Introducción

Qué es Machine Learning

Para comprender a qué se llama comúnmente Machine Learning es conveniente revisar otras expresiones y términos históricamente utilizados en el ámbito de análisis de datos.

Estadística

Es un conjunto de técnicas descriptivas, de tratamiento de datos y modelización predictiva. Tradicionalmente comprende tanto técnicas de estimación y contrastes de hipótesis (inferencia) como visualización de datos, modelos predictivos (regresión, modelos para series temporales, etc.), y técnicas de recogida de datos como son el muestreo y diseño de experimentos.

A medida que fueron apareciendo nuevas técnicas necesarias para el tratamiento de la información se fueron incorporando al corpus de la estadística. Así, técnicas multivariables como análisis factorial , análisis cluster y análisis de correspondencias fueron engrosando el campo de la estadística.

Data Mining

En un momento dado y con la utilización masiva de los ordenadores PC, fue necesario establecer esquemas de trabajo para orientarse en un archivo de datos estadísticos con un gran número de variables y observaciones, a menudo sin un objetivo concreto o varios objetivos simultáneos. Así aparecieron esquemas como SEMMA y CRISP, y se añadieron a las técnicas estadísticas habituales herramientas de tipo informático como estructuras de bases de datos, herramientas de búsqueda, reglas, etc. Esta unión de esquema de trabajo, estadística e informática configura el data mining o minería de datos.

Machine Learning

Machine Learning o aprendizaje automático es un conjunto de técnicas que comprendería métodos tradicionales predictivos, como la regresión, como modernos como redes neuronales, gradient boosting o support vector machines. Según las fuentes, Machine Learning incluiría también aprendizaje no supervisado como técnicas cluster, o conceptos del campo de la informática como creación automática de reglas, reinforcement learning, deep learning, etc.

Big Data

La existencia de grandes volúmenes de datos, y formatos más complejos como texto, imágenes, sonido, etc. y la posibilidad de trabajar con datos online, lleva a añadir herramientas específicas a los conjuntos de técnicas del Data Mining y Machine Learning. En las herramientas del Big Data se incluyen Hadoop, Spark, text mining, web scrapping, técnicas de tratamiento de imágenes, sonido y video, etc.

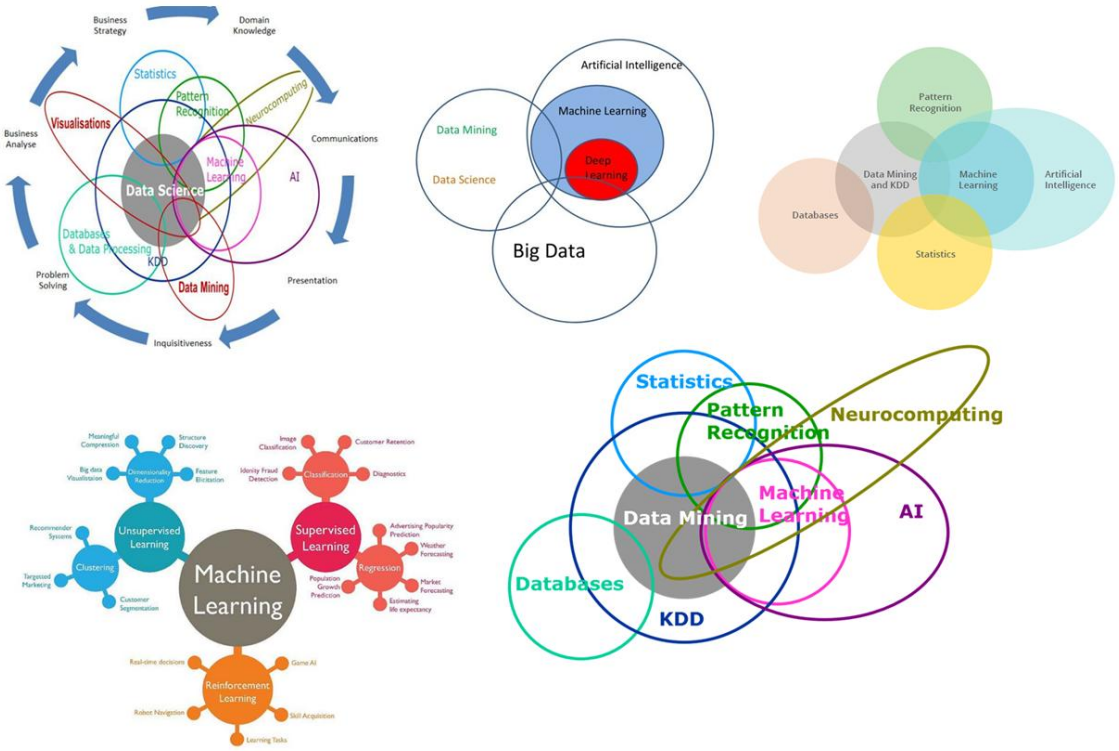
Data Science

La necesidad de contar con expertos en todas las áreas mencionadas en el Big Data aboca a la definición de una nueva disciplina llamada Data Science, cuyos actores son los Data Scientists, capaces de orientarse en las herramientas y conceptos anteriores y adquirir habilidades específicas en un conjunto amplio de lenguajes de programación y plataformas (R, Python, Java, Scala, Spark, etc.)

La ausencia de consenso e incoherencia en las definiciones

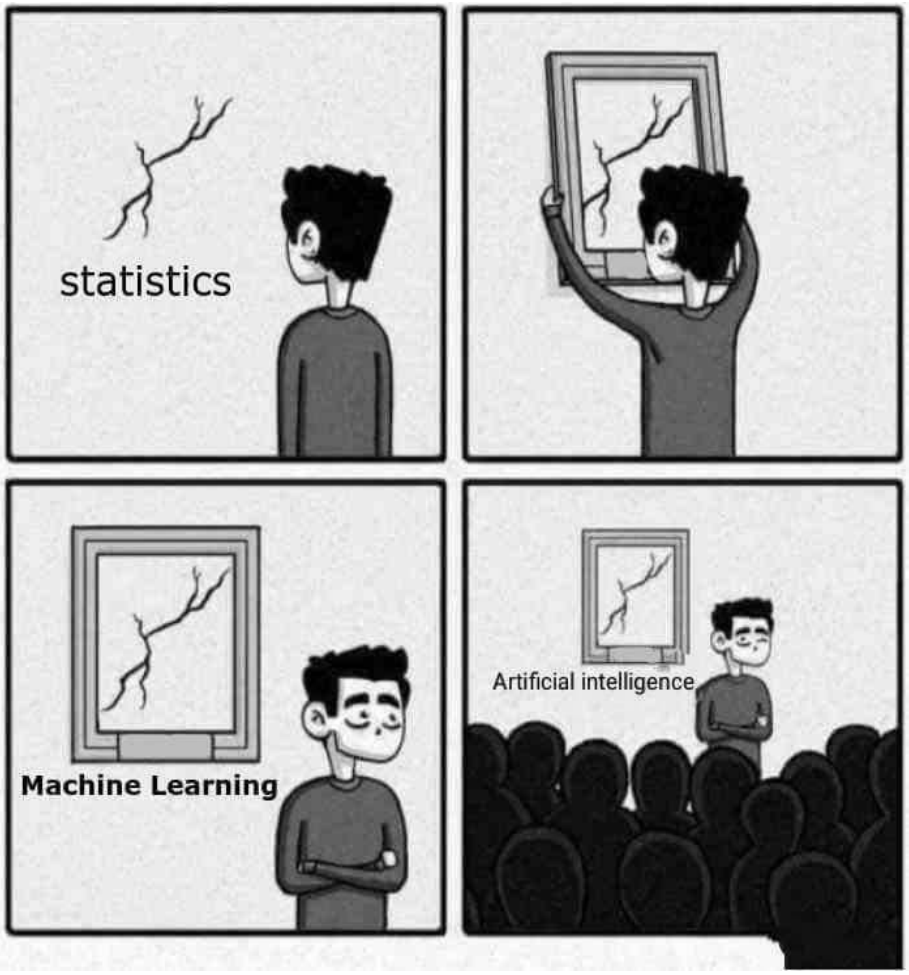
Como cada uno de los términos anteriores incluye un conjunto de técnicas, no hay un consenso generalizado acerca de qué técnicas o herramientas habría que incluir en cada uno de ellos pues no son excluyentes. Así, cada profesional o científico va a dar una definición ligeramente diferente de cada uno de los términos anteriores.

En la imagen siguiente, donde hay diagramas tomados de diferentes páginas web, se observa la falta de coherencia entre las intersecciones de los términos.

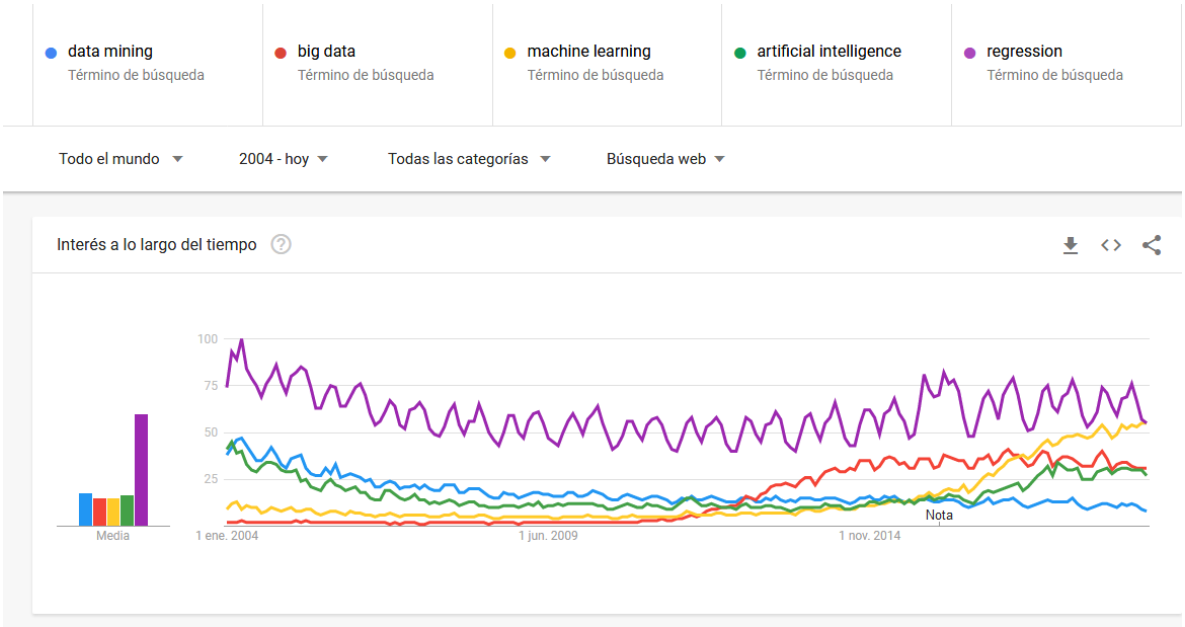


El marketing del tratamiento de datos

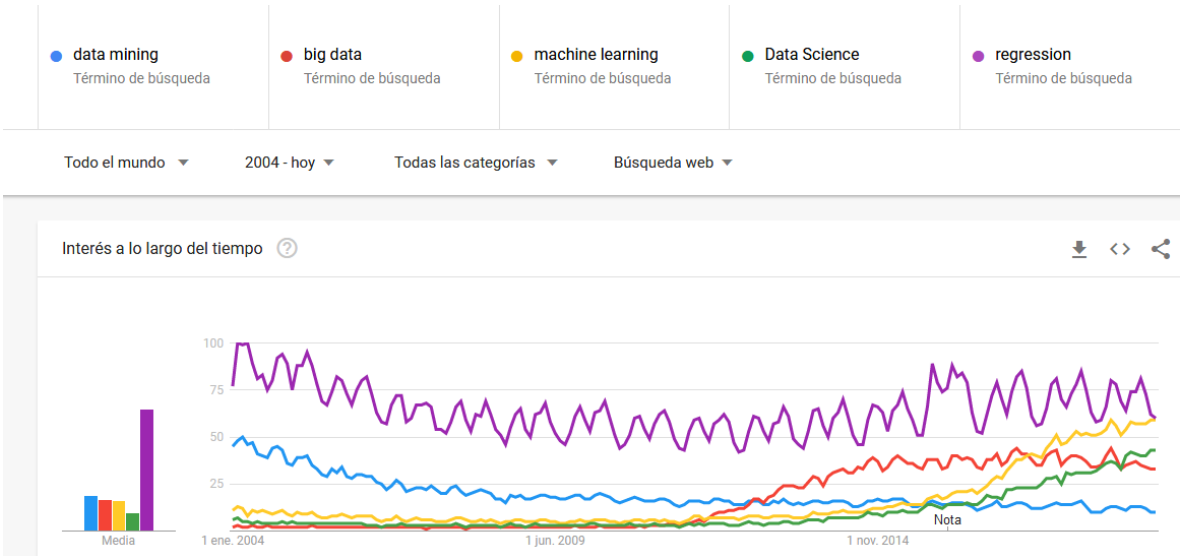
Muchos piensan que se han ido incorporando términos nuevos sin necesidad, con el objetivo corporativo de hacer más atractivas las técnicas de tratamiento de datos para el público y empresas, dando una apariencia futurista a modelos y herramientas matemáticas que a veces tienen hasta 200 años (como la regresión, usada por Gauss en 1809). El caso es que funciona y al irse incorporando nuevas técnicas al campo tradicional de la estadística (término que para muchos evoca simples enumeraciones o tablas, y tiene un aire antiguo y de función pública), la invención de nuevos términos está, al menos parcialmente, justificada.



En la imagen siguiente, búsquedas en Google Trends, se observa como el término Data Mining ha dejado de ser atractivo, mientras que la moda del Big Data e Inteligencia Artificial está estabilizándose. Machine Learning es un conjunto de técnicas muy frecuentemente utilizadas en la práctica empresarial y posiblemente por ello goce de más búsquedas. El modelo predictivo estadístico por excelencia, la regresión, sigue siendo más buscado que los otros términos pero se intuye que está comenzando a formar parte de algo más amplio y con los mismos objetivos, que es el Machine Learning.



En la siguiente imagen se sustituye Artificial Intelligence por Data Science (línea verde) y se observa que es también un término de moda.



Entonces…¿Qué es Machine Learning?

Aunque sea un término sin consenso concreto en cuanto a su definición, sí hay algo que está claro: todas las técnicas de predicción y clasificación supervisada están incluidas en cualquier definición que se pueda encontrar del término Machine Learning. En esta materia llamada Machine Learning se impartirán las técnicas predictivas más modernas, puesto que regresión y regresión logística han sido tratadas en otro módulo. Se estudiarán Redes Neuronales, modelos de árboles, Random Forest, Gradient boosting, Support Vector Machines y técnicas de ensamblado.

Software más utilizado para Machine learning

Lenguajes puros de programación más utilizados

- R
- Python
- Java
- C, C++
- SQL

Entornos o paquetes-plataformas, en general con lenguaje de programación integrado o con compilador de otros lenguajes (en cursiva los exclusivamente comerciales)

- SAS
- SPSS
- MATLAB
- EXCEL
- Spark
- H2o
- Rapidminer
- Knime
- Tableau

Datos de 2019, tomados de la web http://r4stats.com/articles/popularity/

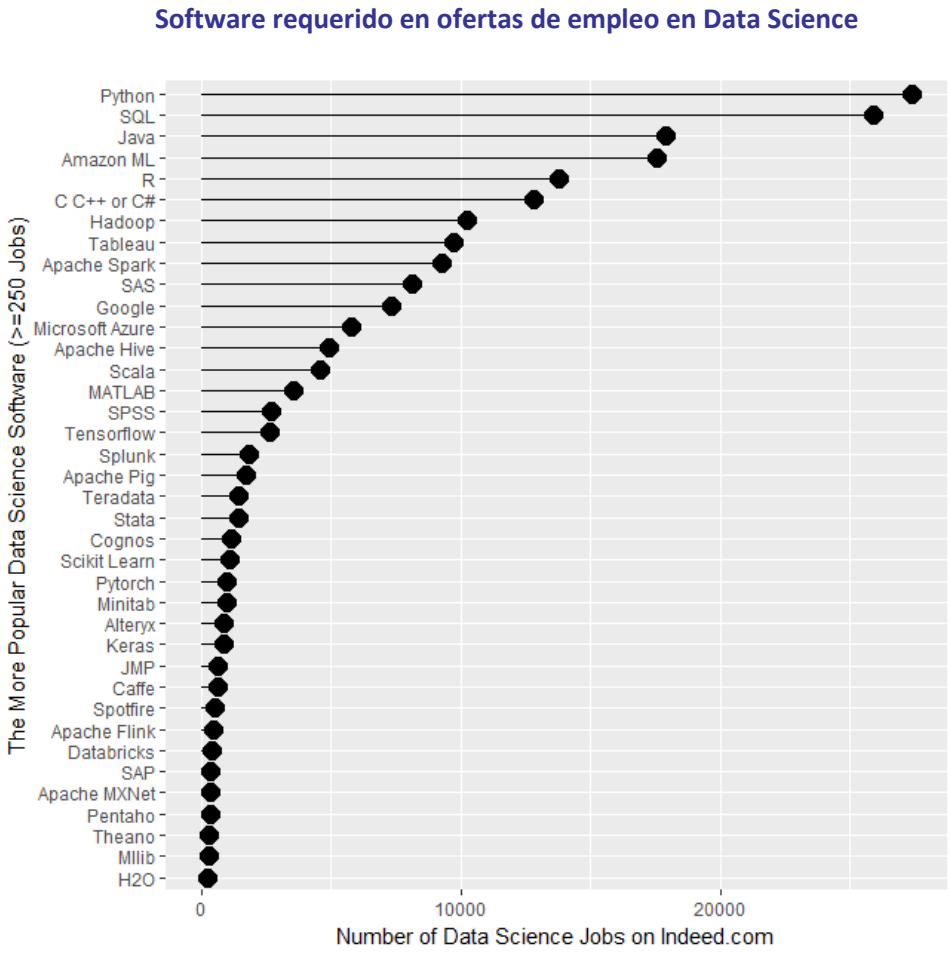
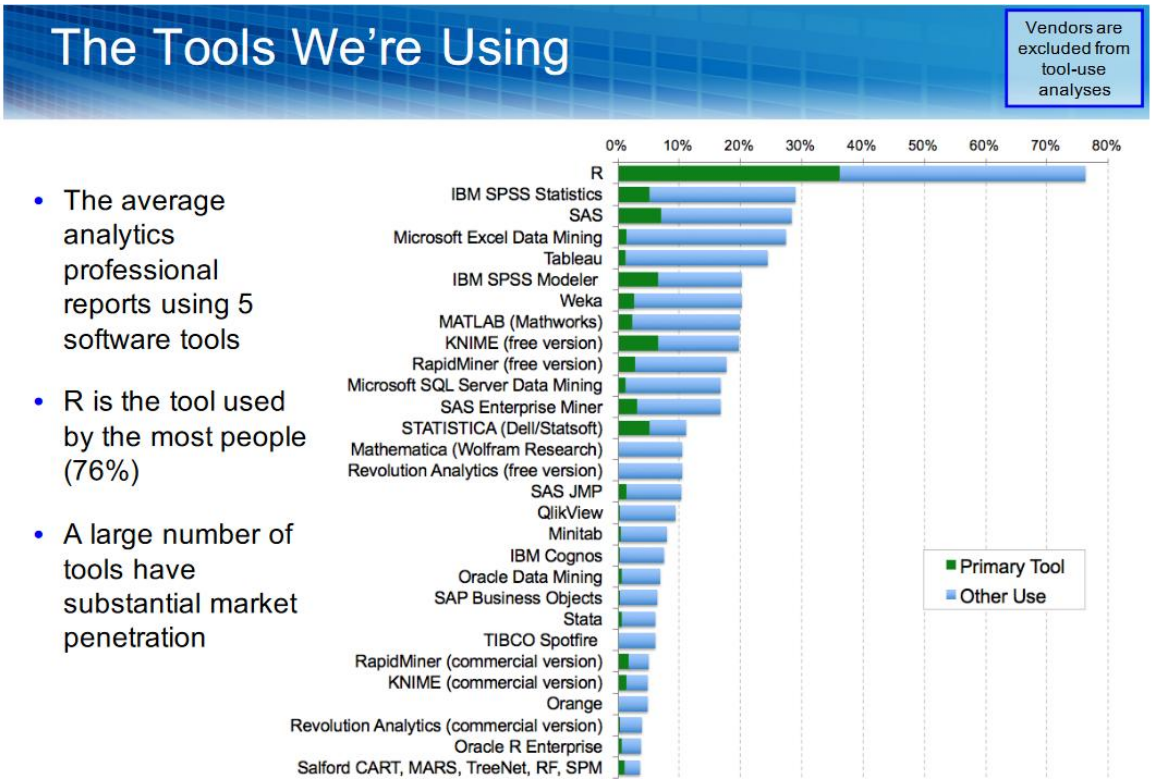


Figure 1a. The number of data science jobs for the more popular software (those with 250 jobs or more, 2019).

Datos de 2015-2017, tomados de Kdnuggets, encuesta a profesionales/usuarios



Temario de Machine Learning

1. Introducción
2. Redes Neuronales
3. Árboles
4. Bagging, Random Forest, Gradient Boosting
5. Support Vector Machines
6. Ensamblado
7. Comparación de algoritmos y estrategias avanzadas en Machine Learning

Evaluación

Será un trabajo de modelización predictiva:

Un trabajo largo de comparación de algoritmos predictivos sobre una variable dependiente binaria, utilizando redes, algoritmos basados en árboles (RF, GBM), SVM, y métodos de ensamblado.

Comentarios

1) Instalar los programas ejecutando en Rstudio el programa paquetes R a instalar master big data.R (h2o se puede dejar para el último tema). Lo mejor es tener la última versión de R antes de instalar programas.

2) Si se quiere, repasar los programas utilidades básicas R y Rstudio.R y utilidades minimas R machine learning.R

3) A lo largo de las lecciones, ir ejecutando los programas de los archivos indicados, observando resultados en la ventana workspace del Rstudio, en la consola, plots, etc. Los ejercicios son opcionales, algunos están corregidos. Se recomienda en todo caso tener un archivo personal de datos con variable dependiente binaria y otro con variable dependiente continua para ir probando sobre esos archivos las técnicas vistas en las lecciones. En el archivo recopilación links datasets machine learning.txt hay repositorios de internet donde se pueden descargar archivos de datos.

4) En cada tema hay bibliografía pero se espera que el módulo sea autocontenido, es decir que toda la información que se necesita está en los documentos y material aportado.