect to its centroid

Risposta corretta.

La risposta corretta è: The *distortion*, that is the sum of the squared distances of each point with respect to its centroid

Domanda **16**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the activities below is part of "Business Understanding" in the CRISP methodology?

○ a. Which machine learning functions are necessary for my problem?

○ b. Which data are available?

○ c. Which data must be collected with a specific campaign?

● d. Which are the resources available (manpower, hardware, software, ...) ✓

Your answer is correct.

La risposta corretta è:
Which are the resources available (manpower, hardware, software, ...)