Curso Actualízate – Machine Learning Gijón (Módulo 5)

Nombre:Gulnat

Apellidos: Pettit

Fecha:07.07.2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. ¿Qué es un ETL?

   Se refiere a un proceso de Extract-Transform-Load, cuando extraemos los datos de fuentes diferentes, luego transformamos (limpiamos, validación, eliminar incompletos), luego ponemos al sistema or db para processar y analizar

2. Enumera las características de Python explicando cada una de ellas (menciona 4 al menos):

   -Multiplataforma -se usa en diferentes plataformas:Windows,Unix.Linux etc. No necesitamos cambiar el codigo

   - De alto nivel, facil para aprender y codigar

   - Tiene una cantidad enorme de librerías. No necesitamos escribir nuestros mismos módulos o funciones

   - Orientado al objeto. Python tiene la concepción de classes, etc.

3. ¿Cuáles son los tipos primitivos en Python y qué valores pueden contener cada uno de ellos? Pista: Son 3 tipos.

   - Entre los tipos primitivos de variables son: texto, número y Booleanos.

     Texto: 'abril', 'Hi, people'

     Número: entero (2, -7), flotantes (4.3, 0.00006)

     Booleanos: True/False, 0/1

4. Menciona alguna estructura de datos más compleja que los tipos primitivos que conozcas.

Listas: de texto, número o mixtas. Son mutables, podemos añadir, eliminar los elementos.

mes=['mayo', 'junio']

5. Escribe la sintaxis para crear variables.

   Usamos '='

   ciudad='Londres'

   x=4

6. Escribe la sintaxis para crear funciones.

   def - define la función sin/con parametres

   def nombreFuncion():

   instrucción 1

   instrucción 2

7. Escribe la sintaxis para llamar a variables.

   Para llamar variable necesitamos utilizar el nombre de la variable

   Ej. mes='julio' # definir variable

   print(mes) #imprimir variable mes con valor 'julio'

8. Escribe la sintaxis para llamar a funciones.

   nombreFuncion() - sin parametres

   nombreFuncion(6) - con parametre

   ej. def cuadrado(n): #creamos una función

   return n*n

   — —

   cuadrado(6) # llamamos la función cuadrado(n)

9. Explica con tus palabras para qué sirven las librerías: Pandas y Numpy.

Pandas se sirve para análisis de datos y proporciona estructura de datos (se organiza en filas/columnas - dataframe) y ofrece las funciones de limpieza, agrupar, visualización y otras.

NumPy es una librería de matemáticas y ofrece muchas funciones de operadores como pi, sqrt, integral. También es posible trabajar con matrices de multidimension.

10. ¿Cómo representamos el valor vacío en Python?

    variable = None # no hay un valor o no es válido

11. ¿Qué es un IDE?

Integrated Development Environment es una aplicación de software que ayuda a los programadores a desarrollar código de software de manera eficiente. Aumenta la

productividad de desarrolladores a través del uso de funciones como edición, creación, prueba y empaquetado de software en una aplicación fácil de usar.

12. ¿Qué es el CRISP-DM?

    Cross-Industry Standard Process for Data Mining

    Es un modelo de proceso estándar que se usa en la minería y análisis de datos. Proporciona un enfoque sistemático para implementar los proyectos de mineria de los datos desde el principio hasta el final, y cubre 6 fases principales: comprensión de negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

13. Explica cada uno de los tipos de Machine Learning.

    1.Supervisado - tu pasas entrada y salida, por ej las fotos de gatos y perros como input y output (etiquetado)

    El conjunto de datos para entrenar el algoritmo incluye la solución que el algoritmo debería de dar a dichos datos.

    2.No supervisado - tu pasas entrada y no decirle al modelo lo que tiene que hacer, por ej solo las fotos de perros y gatos de input

    El conjunto de datos no tiene porque etiquetado, el modelo intenta aprender sin que le digan que tiene que aprender

    3.Semi Supervisado - ej Chat GPT , si recibes muchos dislikes (penalties), mejoras.

    Un agente que es capaz de observar el entorno, realizar acciones y recibir premios o penalizaciones como respuesta a sus actos. An agent that is able to observe the environment, perform actions and receive rewards or penalties in response to its actions.

    4.Aprendizaje reforzado - no usan datos para aprender, aprobar muchas veces, hay premios o castiga.

    Es una mezcla de métodos de aprendizaje. Si el modelo falla, es corregido por el humano y el modelo aprende.

14. ¿Con qué tipo de Machine Learning hemos estado trabajando nosotros?

    Supervisado porque dimos las valores de X e Y.

15. ¿Qué es el Prophet? ¿Y Scikit Learn, Keras y Tensorflow?

    Prophet es una lib re FB para el pronóstico de series temporales.

    Scikit-Learn es una lib de aprendizaje automático en Python, que usamos para clasificación, regresión, agrupación.

    Keras es una lib de aprendizaje profundo de código abierto, que se usa para crear y entrenar redes neuronales artificiales.

TensorFlow es una lib de Google que permite entrenar redes neuronales profundas de procesamiento de imágenes, lenguaje natural, etc.

16. ¿Qué significa en ML regresión? Responde brevemente.

    Es un método estadístico para modelar la relación entre una variable dependiente y una o más variables independientes. Se usa para hacer predicciones sobre nuevos datos en los que no se conoce el valor de la variable dependiente.

17. ¿Cuál es el tipo de predicción más sencilla, pero a la vez la más usada?

    Regression es más sencilla

18. ¿Cuáles son los problemas principales del Machine Learning? Justifica tu respuesta.

    Datos insuficientes - necesitamos obtener más datos de las fuentes diferentes, si es una fuente no es suficiente.

    Variables irrelevantes - columnas innecesarias, procesado de que require mucho tiempo

    Overfitting - este problema cuando modelo funciona con datos de entrenamiento, pero no funciona con los datos de test.

    Underfitting - un ej de esto es el problema con tipos de variables, cuando tenemos números en en caso de texto

    Datos de mala calidad - necesitamos limpiar primero

    Todas estos problemas afectan el proceso de análisis y podrían dar faltas y conclusiones incorrectas, por esto hay que eliminarlos.

19. ¿Qué es el residuo en Machine Learning?

    Es la diferencia entre el valor real Y_test y el valor predicho Y_pred.

For me))

In the context of machine learning, the term "residue" is not commonly used. However, there is a related concept called "residuals" that is frequently discussed in machine learning and statistical modeling.

Residuals represent the differences between the observed values and the predicted values produced by a machine learning model. In other words, they are the errors or discrepancies between the actual data points and the values predicted by the model.

Residuals are often used to evaluate the performance and accuracy of a machine learning model. By analyzing the pattern of residuals, you can determine if the model is adequately capturing the underlying relationships in the data. Ideally, the residuals should be random

and exhibit no discernible pattern. If there is a clear pattern or structure in the residuals, it suggests that the model is not fully capturing all the relevant information in the data.

Analyzing the residuals can help identify issues such as underfitting or overfitting. Underfitting occurs when the model is too simple and fails to capture the complexities of the data, leading to large and systematic residuals. Overfitting, on the other hand, happens when the model is too complex and starts to memorize the noise or idiosyncrasies in the training data, resulting in small but erratic residuals.

To summarize, while "residue" is not a widely used term in machine learning, "residuals" are an important concept that helps assess the performance and quality of machine learning models.