



CAPIBArA

Conceptos de Aprendizaje Profundo para la Investigación en Biología y Áreas Afines

Introducción

Varias cajas de Pandora fueron abiertas en las ciencias biológicas mediante la aplicación de nuevas tecnologías: desde el descubrimiento de los microorganismos a la dilucidación de la maquinaria celular y molecular que hace posible la vida, un sinfín de ejemplos son enumerables. El aprendizaje profundo, una disciplina que se encuentra en ebullición en los últimos años, parece ser una de esas cajas cargadas de herramientas que serán moneda corriente para los/as biólogos/as del futuro.

Este curso se plantea introducir y ejercitar el uso de herramientas de aprendizaje profundo para el análisis de problemas biológicos. Para ello, se introducirán conceptos teóricos básicos de la disciplina, los cuales serán aplicados a problemas biológicos concretos obtenidos de la literatura reciente.

El curso no pretende ser una introducción formal y exhaustiva sobre el tópico (aunque no se evitará la rigurosidad cuando sea necesaria), teniendo por objetivo que el/la estudiante obtenga una noción básica sobre los conceptos claves del área y la aplicación a diversas áreas de la biología. En lo referente al desarrollo práctico, se ejercitará la utilización de herramientas de la librería PyTorch (Python) para aplicar los modelos presentados a datos obtenidos de la literatura.

El curso contará con un equipo interdisciplinar, reuniendo a docentes de las áreas matemática e ingeniería junto a docentes del área de biología computacional,

Objetivo

- Comprender principios fundamentales de aprendizaje automático/profundo
- Aprender a seleccionar y aplicar modelos de aprendizaje profundo adecuados para diferentes tipos de datos y contextos.

- Aprender a utilizar la librería PyTorch (lenguaje de programación Python) para el desarrollo de modelos de aprendizaje profundo
- Ser capaces de interpretar los resultados de los modelos de aprendizaje profundo y considerar las limitaciones e implicancias éticas de sus aplicaciones.
- Explorar casos de estudio reales de aplicación de aprendizaje profundo en ciencias biológicas.

Temario General

· Módulo 1. Fundamentos para el Aprendizaje Profundo

- *Repaso de Aprendizaje Automático.* Flujo de trabajo de AA aplicado a biología: desde datos crudos (ómicos, imágenes/microscopía, audio de campo, etc.) hasta métricas. Función de ajuste y representación de hipótesis. Aprendizaje supervisado, no supervisado y por refuerzo. Ajuste de modelos: overfitting, underfitting, bias–variance tradeoff. Validación cruzada y métricas comunes.
- *Redes Neuronales Artificiales.* Perceptrón monocapa y multicapa. Funciones de activación. Teorema de aproximación universal. Conceptualización de descenso por gradiente y retropropagación. Inicialización, normalización y regularización.

· Módulo 2. Arquitecturas de Deep Learning y Aplicaciones Biológicas

- *Redes Convolucionales (CNN)* Convolución, pooling, padding, stride. Datos con estructura espacial (imágenes, mapas). Patrones locales y jerárquicos en datos 1D y 2D no visuales. Arquitecturas clásicas: LeNet, AlexNet, ResNet. Transfer learning y fine-tuning en contextos biológicos. Casos biológicos: Detección de motivos en secuencias de ADN, predicción de estructura secundaria a partir de la secuencia de aminoácidos, reconocimiento de patrones auditivos.
- *Autoencoders.* Concepto y arquitectura básica. Autoencoders para reducción de dimensionalidad (p. ej., expresión génica, metagenómica). Denoising autoencoders, variational autoencoders (VAE). Casos biológicos.
- *Transformers y Embeddings.* Redes que procesan secuencias y sus limitaciones. Word embeddings no contextuales (word2vec). Mecanismo de atención. Arquitectura transformer (encoder/decoder). Embeddings dependientes del contexto. Casos biológicos: clasificación de secuencias genómicas (DNABERT), anotación de proteínas (ESM models), análisis de cantos de aves con audio embeddings. Análisis del habla.
- *Redes Neuronales en Grafos (GNNs).* Representación de datos no euclidianos. Mensaje y agregación. Casos biológicos: redes de interacción ecológica, redes de co-expresión génica, redes de asociación de palabras..
- *Interpretabilidad y reproducibilidad.* Grad-CAM en CNNs. Visualización de atención en transformers. Autoencoders sparse para la interpretabilidad. Alucinaciones. Aspectos éticos del uso de modelos de DL en CB. Bonus: “cómo leer un paper de deep learning en biología”.

· Módulo 3. Casos de Estudio y Proyecto Final

- Casos de estudio. i. Neurociencias y Lenguaje ii. Genómica y proteómica iii. Aplicaciones en Ecología.

- Diseño de un flujo de trabajo completo: Planteo del problema. Preparación y exploración de datos. Entrenamiento y ajuste de modelo. Evaluación y visualización de resultados.

Público esperado

El curso está dirigido a estudiantes de grado avanzado, posgrado o egresados de las áreas biológicas, (e.g. Ciencias Biológicas, Bioquímica, Medicina, Veterinaria, Agronomía, etc) con experiencia básica de programación y conocimientos matemáticos (álgebra lineal, probabilidad y estadística). Con respecto a programación, es necesario haber cursado Aprendizaje Automático Básico para Científicos, Introducción a la Ciencia de Datos (cursos de CICADA), introducción a la línea de comandos (Instituto de Higiene), o contar con formación equivalente. Los cursos de matemáticas, probabilidad y estadística usuales en las carreras de grado que incluyen estos temas son suficientes.

El curso puede ser también útil a estudiantes con trayectorias fuera de las biociencias (e.g ingeniería, física o matemática). Son deseables, aunque no excluyentes, conocimientos básicos de biología (equivalentes a Introducción a la biología I o Biología general, e.g. conocimientos básicos de genética, biología molecular, ecología y evolución).

Aprobación

Se realizarán entregas de tareas asignadas durante el curso y luego se presentará un proyecto final.

Docentes

Héctor Romero, Juan Valle Lisboa. Colaboran: Mauricio Langleib, Felipe Maresca, Francisco de Izaguirre.

Invitados

Paola Bermolen, Álvaro Cabana, Federico Lecumberry, Bernardo Marengo.

Contacto: eleitor@fcien.edu.uy, cicada@ei.udelar.edu.uy.

Fechas, lugar, duración y carga horaria

El curso se dictará entre el 17 de agosto y el 30 de noviembre, y tendrá una carga horaria de 5 horas por semana (75 hs en total, **10 créditos**). Será dictado en forma presencial en Facultad de Ciencias.

Bibliografía

Deep Learning for Biology: Harness AI to Solve Real-World Biology Problems. Charles Ravarani, Natasha Latysheva. O'Reilly (2025)

Deep Learning. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Massachusetts Institute of Technology (2016)

Deep Learning with PyTorch. Eli Stevens, Luca Antiga and Thomas Viehmann. Manning Publications Co (2023)

Deep Learning for Coders with fastai and PyTorch. Jeremy Howard and Sylvain Gugge. O'Reilly (2020)

The Elements of Statistical Learning. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer (2009)

Mathematics for Machine Learning. Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong Cambridge University Press (2020)

Understanding Deep Learning. Simon J. D. Prince. MIT Press (2023)