Ciencia de Datos (online)

Introducción teórica a los contenidos

En este apartado se introducirán los distintos conceptos que van apareciendo en el módulo para contextualizarlos.

Introducción al Test de Hipótesis.

El test de hipótesis, también conocido como prueba t de Student, es un test estadístico empleado para analizar si dos muestras proceden de poblaciones con la misma media. Por ello, cuantifica la diferencia entre la media de las dos muestras y, teniendo en cuenta la varianza de éstas, estima lo probable que es obtener una diferencia igual o mayor que la observada si la hipótesis nula que las medias poblacionales son iguales fuera cierta. A la probabilidad estimada por el test se le conoce como p-value.

Un p-value mayor que un determinado límite, por ejemplo 5% o 1%, indica que la diferencia observada puede deberse al azar, por lo que no se rechaza la hipótesis nula. Por el contrario, cuando el p-value es menor que el límite seleccionado, se considera que existen evidencias suficientes para rechazar que las muestras proceden de poblaciones con la misma media.

Cuando se dispone de dos muestras, el hecho de que sus valores de media no sean exactamente iguales no implica que existan evidencias de una diferencia real. Dado que cada muestra tiene su propia variabilidad debida al muestreo aleatorio, aunque procedan de la misma población, las medias muestrales no tienen por qué ser iguales. Es ahí donde el test de hipótesis aporta va
Así, es un test estadístico para comparar la media entre dos muestras.

Existen múltiples adaptaciones del test de hipótesis en función de si los datos son independientes o dependientes, si la varianza es la misma en las dos muestras, o qué tipo de diferencias se desea detectar. En este módulo se trabajará el test de hipótesis con Python a partir del p-value.

Introducción al Machine Learning.

Machine Learning es una disciplina del campo de la inteligencia artificial que, mediante algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo). Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados.

El término se utilizó por primera vez en 1959. Sin embargo, ha ganado relevancia en los últimos años debido al aumento de la capacidad de computación y al boom de los datos. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del Big Data.

Los algoritmos de Machine Learning se dividen en tres categorías, siendo las dos primeras las más comunes:

- **Aprendizaje supervisado** : estos algoritmos tienen un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de spam que etiqueta un correo electrónico como spam o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).
- **Aprendizaje no supervisado** : estos algoritmos no tienen un conocimiento previo. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna forma. Por ejemplo, en el campo del marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas.
- **Aprendizaje por refuerzo** : su finalidad es que un algoritmo aprenda a partir de la propia experiencia. Es decir, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas. Actualmente, se está utilizando para posibilitar el reconocimiento facial, realizar diagnósticos médicos o clasificar secuencias de ADN.

Introducción al método Train-Test.

En torno al Machine Learning se crean modelos para predecir el resultado de ciertos eventos o valores, como podría ser prever cuál será la emisión de CO_2 de un automóvil cuando sabemos el peso y tamaño del motor.

Para medir si un modelo de previsión es suficientemente bueno, se puede utilizar un método llamado Train/Test. Este método Train/Test es pues un método para medir la precisión del modelo.

A este modelo se le llama Train/Test porque divide el conjunto de datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. Normalmente, se suele realizar un reparto de un 80% para entrenamiento y un 20% para pruebas.

El modelo se entrena con el conjunto de entrenamiento y se prueba usando el conjunto de prueba.

Entrenar el modelo significa crear el modelo y probar el modelo significa probar la precisión del modelo.

Así pues, estamos ante un método en el que, a partir de los propios datos conseguimos conocer la precisión de nuestro modelo.

Librería Scikit-Learn.

La librería Scikit-learn es uno de los open-source y bibliotecas de aprendizaje automático más populares de Python.

Esta biblioteca contiene muchas herramientas eficientes para aprendizaje automático y modelado estadístico, incluyendo clasificación, regresión, agrupación y reducción de dimensionalidad. Está basada en NumPy, SciPy y Matplotlib, por lo que es fácil reaprovechar el código que se utiliza en estas librerías.

Además, incorpora varias funciones para preprocesar los datos:

- **Normalización** : ajustar las variables numéricas para que tengan media 0 y varianza 1, o bien que estén en un rango como [0,1]. También permite normalizar vectores para que tengan norma.

- **Transformaciones no lineales** : basadas en cuantiles y exponentes, para transformar variables con distribuciones muy sesgadas, por ejemplo. Entre otras incluye la transformada de Yeo-Johnson y la de Box-Cox.
- **Discretización** : se trata de convertir una variable numérica en un conjunto de valores posibles según algún criterio. Un caso extremo es cuando una variable es convertida a sólo dos valores posibles, lo que se conoce como binarización.
- **Valores perdidos** : cuando algunos registros faltan datos de alguna variable (p. ej., un usuario/a no responde a alguna pregunta de una encuesta), es posible imputar un valor en función de algún criterio automatizable, por ejemplo, sustituir -lo por la media.
- **Creación de interacciones** entre variables mediante el uso de polinomios.

Así pues, nos encontramos frente a una de las bibliotecas de referencia en Python.

Unió Europea
Fons Europeu
de Desenvolupament Regional

Síguenos

Ha iniciado sesión como CARLOS RAUL ARDILA PERDOMO (Sale)

Política de Cookies - Política de Privacidad - Condiciones Generales de Uso