Interpretabilidad del Deep Learning

Introducción

Christian Oliva Moya

Grado en **Ingeniería Informática** UAM

Máster en **Inteligencia Computacional y Sistemas Interactivos** UAM

Doctorando en **Ingeniería Informática y Telecomunicaciones** UAM

XAI de redes neuronales profundas

Profesor Ayudante Dpto. Ingeniería Informática

Profesor MIAX desde la edición 8 - ML y Redes Neuronales

Profesor Máster Big Data UAM - ML y Clustering

christian.oliva@uam.es

https://www.linkedin.com/in/christian-oliva-moya-ingeniero/

Aprendizaje Automático (Machine Learning)

Redes Neuronales (Neural Networks)

Aprendizaje Profundo (Deep Learning)

Interpretabilidad-Explicabilidad (XAI)

Investigación

Formación

Objetivos del curso

- Comprender los conceptos básicos de la IA Explicable (XAI)
- Diferenciar Selección de atributos vs Explicabilidad vs Interpretabilidad
- Comprender las mecánicas de Explicabilidad Global
- Comprender las mecánicas de Explicabilidad Local
- Comprender las dificultades de la Interpretabilidad a nivel global
- Aprender a Interpretar una Red Neuronal Recurrente de forma global

Requisitos previos

- Conocimiento en **programación** (vamos a usar Python)
    - Librerías típicas como numpy, pandas, etc.
- Conocimiento en derivación automática (vamos a usar **Tensorflow y Keras**)
- Conocimiento en Machine Learning (usaremos varios modelos sencillos)
- Conocimiento en Deep Learning (vamos a usar MLPs, CNNs y RNNs)

Contenido del curso (I)

- Primera semana:
  - Introducción de conceptos
  - Explicabilidad genérica
    - SHAP
    - LIME
  - Coeficiente de Explicabilidad-Rendimiento

Contenido del curso (II)

- Segunda semana:
  - Explicabilidad específica para redes neuronales
  - Métodos basados en gradientes: *Gradient x Input*
  - Métodos basados en relevancias: Layerwise Relevance Propagation
  - ¿Es LRP equivalente a *Gradient x Input*?
  - Dificultades del Deep Learning

Contenido del curso (III)

- Tercera semana:
  - Interpretabilidad a nivel global y sus dificultades
  - Interpretabilidad global de las redes neuronales recurrentes
- Talleres:
  - Explicabilidad MNIST
  - Interpretabilidad P300