Universidad Nacional
Autónoma de México
UNAM

Temario del Curso: Aprendizaje Automático en Cosmología

Sebastien Fromenteau (IA-UNAM)

La cosmología busca extraer información estadística a partir de datos astrofísicos con el objetivo de restringir el modelo que describe la evolución del Universo, conocido como el modelo estándar de la cosmología o modelo del Big Bang. Los métodos tradicionales se basan en la compresión de la información observacional, permitiendo así contrastar las predicciones teóricas con estimadores derivados de los datos.

A medida que la calidad y cantidad de las observaciones aumenta, se vuelve necesario mejorar la estimación de las estadísticas comprimidas, que suelen estar basadas en funciones de correlación de N puntos. Sin embargo, los efectos astrofísicos no lineales empiezan a desempeñar un papel relevante en los modelos, lo que exige descripciones más complejas y computacionalmente costosas.

En este contexto, se hace indispensable el uso de emuladores capaces de aproximar modelos complejos, reduciendo así sus tiempos de evaluación. También se requieren métodos automatizados para analizar el enorme volumen de datos observacionales (por ejemplo, LSST generará alrededor de 2 TB de datos por noche), así como técnicas avanzadas para clasificar objetos astrofísicos y estimar de manera precisa sus corrimientos al rojo.

Todos estos desafíos han motivado, en los últimos años, una creciente adopción de técnicas de aprendizaje automático en cosmología. Más aún, una parte importante de los esfuerzos actuales se centra en la Simulación-Based Inference (SBI), el enfoque más popular dentro de la familia de métodos Likelihood-Free Inference, que permite incorporar efectos astrofísicos difíciles de modelar explícitamente en las funciones de correlación.

Este curso tiene como objetivo introducir primero los fundamentos estadísticos necesarios, para luego construir progresivamente las herramientas que permiten comprender y aplicar técnicas modernas de aprendizaje automático, hasta llegar a la exploración de conceptos avanzados de inferencia sin modelización explícita de la función de verosimilitud.

Bloque 1 – Introducción general al aprendizaje automático en cosmología

- Motivaciones: ¿Por qué ML en cosmología moderna?
- Tipos de problemas cosmológicos donde se aplica:
  - Emulación rápida de modelos cosmológicos (e.g. $P(k)$, C_ℓ, halo mass functions)
  - Clasificación de morfologías de galaxias
  - Generación de datos simulados
  - Inferencia Bayesiana sin verosimilitud (likelihood-free inference)
- Panorama de métodos: emuladores, autoencoders, normalizing flows, redes convolucionales, transformers.
- Revisión de bibliografía clave y bases de datos disponibles (e.g. CAMELS, SIMBIG, LFI, Cosmoflow)

Bloque 2 – Fundamentos de estadística Bayesiana

- Reglas de Bayes y marco probabilístico
- Verosimilitud, prior, posterior y evidencia bayesiana
- Métodos numéricos: MCMC, Nested Sampling
- Aplicación concreta: estimación de parámetros cosmológicos desde un $P(k)$ sintético

Bloque 3 – Construcción manual de una red neuronal básica

- Construcción "from scratch" sin librerías de ML (sólo NumPy)
- Pesos, sesgos, funciones de activación (sigmoide, ReLU, tanh)
- Función de costo (MSE, cross-entropy), gradiente descendente y tasa de aprendizaje
- Retropropagación del error (Graident descent)
- Entrenamiento en mini-batches vs batches completos
- Evaluación sobre datos sintéticos 1D

Bloque 4 – Emulador con red neuronal densa (MLP)

- Introducción a redes densas profundas (MLP: MultiLayer Perceptron)
- Emulación de $P(k)$ a partir de parámetros cosmológicos
- Funciones de activación especiales (Softplus, GELU)
- Regularización: dropout, early stopping, weight decay
- Métricas de desempeño: RMSE, χ^2 efectivo

Bloque 5 – Autoencoder (AE) para reducción de dimensionalidad

- Autoencoder denso sobre datos 1D y 2D
- Estudio del espacio latente (PCA, t-SNE, interpretabilidad física)
- Clasificador usando el encoder; generador usando el decoder
- **Aplicaciones adicionales: reconstrucción de datos faltantes y filtrado de ruido**

Bloque 6 – Redes convolucionales (CNN) en 1D y 2D

- Motivación: extracción de características invariantes por translación
- Filtros, operaciones de convolución
- MaxPooling, arquitectura CNN básica
- Aplicación a mapas 2D (CMB, galaxias), predicción de parámetros cosmológicos

Bloque 7 – Autoencoder convolucional (CAE)

- Combinación de CNN y autoencoder
- Compresión eficiente de mapas cosmológicos 2D
- Comparación con AE denso: reconstrucción, latente, robustez

Bloque 8 – Métodos de inferencia "likelihood-free"

- Motivación: verosimilitud inexacta o costosa
- Introducción a Normalizing Flows y Diffusion Models
- Entrenamiento para aproximar distribuciones a posteriori
- Aplicación a problema cosmológico simple (e.g. inferir Ω_m, σ_8)
- Comparación con MCMC tradicionales usand $P(k)$

Extras (opcional)

- Visualización de modelos (TensorBoard)
- Procesos Gausianos
- PyTorch Lightning
- Breve sobre Transformers, GNNs y simuladores diferenciables
- Proyecto final aplicado