Tecnologías de datos masivos

Doble Grado en Ingeniería en Tecnologías de Telecomunicación y Business Analytics

Machine Learning

Machine Learning - Historia

- El Machine learning nace a principios de los años 50 con los test de Turing
- En sus inicios, el Machine Learning seguía el principio de aprendizaje orientado al conocimiento
- A finales de la década de los 70 se inicia el primer “invierno” del Machine Learning debido a problemas de financiación
- La década de los 80 viene marcada por el primer boom del Machine Learning
- Se extienden los sistemas basados en reglas y estos mismo son adoptados en el mundo empresarial para la toma de decisiones

Machine Learning - Historia

- Tras el boom de la década de los 80 viene el segundo invierno del Machine Learning
- No se le dedica tanta investigación pero hay avances como el ordenador *Deep Blue* capaz de vencer al *gran maestro* del ajedrez Gary Kasparov
- Se empieza a cambiar el enfoque del Machine Learning del knowledge-based a un enfoque basado en el dato
- Hasta los primeros años de los 2000 no hay grandes avances
- Con la popularización de las tecnologías **Big Data** y la capacidad de procesar cantidades masivas de datos se impulsa el Machine Learning orientado al dato

Machine Learning - Conocimiento vs datos

- En un principio el Machine Learning se basa en el aprendizaje por conocimiento
    - Basado en el conocimiento:
        - Enfatiza la lógica como una herramienta para representar creencias sostenidas por un agente.
        - Ej: Casa es donde la gente está a la hora del desayuno, la comida o la cena y trabajo es aquellos sitios donde te encuentras en horario laboral
    - Basados en los datos:
        - La principal fuente de conocimiento está formada por datos observados y, en general, no utiliza la lógica como herramienta de modelado.
        - Ej: Estudio de los patrones más comunes de estancia de los individuos

Machine Learning - Pasos

- Cuando se usan aproximaciones basadas en el dato más que en la lógica el estudio y procesamiento de cantidades masivas de datos se hace fundamental
- Tener cantidades masivas de datos también hace tener datos no útiles
- Para obtener las principales características de un problema no se requieren toda la muestra sino una parte controlada
- Con los datos de entrenamiento obtenidos se crea un modelo matemático para cada casuística
- Tras aplicar y obtener un modelo, este se tiene que usar para predecir el resto de datos de nuestro dataset
- Todas estas fases puede beneficiarse de herramientas de cálculo masivo

Machine Learning - MlLib

- Spark, además de ayudar en las tareas de limpieza y selección de datos, tambien tiene su propia librería de generación y uso de algoritmos
- Dicha librería se conoce como MlLib cuando usamos Spark Core y ML en caso de usar SparkSQL
- Para la resolución de estos algoritmos Spark se basa en su Core, los RDD
- Los algoritmos de Machine Learning desarrollados por Spark requieren que el problema se pueda tratar de manera distribuida
- No todos los algoritmos de Machine Learning pueden ser resueltos con Spark

Machine Learning - MlLib

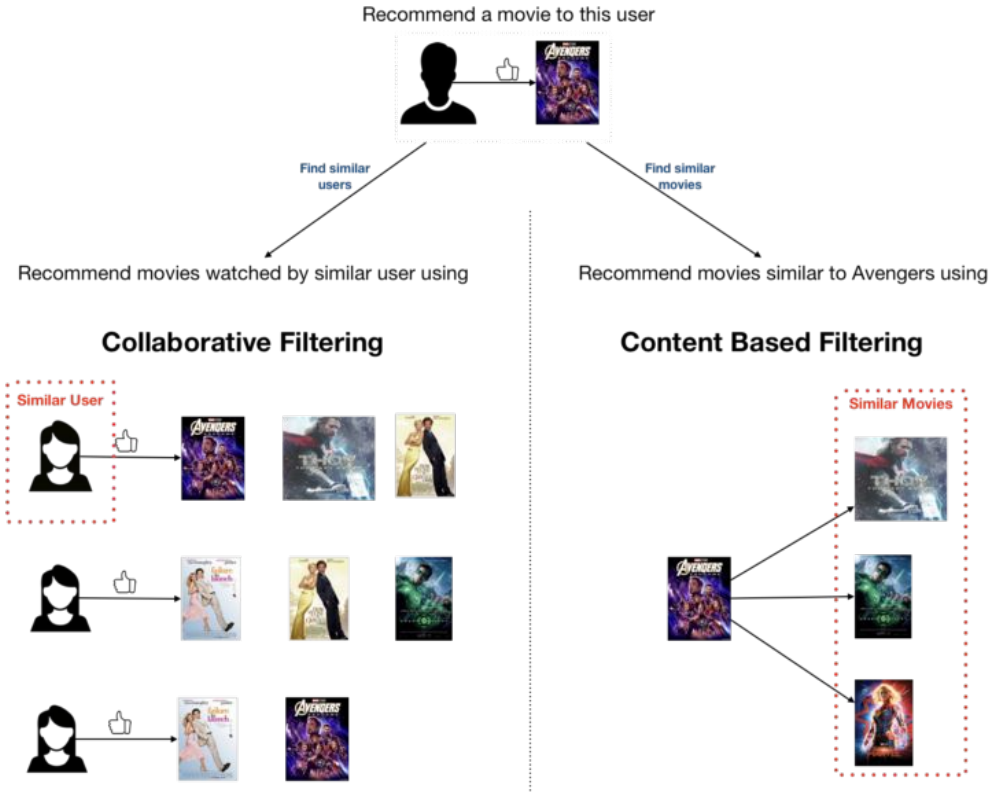
- Ml y MlLib requieren del uso de **Models**
- Hay modelos de Ml/MlLib para las siguientes casos de uso:
  - Clasificación y regresión
  - Filtro Colaborativo
  - Clustering
  - Minería de patrones frecuentes
- Cada uno de ellos conlleva la creación y uso de distintos tipos de variables

Ejemplos MILib

Machine Learning - MlLib- ALS

- Uno de los problemas que tienen las empresas es que productos recomendar a cada uno de sus clientes
- El sistema más sencillo sería recomendar aquellos productos mejor valorados
- Con esta aproximación perderíamos la tipología del usuario, no son lo mismo las notas de los turistas que de los locales o de la gente mayor que la gente joven
- Otra opción es recomendar en base a los productos que has consumido, pero asi asumimos que al usuario “solo” le gusta una cosa, por ejemplo sólo recomendar productos de deporte a alguien que le guste ver el futbol
- ¿Por qué no usar todas las notas de nuestros clientes para determinar que productos recomendar?

Machine Learning - MlLib- ALS



Machine Learning - MlLib- ALS

Machine Learning - MlLib- ALS

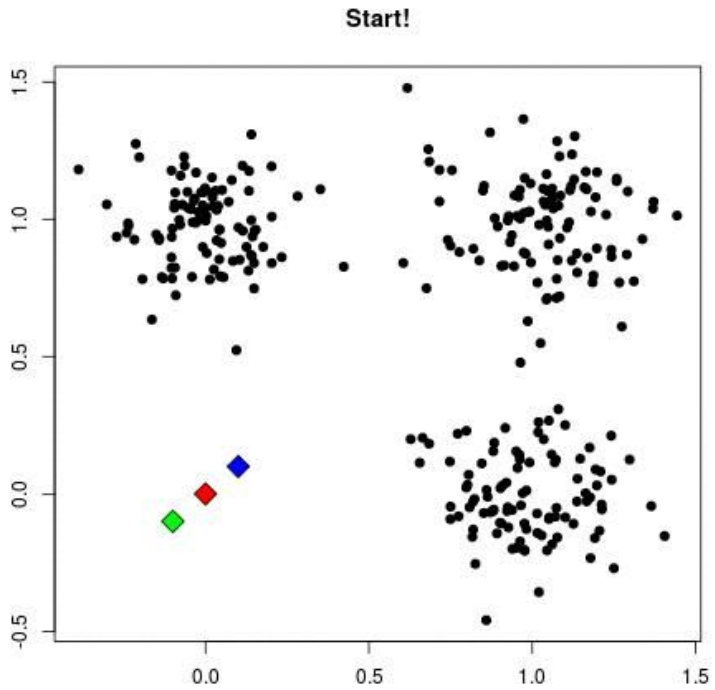
- Spark Ml tiene ya predefinido un algoritmo ALS mediante el cual con configurar unos parámetros podríamos hacer recomendaciones para millones de usuarios

- Usando este algoritmo en conjunción con SparkStreaming podríamos dar recomendaciones en “tiempo real” a usuarios basandonos en el uso de los productos del cliente

Machine Learning - MlLib- Kmeans

- En el pasado se usaba el conocimiento experto para agrupar el conocimiento
- Para analizar un nuevo patrón o agrupación se requería:
    - que un conjunto de expertos evaluara los datos
    - Les diera un significado
    - Determinen las normas para que un evento pertenezca a ese nuevo grupo
- Este proceso era muy lento y estaba sesgado por el conocimiento previo de los expertos
- Con la nuevas tecnologías se nos permite usar los propios datos para que nos descubran nuevos patrones en nuestros conjuntos de datos

Machine Learning - MlLib- KMeans



Machine Learning - MlLib- KMeans

- En Spark tenemos varios algoritmos de clusterización ya predefinidos
- Para ilustrar cómo se agrupan datos en Spark vamos a usar el algoritmo KMeans para detectar casa y trabajo
- Una vez detectado esto podríamos añadir un nuevo insights a nuestra base de conocimientos que enriquecerá el resto de casos de usos que queramos resolver con nuestros datos
- Aunque la tecnología nos determine el número de clusters de nuestros datos se sigue teniendo que hacer un estudio de los datos para darles un significado

MLlib