uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

Introducción a técnicas de Big Data :

MapReduce y Spark

MAPREDUCE

- Primero se explica qué se entiende por aprendizaje automático a gran escala y se muestra que hay varias formas en que los algoritmos de aprendizaje automático se pueden paralelizar: task-parallelism, data-parallelism, pipeline-parallelism.
- Se comentan algunos ejemplos de task-parallelism (principalmente, embarrising parallelism o paralelismo obvio).
- Pero el principal tipo de paralelismo que se usa hoy en día es el paralelismo de datos.
- Uno de los paradigmas principales para el paralelismo de datos es MapReduce.
- MapReduce es particularmente útil cuando se dispone de cientos o miles de ordenadores conectados a través de una red, y los datos se pueden dividir entre los diferentes ordenadores. La idea principal de MapReduce no es mover los datos por la red, sino mover procesos a donde se encuentran los datos.
- El modelo MapReduce se explica explicando sus procesos principales: map, sort and shuffle y reduce. Se explica un ejemplo para contar palabras.
- Se muestra como se puede programar KNN con MapReduce.
- Finalmente, se explica que hoy en día el paralelismo de datos se está moviendo hacia un nuevo modelo de programación llamado Spark, aunque muchas de las ideas de MapReduce siguen siendo válidas.

SPARK

- Se explican las limitaciones del modelo de programación MapReduce, y como Spark puede resolverlos.
- Se presentan conceptos básicos, especialmente el RDD (resilient distributed dataset) y los conceptos de transformación y de acción.
- Las transformaciones transforman un RDD en otro RDD, pero su ejecución es lenta.

- Solo cuando se ejecuta una acción, todas se aplican y ejecutan todas las transformaciones.
- Se presentan dos de las principales bibliotecas en R de Spark : sparkR y sparklyr.

Material asociado

Además de las diapositivas de la clase y algunos ejercicios (consulte la guía del curso)