

PROGRAMA DE CURSO
Métodos para ciencia de datos
XS-0132

Información General:

Horas: 4 horas semanales teorico-prácticas

Créditos: 4 créditos.

Requisitos: XS4110, XS2230

Correquisitos: Ninguno

Descripción:

Este curso está dirigido a estudiantes de cuarto año del Bachillerato en Estadística. Procura que el/la estudiante aprenda los conceptos e implementación de redes neuronales como técnica de aprendizaje supervisado en el contexto de ciencia de datos, así como su aplicación a datos de imágenes y texto.

Objetivo General:

Conocer los conceptos básicos sobre técnicas de aprendizaje profundo, así como de herramientas para ejecutar dichas técnicas, con el fin de aplicarlas a datos de imágenes y texto.

Objetivos:

Al finalizar el curso el/la estudiante estará en la capacidad de:

1. Aplicar redes neuronales profundas a problemas de aprendizaje supervisado mediante el uso de TensorFlow y Keras con el fin de implementar la teoría de aprendizaje profundo.
2. Identificar los usos de distintos tipos de funciones de activación, optimizadores, funciones de pérdida e hiperparámetros de redes neuronales para modelar de manera adecuada y eficiente.
3. Aplicar técnicas de procesamiento de imágenes con el fin de facilitar el análisis de datos procedentes de imágenes.
4. Procesar datos de texto mediante métodos de procesamiento de lenguaje natural (NLP) con el fin de estructurar los datos para el modelado.
5. Identificar las arquitecturas de redes neuronales apropiadas para su aplicación en datos provenientes de texto e imágenes para su mejor funcionamiento.

Contenidos:**1. Principios de redes neuronales**

- a. Áreas de ciencia de datos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado.
- b. Conceptos básicos de redes neuronales: neuronas, capas, funciones de activación, funciones de pérdida.
- c. Arquitectura de redes neuronales: perceptrón y redes neuronales profundas (DNN).
- d. Métodos de optimización: Descenso de gradiente, propagación hacia atrás y descenso de gradiente estocástico, momentum y decline en la tasa de aprendizaje
- e. Problemas de ajuste en modelos: Sobreajuste y balance entre varianza y sesgo
- f. Validación cruzada y selección de hiperparámetros.
- g. Métodos de regularización y fin temprano.
- h. Métodos de optimización con tasas de aprendizaje adaptativas (adaptive learning rates): AdaGrad, RMSProp, Adam.
- i. Aplicaciones para regresión y clasificación.

2. Redes neuronales aplicadas a imágenes

- a. Procesamiento de imágenes: tipos de datos, cargar imagenes, convertir imagenes en vectores, reducción de dimensionalidad, cambio de tamaño, rotar o cortar imágenes, normalizar pixeles, standarizar pixels, manipulación de canales de una imagen.
- b. Estructura de redes neuronales convolucionales.
- c. Conceptos para redes neuronales convolucionales (CNN): convolución, filtros, capas de convolución, capas de pooling, padding, stride.
- d. Introducción a arquitecturas de redes neuronales convolucionales: AlexNet, VGG, Inception y Residual Net o similares.
- e. Introducción a la utilización de modelos de redes neuronales convolucionales pre-entrenados.

- f. Introducción a los conceptos de autoencoders, autoencoders de variación, GANs y otras aplicaciones.

3. Redes neuronales aplicadas a texto

- a) Procesamiento, limpieza y organización de texto: cargar datos de texto, tokenización, stemming, lematización, palabras vacías (stop words), expresiones regulares básicas, separación de oraciones, entre otros.
- b) Modelo bolsa de palabras (bag-of-words) y TF-IDF.
- c) Introducción al análisis de sentimientos.
- d) Word embeddings: Word2Vec y/o GloVe.
- e) Introducción de las diferentes aplicaciones de procesamiento de lenguaje natural (NLP): resumen de texto, traducción, generación de texto, entre otros.
- f) Estructura y aplicación de redes neuronales recurrentes: RNN, LSTM, GRU.
- g) Estructura y aplicación de redes neuronales recurrentes bidireccionales.
- h) Estructura y aplicación de introducción a transformadores (BERT).

Metodología:

- El curso combina enfoques teórico-prácticos para cubrir los conceptos así como sus aplicaciones con el uso de lenguaje de programación. Dado lo anterior, exige el uso frecuente de la computadora.
- **Presentaciones teóricas:** se impartirán clases magistrales donde se explicará los conceptos de redes neuronales.
- **Componente práctico:** se realizarán laboratorios estructurados con ejercicios sobre los contenidos desarrollados en las clases teóricas. En estas sesiones de laboratorio se utilizará el lenguaje de programación Python en la plataforma Google Colab o similar.
- **Proyectos prácticos:** con el objetivo de poner en práctica los conocimientos, los/las estudiantes se enfrentarán a problemas reales en donde apliquen los conocimientos obtenidos durante el curso y presentar sus hallazgos en forma de artículo y presentación oral.

Se planea utilizar como referencia, para las clases teóricas y los laboratorios, los siguientes capítulos o sus respectivos equivalentes en las ediciones más actualizadas que se encuentren disponibles a la hora de impartir el curso:

1. Principios de redes neuronales:

- a. Brownlee (2016) capítulos del 6 al 17
- b. Goodfellow et al. (2016) capítulos del 5 al 8.

2. Redes neuronales aplicadas a imágenes:

- a. Brownlee (2016) capítulos del 18 al 22
- b. Brownlee (2019) capítulos del 4 al 22
- c. Goodfellow et al. (2016) capítulo 9.

3. Redes neuronales aplicadas a texto:

- a. Brownlee (2016) capítulos del 23 al 28
- b. Brownlee (2017) capítulos del 5 al 13 y del 17 al 19
- c. Goodfellow et al. (2016) capítulo 10.
- d. Jurafsky (2021) capítulos: 2 al 4 y 7

Evaluación:

- **Exámenes:** Se realizarán dos exámenes parciales que tendrán un valor total de 50%.
- **Proyecto:** Durante el semestre los estudiantes deberán trabajar en dos proyectos de análisis de datos el cual utilice una de las técnicas de aprendizaje profundo vistas en clase. Cada proyecto se enfoca en aplicación a imágenes y de análisis de texto respectivamente, las cuales pueden utilizar una base de datos existente.

Ambos proyectos concluyen con una presentación en clase y un informe escrito en formato de artículo científico en el cual detallen los procedimientos de análisis utilizados.

TIPO DE EVALUACIÓN	PORCENTAJE
Examen 1	25%
Examen 2	25%
Proyecto 1	25%
Proyecto 2	25%
Total	100%

Cronograma:

SEMANA / FECHA	CONTENIDO	ACTIVIDAD
1	Principios de redes neuronales	
2		
3		
4		
5		
6		
7	Redes neuronales aplicadas a imágenes	
8		Examen I
9		
10		
11		Presentación de proyecto I (escrito y oral)
12	Redes neuronales aplicadas a texto	
13		
14		
15		
16		Examen II
17	-	Presentación de proyecto II (escrito y oral)

Bibliografía:

- Brownlee, J. (2016). Deep Learning With Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras. Machine Learning Mastery.
- Brownlee, J. (2017). Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems. Machine Learning Mastery.
- Brownlee, J. (2019). Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. Machine Learning Mastery.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. The MIT Press
- Jurafsky, D., Martin, J. H. (2021). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Pearson Prentice Hall
- Szeliski, R. (2011). Computer vision algorithms and applications Springer