Módulo: ING1906 - Python aplicado a ciencia de datos - (A49)

Actividad: Reto de aprendizaje 5. Síntesis: Herramientas de Python aplicadas a la ciencia de datos

Nombre: Roberto Mora Balderas

Asesor: Adriana Ávalos Vargas

Fecha: 14 de mayo de 2023

Título: Herramientas de Python para la visualización de datos, limpieza de datos, modelos de regresión y clasificación mediante técnicas de machine learning, y el funcionamiento de las redes neuronales.

Python es un lenguaje de programación muy popular en el ámbito del análisis de datos y machine learning, debido a la gran cantidad de librerías y herramientas disponibles que facilitan estas tareas. A continuación, se explican algunas de las principales herramientas que Python ofrece para la visualización de datos, la limpieza de datos y la creación de modelos de regresión y clasificación mediante técnicas de machine learning, así como el funcionamiento de las redes neuronales y sus elementos.

Visualización de datos:

Python cuenta con varias librerías que permiten crear gráficos y visualizaciones de datos de manera sencilla. Algunas de las herramientas más populares son:

- Matplotlib: permite crear gráficas 2D y 3D, histogramas, diagramas de barras, entre otros tipos de gráficos.
- Seaborn: esta librería está basada en Matplotlib y proporciona un conjunto de gráficos más avanzados, como diagramas de dispersión con ajustes de regresión, mapas de calor y gráficos de distribución.
- Plotly: es una librería que permite crear gráficos interactivos y animados que se pueden compartir y editar en línea.

La elección de la herramienta de visualización dependerá de los datos y el tipo de gráfico que se quiera crear. Siendo el histograma el indicado para mostrar distribuciones, el diagrama de caja para mostrar la proporción de las poblaciones y la gráfica de línea para mostrar el cambio con respecto al tiempo.

Limpieza de datos:

Antes de crear modelos de machine learning es necesario preparar y limpiar los datos. Python ofrece diversas herramientas para esta tarea, algunas de ellas son:

- Pandas: es una librería que permite manipular y analizar datos en forma de tablas llamadas DataFrames. Con Pandas se pueden realizar operaciones como eliminar filas con datos faltantes, cambiar el tipo de dato de las variables y seleccionar y filtrar datos de una tabla, además de manejar los valores nulos o valores extremos.
- NumPy: es una librería que proporciona herramientas para trabajar con matrices y arrays numéricos. Es muy útil para operaciones matemáticas y estadísticas en grandes conjuntos de datos. Además de generar datos escalados y normalizados.

Modelos de regresión y clasificación mediante técnicas de machine learning:

Python cuenta con varias librerías para crear modelos de regresión y clasificación mediante técnicas de machine learning, entre ellas se encuentran:

- Scikit-learn: es una librería muy popular que contiene herramientas para realizar tareas de aprendizaje supervisado y no supervisado. Proporciona modelos de regresión lineal, regresión polinomial, árboles de decisión, redes neuronales, entre otros.
- TensorFlow: es una librería de código abierto desarrollada por Google para crear modelos de machine learning y deep learning. Proporciona herramientas para construir y entrenar redes neuronales.
- PyTorch: es una librería de código abierto desarrollada por Facebook para crear modelos de machine learning y deep learning. Proporciona una interfaz fácil de usar y permite construir y entrenar redes neuronales.
- Keras: Es una biblioteca de aprendizaje profundo de alto nivel que se ejecuta en TensorFlow. Proporciona una API sencilla para construir y entrenar modelos de aprendizaje profundo para regresión y clasificación.

Los métodos de regresión se utilizan para predecir un valor numérico continuo en función de una o más variables independientes. La regresión lineal es una técnica de modelado estadístico que utiliza una relación lineal entre la variable independiente y la variable dependiente. La regresión polinomial es un tipo de regresión en el que la relación entre la variable independiente y la variable dependiente se modela como una función polinómica.

Clasificación:

Los métodos de clasificación se utilizan para predecir una clase o categoría a partir de un conjunto de variables de entrada. La clasificación es una técnica de aprendizaje supervisado que se utiliza para predecir una clase o categoría a partir de un conjunto de variables de entrada.

El aprendizaje supervisado, se entrena un modelo utilizando un conjunto de datos etiquetados, donde se conocen las respuestas o las salidas esperadas para cada muestra. El modelo aprende a mapear las características de entrada a las salidas conocidas y luego puede predecir las salidas para nuevos datos. Los algoritmos de aprendizaje supervisado incluyen regresión lineal, regresión logística, regresión polinomial, SVM, árboles de decisión, entre otros.

El aprendizaje no supervisado el modelo se entrena utilizando un conjunto de datos no etiquetados, es decir, no se conocen las salidas esperadas. El objetivo principal del aprendizaje no supervisado es descubrir patrones, estructuras o grupos inherentes en los datos. Algunos algoritmos de aprendizaje no supervisado incluyen el clustering (agrupamiento), análisis de componentes principales (PCA), y algoritmos de reducción de dimensionalidad.

Los métodos de clasificación se utilizan para predecir una categoría o clase específica a la que pertenece un dato, mientras que los métodos de regresión se utilizan para predecir un valor numérico o continuo. La elección entre estos métodos depende del tipo de problema y de los datos disponibles.

Redes neuronales:

Las redes neuronales son un tipo de modelo de aprendizaje profundo que se inspira en la estructura y función del cerebro humano. Las redes neuronales consisten en una capa de entrada, una o varias capas ocultas y una capa de salida. Cada capa está compuesta por neuronas artificiales que están conectadas a las neuronas de la capa siguiente mediante conexiones sinápticas ponderadas.

. Las redes neuronales se componen de varios elementos clave:

- Neuronas: Son las unidades fundamentales de una red neuronal. Cada neurona tiene pesos asociados que determinan la importancia de las entradas y una función de activación que determina la salida de la neurona.

- Capas: Las neuronas se organizan en capas, que pueden ser de diferentes tipos. La capa de entrada recibe los datos de entrada, las capas intermedias (llamadas capas ocultas) realizan cálculos intermedios y la capa de salida produce la salida final de la red.

- Conexiones: Las neuronas están interconectadas mediante conexiones, que representan los pesos que determinan la influencia de una neurona en otra. Cada conexión tiene un peso asociado que se ajusta durante el entrenamiento de la red.

- Funciones de activación: Las funciones de activación determinan la salida de una neurona en función de sus entradas y pesos. Algunas funciones de activación comunes incluyen la función sigmoide, la función ReLU (Rectified Linear Unit) y la función softmax.

- Retropropagación: Es un algoritmo utilizado para entrenar redes neuronales. Consiste en propagar el error desde la capa de salida hacia las capas anteriores, ajustando los pesos de las conexiones en función del error calculado. La retropropagación se repite iterativamente hasta que la red aprende a producir salidas precisas.

Las redes neuronales son capaces de aprender patrones y características complejas en los datos, lo que las hace adecuadas para una amplia gama de tareas de machine learning, como reconocimiento de imágenes, procesamiento del lenguaje natural, análisis de datos secuenciales y más. Su capacidad para modelar relaciones no lineales y adaptarse a los datos las convierte en una herramienta poderosa en el campo del aprendizaje automático.

Conclusiones

Actualmente el área de data ha ido creciendo y siendo cada día mas importante en la vida cotidiana, desde decisiones basadas en información para las empresas, desde algoritmos para las redes sociales que utilizamos y en los servicios de streaming. Como podemos observar Python es una herramienta bastante robusta en lo que concierne al área de datos, nos ofrece bastante posibilidades con todas las librerías que soporta, siendo una opción muy importante para el desarrollo de todas estas nuevas tecnologías.

Propuesta de trabajo - Cáncer de mama

Antecedentes del problema:

En el campo de la salud, es de vital importancia tener herramientas que permitan identificar patologías de manera temprana para poder brindar tratamientos oportunos y mejorar la calidad de vida de los pacientes. Una de estas patologías es el cáncer de mama, que afecta principalmente a mujeres de todo el mundo.

Planteamiento del problema:

El cáncer de mama es una de las principales causas de muerte en mujeres en todo el mundo. En la actualidad, existen varias técnicas de detección temprana, pero es necesario contar con herramientas que permitan identificar patrones de la enfermedad para poder brindar tratamientos efectivos. Por lo tanto, se requiere una aplicación que permita analizar datos clínicos de pacientes con cáncer de mama y realizar predicciones mediante técnicas de machine learning.

Justificación de la aplicación a implementar:

La implementación de una aplicación que permita realizar predicciones en el diagnóstico del cáncer de mama, puede contribuir significativamente a mejorar la detección temprana de esta enfermedad, lo que permitiría un tratamiento oportuno y una mayor tasa de supervivencia de los pacientes. Además, al utilizar técnicas de machine learning, se pueden obtener resultados más precisos y confiables que los métodos tradicionales de análisis.

Objetivos:

El objetivo principal de esta propuesta es desarrollar una aplicación que permita realizar predicciones en el diagnóstico del cáncer de mama utilizando técnicas de machine learning en Python. Para lograrlo, se plantean los siguientes objetivos específicos:

- Recopilar y limpiar datos clínicos de pacientes con cáncer de mama.
- Realizar una exploración de los datos y visualizarlos para identificar patrones y relaciones.
- Implementar modelos de regresión y clasificación mediante técnicas de machine learning, utilizando librerías de Python como Scikit-learn y Keras.
- Evaluar el rendimiento de los modelos utilizando técnicas de validación cruzada y métricas de evaluación como el área bajo la curva ROC.
- Desarrollar una interfaz gráfica de usuario que permita a los usuarios cargar datos y realizar predicciones mediante los modelos implementados.

Resultados esperados:

Se espera obtener una aplicación funcional que permita a los usuarios cargar datos clínicos de pacientes con cáncer de mama y realizar predicciones precisas en el diagnóstico de la enfermedad. Además, se espera que los modelos implementados sean robustos y confiables, con una tasa de acierto alta. Esta aplicación puede ser de gran utilidad para el personal médico que se encarga del diagnóstico y tratamiento del cáncer de mama, y para los pacientes que buscan obtener un diagnóstico temprano y preciso.