Nombre Deybit Gutierrez Fecha 03/09/2021

Data Science Skills Test

Read carefully each question, then answer.

1. Select the statement that limits both x and y axes to the interval [0, 6]? puntos: 1

plt.axis([0, 6, 0, 6])

plt.xlim(0, 6)
plt.ylim(0, 6)
plt.xylim(0, 6)

2. In NumPy, what does the shape (2,3,2) indicate? puntos: 1

2 rows, 3 columns, 2 ranks
2 rows, 3 ranks, 2 columns
2 rows, 3 columns, 2 elements
2 rows, 3 elements, 2 ranks

3. In statistics, a Type II error occurs when: puntos: 1

a null hypothesis is rejected but should not be rejected
a null hypothesis is not rejected but should be rejected
a test statistic is incorrect
a hypothesis is chosen incorrectly

4. In BeautifulSoup, what are the options to search a web tree? puntos: 2

find()
findall() -> **find_all()**
search()
searchall()

5. In Pandas series, data can be accessed through different functions such as: puntos: 2

loc()
iloc()
access()
get()

6. What will be the result in vector addition if labels are not found in a series? puntos: 1

Will be marked as Zeros
Will be skipped
Will be marked as NaN
Will throw an exception, index not found

7. What does the following code do? import pandas as pd df = pd.read_csv('log-access_file.csv') puntos: 1

Imports pandas directory
Loads a csv file into a DataFrame
Initiates a program in pandas
Ends a program in pandas

8. In NLP, stemming is a technique to: puntos: 1

split words, phrases, idioms

map valid word root

discover topics in a collection of documents

determine where one word ends and other begins

9. If a dataframe is imported, and there is no index for its values and dates, how do you create one? puntos: 1

df.set_index('Date' = True)

df.set_index('Date'= true, inplace = True)

df.setting_index('Date'= true, in place = True)

10. A classifier that predicts if an image contains only a cat, a dog, or a llama produced the following confusion matrix: What is the accuracy of the model, in percentages? puntos: 2

		True values		
		Dog	Cat	Llama
Predicted values	Dog	14	2	1
	Cat	2	12	3
	Llama	5	2	19

75%

0.75%

70%

0.71%

98%

Accuracy can only be calculated on binary problems.

11. What is recall and precision? puntos: 2

Precisión: Es la proporción que resulta de tomar los Verdaderos positivos entre los datos recuperados.

Recall: Proporción resultante de tomar los verdaderos positivos y los elementos relevantes

12. What is a normal distribution? puntos: 2

Es una distribución de probabilidad continua. Su gráfica de distribución tiene forma de campana, muchos de los fenómenos de la probabilidad se pueden expresar utilizando una distribución normal.

13. What are some methods you might use to fill in missing data and what are the consequences if you fill in missing data uncarefully? puntos: 2

Podemos rellenar valores utilizando la media de los valores existentes, eliminar aquellas columnas cuya cantidad de datos faltantes sea muy grande, predecir los valores faltantes. Como consecuencia de no tratar de forma correcta los datos podemos hacer que el modelo pierda precisión y perder data relevantes.

14. What is PCA and how can it help? puntos: 2

PCA es un método de reducción dimensional: Análisis de los componentes principales: Estandarizar un rango de variables iniciales continuas. Puede ayudarnos para calcular la matriz de covarianza para identificar relaciones, calcular valores y vectores propios de la matriz de covarianza para identificar los componentes principales. Reposicionar los datos a lo largo de los ejes principales.

Tomamos un dataset grande, se simplifica en uno más pequeño pero que sigue conteniendo la información relevante. Al hacerse más simple, ayuda con el análisis para los algoritmos de machine learning.

15. How do you choose the k-value in K Means Clustering without looking at the clusters? puntos: 2

16. Explain what is overfitting. puntos: 2

Sobreajuste: Cuando un modelo de inteligencia artificial aprende muy bien el conjunto de entrada de entrenamiento incluyendo el ruido de la data, lo que provoca que esta fluctuación en los datos sea tomada como parte real de los datos y afecta negativamente el análisis de la nueva data.

17. What is the main difference between supervised and unsupervised learning? puntos: 2

Aprendizaje supervisado: Los modelos aprenden con data clasificada, previamente tratada por un humano y arroja resultados esperados.

Aprendizaje no supervisado: No contamos con datos etiquetados para el análisis, solo conocemos los datos de entrada, describe su estructura y con el modelo se intenta conseguir un tipo de relación/organización, sin garantizar que estos resultados tengan utilidad o significado.

18. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it? puntos: 2

Es posible que el modelo sufra de Overfitting. Con un porcentaje tan alto de precisión muy probablemente tome la mayoría de los casos como casos favorables tanto los falsos negativos que deberían ser descartados como los verdaderos negativos.

Para este caso es recomendable tener un high recall en lugar de high precision, para esto podemos ajustar los parametros y la probabilidad para alcanzar el recall.