

Desafío - Introducción al Machine Learning

En este desafío tendrás la oportunidad de poner a prueba los conceptos aprendidos durante la sesión. Los ejercicios están diseñados para reforzar ciertos conceptos básicos que probablemente podrían ser consultados en una entrevista técnica.

Lee todo el documento antes de comenzar el desarrollo **individual**, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos. Asegúrate de seguir las instrucciones específicas en cada ejercicio y de completar los requerimientos adicionales, si los hubiera. ¡A disfrutar aprendiendo!

Tiempo asociado: 2 horas cronológicas

Descripción

El siguiente desafío consta de dos partes: una de 5 preguntas conceptuales, y otra en la que pondremos a prueba tus habilidades en la implementación de modelos de Machine Learning, específicamente en el área de regresión lineal. El objetivo es desarrollar un modelo que pueda predecir el precio de las casas en función de sus características.

Parte I

Responde las siguientes preguntas:

1. ¿Cuál es la diferencia entre el aprendizaje supervisado y el aprendizaje no supervisado?
2. En la industria de la salud, a menudo se utiliza el machine learning para predecir la probabilidad de enfermedades o condiciones médicas en función de diferentes variables de entrada, como síntomas, antecedentes médicos y factores de estilo de vida. Imagina que tienes un conjunto de datos que contiene información de pacientes y su diagnóstico (enfermo/no enfermo). ¿Qué tipo de aprendizaje utilizarías para construir un modelo predictivo?
3. ¿Por qué razón es necesario dividir el set de datos en 2 set de datos al momento de realizar algún modelamiento supervisado? ¿Cuáles son los 2 set de datos y cuál es la función de cada uno?
4. ¿Cuál es la principal diferencia en el modelamiento de la regresión lineal para un enfoque estadístico y un enfoque de aprendizaje de máquinas?

5. Escoge una industria de acuerdo a su área de interés, rubro de trabajo o área de estudios (por ejemplo, salud, telecomunicaciones, retail, medioambiente, etc). Menciona al menos 1 aplicación en esa industria de:
 - a. Aprendizaje Supervisado (Clasificación)
 - b. Aprendizaje Supervisado (Regresión)
 - c. Aprendizaje No Supervisado (Clustering)
 - d. Aprendizaje No Supervisado (Reducción de dimensionalidad)

Parte II

En este ejemplo, utilizaremos Python y la biblioteca Scikit-learn para construir y entrenar nuestro modelo de regresión lineal. A continuación, se presenta un paso a paso para llevar a cabo este desafío:

6. Carga los datos de la base **data_housing.csv** e importa las librerías necesarias para cargar los datos en un DataFrame, manipularlos, hacer cálculos, analizarlos y construir modelos de regresión lineal. Utiliza Pandas, NumPy y Scikit-learn.
7. Genera un informe de calidad de datos de la base. Analiza la existencia de datos atípicos, extraños o nulos. Además, analiza las posibles correlaciones entre las variables.
8. Genera un modelo de regresión lineal. Para ello:
 - a. Importa el estimador de regresión lineal de Scikit-learn.
 - b. Divide los datos (80% train y 20% test)
 - c. Crea una instancia del modelo de regresión lineal.
 - d. Ajusta el modelo utilizando los datos de entrenamiento.
9. Evalúa el modelo. Para ello:
 - a. Utiliza los datos de entrenamiento y calcula métricas de evaluación, como el error cuadrado medio (MSE) o el coeficiente de determinación (R^2). Grafica los coeficientes.
 - b. Utiliza los datos de prueba para realizar predicciones con el modelo entrenado, y evalúa el rendimiento del modelo comparando las predicciones con los valores reales.

Requerimientos

1. Comprende las diferencias entre los tipos de aprendizaje, y es capaz de identificarlos y ejemplificarlos. **(3 puntos)**
2. Aplica los pasos de la regresión lineal para generar un modelo. **(5 puntos)**
3. Evalúa un modelo de regresión lineal, calculando métricas y coeficientes. **(2 puntos)**

😊 ¡Mucho éxito!

Consideraciones y recomendaciones

- Debes entregar tu trabajo en un archivo de Jupyter Notebook, con las respuestas a las preguntas, y el código y las explicaciones respectivas para desarrollar tu trabajo.