INTELIGENCIA ARTIFICIAL

UNINPAHU
INGENIERÍA Y TECNOLOGÍAS DE LA INFORMACIÓN
INGENIERÍA DE SOFTWARE

DANIEL STIVEN GARCIA MIRANDA
2021

TALLER 1

STATEMENT OF ACCOMPLISHMENT

#18,124,524

HAS BEEN AWARDED TO

DANIEL STIVEN GARCIA MIRANDA

FOR SUCCESSFULLY COMPLETING

Introducción a Python

COMPLETED ON

FEB 28, 2021

datacamp

Jonathan Cornelissen, CEO

UNINPAHU

INSTITUCIÓN UNIVERSITARIA

1. Investigar sobre las estructuras de datos de árbol y grafo y describir:

- La definición formal de la estructura de datos

En programación, una estructura de datos es una forma de organizar un conjunto de datos elementales con el objetivo de facilitar su manipulación. Un dato elemental es la mínima información que se tiene en un sistema.

Una estructura de datos define la organización e interrelación de éstos y un conjunto de operaciones que se pueden realizar sobre ellos. Las operaciones básicas son:

Alta, adicionar un nuevo valor a la estructura.

Baja, borrar un valor de la estructura.

Búsqueda, encontrar un determinado valor en la estructura para realizar una operación con este valor, en forma SECUENCIAL o BINARIO (siempre y cuando los datos estén ordenados).

Otras operaciones que se pueden realizar son:

Ordenamiento, de los elementos pertenecientes a la estructura.

Apareo, dadas dos estructuras originar una nueva ordenada y que contenga a las apareadas.

Cada estructura ofrece ventajas y desventajas en relación a la simplicidad y eficiencia para la realización de cada operación. De esta forma, la elección de la estructura de datos apropiada para cada problema depende de factores como la frecuencia y el orden en que se realiza cada operación sobre los datos.

- Las partes que componen dicha estructura de datos
- Casos de uso de dicha estructura de datos

La estructura de datos que se aplican en los equipos y que se encuentran basados sobre otras estructuras simples:

Vector

- Es un conjunto de elementos que se encuentran estructurado de una forma especial y especifica
- De una forma general cada elemento que se disponga son del mismo tipo
- Se puede acceder a estos elementos mediante l aplicación de un entero como un índice de manera que se tenga que señalar el elemento que se desee

- Puede presentar algunas implementaciones básicas las cuales pueden dar las palabras de la memoria adyacente de los elementos que se encuentre en cada rearreglo
- Con cada modificación que se realice se puede cambiar o variar el tamaño de la longitud
- También puede disponer de una longitud fija determinada

Vector Asociativo

- Es una variable caracterizada por ser flexible
- Su flexibilidad es mayor que el de una matriz
- Da la opción de agregar pares nombre valor
- También permite eliminar pares nombre valor
- Cuenta con una tabla de hash
- Facilita el arreglo asociativo que se realiza

Registro

- Es también conocido como estructura o como tupia
- Consiste en una estructura de datos que se pueden anexar
- Basado en un valor el cual dispone de otros valores
- Generalmente su forma básica es un número fijo
- Su valor puede ser en secuencia
- Cuenta con un índice por nombres para facilitar la búsqueda de valores y variables caracterices
- Dispone de elementos que son denominados como Campos y también como Celdas

Unión

- Es una estructura de datos que señala de forma esencial el conjunto de tipos de datos que pueden ser guardadas en un lugar en especifico
- Dispone de algunas funciones diferentes al Registro
- Cuenta de un solo valor que se aplica a la vez
- Permite asignar el espacio requeridos para almacenar los tipos de datos, es por ello que dicho lugar debe ser suficiente para contener los datos y la información especifica

Tipo Variante

- Se conoce como el registro variante
- También es llamado como la unión discriminada
- Dispone de un campo adicional
- Se encarga de indicar y resaltar el tipo que presente a tiempo real

Conjunto

- Es un tipo de datos abstracto
- Da la capacidad de guardar valores específicos
- No requiere que al guardar los datos se disponga de un orden especifico y particular
- Tampoco almacena valores que se encuentre duplicados

Multiconjunto

- Es otro tipo de datos abstracto
- Se encarga de guardar y ubicar los diversos valores específicos dados
- No almacena los valores por un orden particular, sino que los almacena a medida que se ingrese
- Permite almacenar valores que estén repetidos

Grafo

- Es una estructura de datos que se encuentra conectada
- Se encuentra constituidas por nodos
- Cada nodo que dispone posee un valor especifico
- También los nodos contienen referencias de otros nodos
- Tiene la capacidad de aplicarse para dar una representación de redes
- Puede dar referencia entre cada nodo
- Dispone de algunas conexiones las cuales contienen direcciones, es decir, algunos de puntos de entrada y salida

Árbol

- Consiste en un caso diferente o especifico de grafo
- Se encuentra en la aplicación de los ciclos que no se permiten
- Dispone de un camino a partir de un nodo hasta otro nodo
- El nodo de partida se conoce como raíz
- Presenta una colección de árboles el cual es comúnmente conocida como bosque

Clase

- Es una plantilla especifica
- Aplicada para la elaboración de objetos de datos
- Está basado en un modelo que es predefinido
- Se emplea como una representación abstracta de conceptos
- Introducen diversos campos como lo son los registros y las operaciones
- Da la posibilidad de realizar una consulta por el valor de dichos campos
- También puede cambiar los valores específicos

Conceptos básicos de Machine Learning

- Investigar sobre el concepto de aprendizaje de máquina y definir los siguientes conceptos:

I. Dataset

Un conjunto de datos o dataset corresponde a los contenidos de una única tabla de base de datos o una única matriz de datos de estadística, donde cada columna de la tabla representa una variable en particular, y cada fila representa a un miembro determinado del conjunto de datos que estamos tratando.

II. Aprendizaje Supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

III. Aprendizaje no Supervisado

Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.

El aprendizaje no supervisado se puede usar en conjunto con la Inferencia bayesiana para producir probabilidades condicionales (es decir, aprendizaje supervisado) para cualquiera de las variables aleatorias dadas. El Santo Grial del aprendizaje no supervisado es la creación de un código factorial de los datos, esto es, un código con componentes estadísticamente independientes. El aprendizaje supervisado normalmente funciona mucho mejor cuando los datos iniciales son primero traducidos en un código factorial.

IV. Datos de entrenamiento y datos de prueba

Los datos de entrenamiento o “training data” son los datos que usamos para entrenar un modelo. La calidad de nuestro modelo de aprendizaje automático va a ser directamente proporcional a la calidad de los datos. Por ello las labores de limpieza, depuración o “data wrangling” consumen un porcentaje importante del tiempo de los científicos de datos.

Los datos de prueba, validación o “testing data “son los datos que nos “reservamos” para comprobar si el modelo que hemos generado a partir de los datos de entrenamiento “funciona”. Es decir, si las respuestas predichas por el modelo para un caso totalmente nuevo son acertadas o no.

Es importante que el conjunto de datos de prueba tenga un volumen suficiente como para generar resultados estadísticamente significativos, y a la vez, que sea representativo del conjunto de datos global.

Normalmente el conjunto de datos se suele repartir en un 70% de datos de entrenamiento y un 30% de datos de test, pero se puede variar la proporción según el caso. Lo importante es ser siempre conscientes de que hay que evitar el sobreajuste u “overfitting”.

V. Cross Validation

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las

medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

VI. Overfitting

El sobreajuste (también es frecuente emplear el término en inglés overfitting) es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. El algoritmo de aprendizaje debe alcanzar un estado en el que será capaz de predecir el resultado en otros casos a partir de lo aprendido con los datos de entrenamiento, generalizando para poder resolver situaciones distintas a las acaecidas durante el entrenamiento. Sin embargo, cuando un sistema se entrena demasiado (se sobreentrena) o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo. Durante la fase de sobreajuste el éxito al responder las muestras de entrenamiento sigue incrementándose mientras que su actuación con muestras nuevas va empeorando.

VII. Red Neuronal (Artificial Neural Networks)

Las redes neuronales artificiales (también conocidas como sistemas conexionistas) son un modelo computacional el que fue evolucionando a partir de diversas aportaciones científicas que están registradas en la historia. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida.

Cada neurona está conectada con otras a través de unos enlaces. En estos enlaces el valor de salida de la neurona anterior es multiplicado por un valor de peso. Estos pesos en los enlaces pueden incrementar o inhibir el estado de activación de las neuronas adyacentes. Del mismo modo, a la salida de la neurona, puede existir una función limitadora o umbral, que modifica el valor resultado o impone un límite que no se debe sobrepasar antes de propagarse a otra neurona. Esta función se conoce como función de activación.

Estos sistemas aprenden y se forman a sí mismos, en lugar de ser programados de forma explícita, y sobresalen en áreas donde la detección de soluciones o características es difícil de expresar con la programación convencional. Para realizar este aprendizaje automático, normalmente, se intenta minimizar una función de pérdida que evalúa la red en su total. Los valores de los pesos de las neuronas se van actualizando buscando reducir el valor de la función de pérdida. Este proceso se realiza mediante la propagación hacia atrás.

VIII. Máquina de soporte vectorial (Support Vector Machines)

Las máquinas de vectores de soporte o máquinas de vector soporte (del inglés Support Vector Machines, SVM) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T.

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las

clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

IX. Bosques Aleatorios (Random Forests)

Random forest (o random forests) también conocidos en castellano como '"Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

El algoritmo para inducir un random forest fue desarrollado por Leo Breiman y Adele Cutler y Random forests es su marca de fábrica. El término aparece de la primera propuesta de Random decision forests, hecha por Tin Kam Ho de Bell Labs en 1995. El método combina la idea de bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de decisión con variación controlada.

La selección de un subconjunto aleatorio de atributos es un ejemplo del método random subspace, el que, según la formulación de Ho, es una manera de llevar a cabo la discriminación estocástica propuesta por Eugenio Kleinberg.

En muchos problemas el rendimiento del algoritmo random forest es muy similar a la del boosting, y es más simple de entrenar y ajustar. Como consecuencia, el Random forest es popular y ampliamente utilizado