



ME-1000

Programación Paralela en R y Python

Duración

Cuatro semanas

Descripción del Curso

La programación paralela es esencial en la ciencia de datos y el aprendizaje automático para el manejo eficiente de grandes volúmenes de datos y la ejecución de modelos computacionalmente intensivos. Este curso explora el uso de herramientas en R y Python para implementar programación paralela y optimizar el rendimiento de algoritmos. Se abordarán paquetes específicos como Snow en R y Joblib en Python, y se explorarán ejemplos prácticos de paralelización de tareas usando scikit-learn y Dask.

Objetivos

1. Introducir los conceptos y técnicas básicas de programación paralela en R y Python.
2. Capacitar a los estudiantes en el uso de herramientas como Snow en R y Joblib en Python para mejorar el rendimiento de tareas computacionales.
3. Explorar la paralelización en el entrenamiento y evaluación de modelos de machine learning utilizando scikit-learn.
4. Enseñar a utilizar documentación técnica para profundizar en el uso de paquetes de programación paralela.

Metodología

El curso combina teoría y práctica para asegurar una comprensión profunda de los conceptos y habilidades prácticas en programación paralela. Se emplearán las siguientes herramientas y recursos:

- 1. Videoconferencias semanales:**
Se ofrecerán sesiones en vivo en Zoom que serán grabadas para que los estudiantes puedan revisarlas cuando deseen.
- 2. Trabajos prácticos semanales:**
Ejercicios y casos de estudio aplicados para implementar programación paralela.
- 3. Foros de discusión:**
Un espacio en Moodle donde los estudiantes podrán plantear dudas y discutir con el tutor y compañeros.
- 4. Aula virtual en Moodle:**
Aula virtual en Moodle: Acceso a materiales, enlaces a documentación, ejemplos de código y otros recursos complementarios.

Luego de este curso el estudiante será capaz de:

- 5.** Implementar programación paralela en R usando el paquete Snow y en Python usando Joblib.
- 6.** Aprovechar la paralelización en modelos de aprendizaje automático en scikit-learn.
- 7.** Comprender y navegar la documentación técnica para aplicar programación paralela en proyectos de ciencia de datos.
- 8.** Optimizar procesos computacionales para reducir tiempos de ejecución en tareas de ciencia de datos.

Contenido

- **Semana 1:**
Fundamentos de Programación Paralela en Ciencia de Datos
 - a. Introducción a la programación paralela y su importancia en ciencia de datos.
 - b. Conceptos básicos de concurrencia y paralelización.
 - c. Herramientas y paquetes principales para programación paralela en R y Python.
- **Semana 2:**
Programación Paralela en R con el Paquete Snow
 - a. Instalación y configuración de Snow en R.
 - b. Creación de clústeres y ejecución de tareas en paralelo.

- c. Ejemplos prácticos de paralelización en análisis de datos y simulaciones.
- d. Documentación: Snow Package Documentation.
- **Semana 3:**
Programación Paralela en Python con Joblib
 - a. Introducción a Joblib en Python.
 - b. Uso de joblib.Parallel para la ejecución de funciones en paralelo.
 - c. Casos prácticos de paralelización de tareas computacionales.
 - d. Documentación: Joblib.Parallel.
- **Semana 4:**
Paralelización en Modelos de Machine Learning con scikit-learn y Dask
 - a. Implementación de programación paralela en scikit-learn usando Joblib.
 - b. Uso de Dask para la escalabilidad en machine learning.
 - c. Ejemplo práctico: Paralelización de la Validación Cruzada en R y Python.
 - d. Ejemplo práctico: Paralelización de calibración y selección de modelos en R y Python.
 - e. Documentación: Scikit-learn con Joblib en Dask.

Evaluación

La evaluación consistirá en 4 tareas prácticas con un valor igual. Cada tarea estará enfocada en implementar y optimizar programación paralela en diferentes contextos de análisis de datos y aprendizaje automático.

Bibliografía

Vahid, F. (2020). Parallel Programming in Science and Engineering. Springer. Libro.

Joblib Documentation. (2023). Joblib.Parallel. <https://joblib.readthedocs.io/en/latest/generated/joblib.Parallel.html>

Dask Developers. (2023). Using Joblib with Dask. <https://ml.dask.org/joblib.html>

Scikit-Learn Developers. (2023). Parallelism in Scikit-Learn 1.5. <https://scikit-learn.org/1.5/computing/parallelism.html>

Scikit-Learn Developers. (2023). Parallelism in Scikit-Learn 1.0. <https://scikit-learn.org/1.0/computing/parallelism.html>

Downey, A. (2016). Think Python: How to Think Like a Computer Scientist. Green Tea Press.

Eddelbuettel, D. (2013). Seamless R and C++ Integration with Rcpp. Springer.

Wilkinson, B. & Allen, M. (2005). Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers. Pearson.



Programa Iberoamericano de
Formación en Minería de Datos