Científico de datos o data scientist

La demanda de científicos de datos se ha incrementado de manera constante en los últimos años, existe en el mercado una amplia oferta para los científicos de datos. Por otro lado las empresas son cada vez más conscientes de la necesidad de aplicar técnicas de Machine Learning para explotar los datos que tienen y no perder el tren de la competencia.

¿Qué aporta Big Data?, para poder contestar a esta pregunta, simplemente debemos tener en cuenta la diferencia fundamental en el cambio de enfoque. Los sistemas tradicionales de BI nos ayudan a contestar que ha pasado mediante el seguimiento, normalmente a través de cuadros de mandos de qué ha ocurrido. El cambio fundamental y el valor que aporta big data no es simplemente el uso de muchas fuentes y diversas de datos, sino el poder preguntar a los datos ¿Qué es lo que va a ocurrir?, la clave es Machine Learning.

Para poder construir sistemas que aporten un valor diferencial debemos construir modelos Machine Learning capaces de contestar a las preguntas de negocio. Esto es la misión de un científico de datos.

Parte del trabajo del científico de datos es la captura, depuración y almacenamiento de la información en un formato adecuado para su tratamiento y análisis.

Datascience se encarga del análisis de información usando técnicas y teorías en muchos campos como las matemáticas, la estadística, reconocimiento de patrones, visualización, etc, con el objetivo final de obtener datos.

Python para científicos de datos

http://python-xy.github.io/

Definiciones de aprendizaje

Una de las tareas más desafiantes en la ciencia de la computación es construir máquinas o programas de computadoras que sean capaces de aprender. El darles la capacidad de aprendizaje a las máquinas abre una amplia gama de nuevas aplicaciones. El entender también cómo estas pueden aprender nos puede ayudar a entender las capacidades y limitaciones humanas de aprendizaje.

Algunas definiciones de aprendizaje son:

Cambios adaptativos en el sistema para hacer la misma tarea de la misma población de una manera más eficiente y efectiva la próxima vez [Simon, 83].

Un programa de computadora se dice que aprende de experiencia E con respecto a una clase de tareas T y medida de desempeño D, si su desempeño en las tareas en T, medidas con D, mejoran con experiencia E [Mitchell, 97].

En general, se busca construir programas que mejoren automáticamente con la experiencia.

Definición de machine learning

https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Machine Learning es un término asociado a la Inteligencia Artificial que trata sobre el estudio de sistemas que pueden aprender a partir del análisis de los datos que reciben.

Con machine learning nos referimos al hecho de que las máquinas puedan aprender mediante programas capaces de generalizar y automatizar comportamientos a partir de unos datos de entrada.

El Machine Learning (ML) o Aprendizaje Autónomo es una rama de la Inteligencia Artificial (IA) que tiene como objetivo crear sistemas capaces de aprender por ellos mismos a partir de un conjunto de datos (data set), sin ser programados de forma explícita.

Para que estos sistemas puedan aprender por ellos mismos, se utilizan una serie de técnicas y algoritmos capaces de crear modelos predictivos, patrones de comportamiento,etc. Aunque no existe en la bibliografía actual un listado concreto y acotado de aquellas técnicas y algoritmos que se enmarcan dentro de la rama del ML (aunque hay algunas técnicas que claramente son propias de dicha área), si que podemos decir que en el área del ML encaja todo proceso de resolución de problemas, basados más o menos explícitamente en una aplicación rigurosa de la teoría de la decisión estadística; por tanto, es muy normal que el área del ML se solape con el área de la estadística.

Algunas definiciones:

Arthur Samuel (1959): El Machine Learning es un campo de estudio que da a las computadoras la capacidad de aprender sin ser programadas de forma explícita.

Tom Michell (1998): Un programa se dice que aprende de una experiencia ‘E’ con respecto a alguna tarea ‘T’ y alguna medida de rendimiento ‘R’, si su rendimiento en ‘T’ medida por ‘R’, mejora con la experiencia ‘E’.

El aprendizaje automático es un conjunto de técnicas pertenecientes al campo de la inteligencia artificial que permiten descubrir patrones y aprender modelos a partir de los datos.Algunos de **ejemplos de aplicaciones** de problemas que se pueden solucionar empleando técnicas de machine learning:

- **Detectar automáticamente si un correo es spam** a partir de su contenido,basándose en los reportes que han hecho los usuarios sobre correos electrónicos anteriores
- Predecir si un gasto realizado con una tarjeta de crédito es legítimo o fraudulento,en función del histórico de gastos del portador de la tarjeta
- **Predecir el gasto** que va a realizar un usuario en nuestro comercio en función de la información geográfica que disponemos sobre ese usuario
- **spam y fraude(problemas de clasficación).** La clasificación consiste en aprender un modelo a partir de datos que están previamente clasificados o etiquetados,que puede explotarse para predecir la clase de nuevos datos ,como pueden ser nuevos correos electrónicos que no sabemos si son spam.u operaciones con tarjeta que en el momento de efectuarse no sabemos si son fraudulentas o no

- **predecir el gasto de usuario(problema de regresión)** .La diferencia entre los problemas de clasificación y los de regresión es que en de clasificación se trata de predecir el valor de una clase,categoría o etiqueta,mientras que en los de regresión se trata de predecir un valor numérico

Muchos de los servicios que utilizamos en nuestro día a día como google, gmail, netflix, spotify o amazon se valen de las herramientas que les brinda el Machine Learning para alcanzar un servicio cada vez más personalizado y lograr así ventajas competitivas sobre sus rivales.

Ejemplos clásicos de machine learning

- Clasificar email como spam/no spam
- Reconocimiento de caracteres
- Detección de patrones en imágenes
- Reconocimiento de voz
- Detección de fraude en transferencias mediante tarjetas de crédito
- Predicción de demanda,impagos,abandono del cliente por parte de compañías telefónicas
- Predicciones económicas
- Sistemas de recomendación(spotify,amazon)
- Clasificar clientes en campañas de marketing
- Diagnóstico de enfermedades

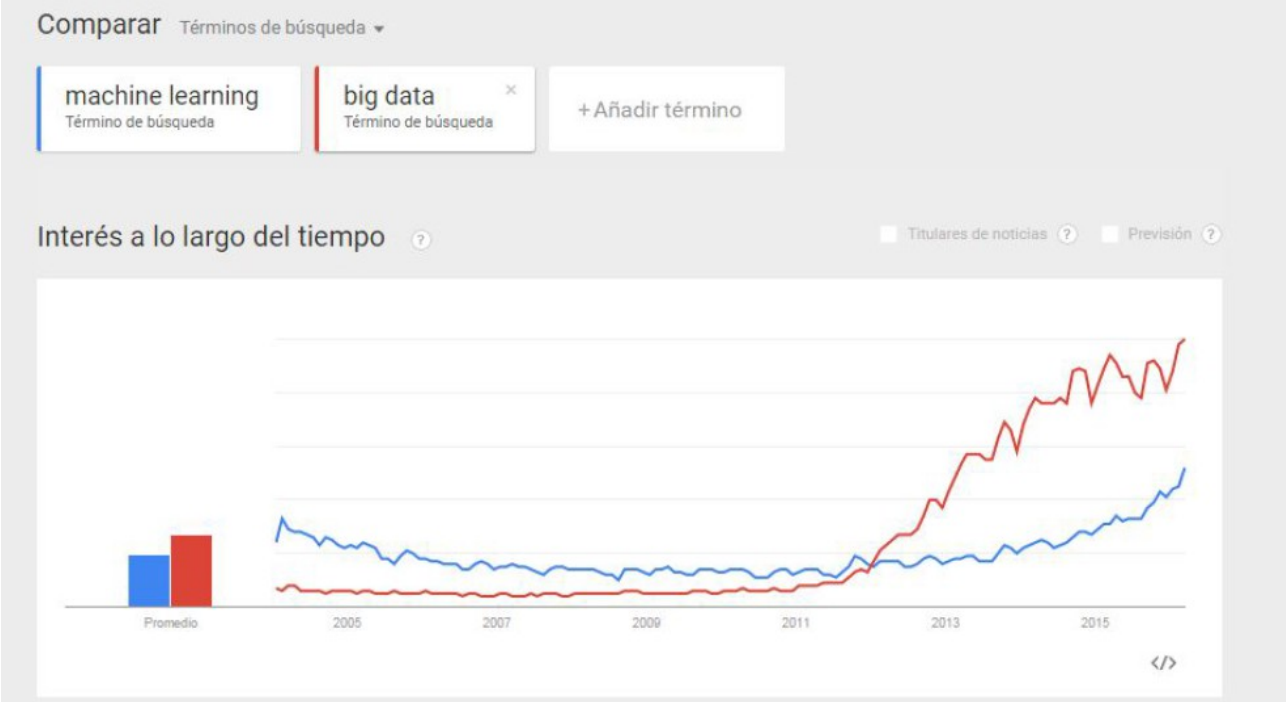
El Machine Learning o " Aprendizaje automático " es un área que lleva con nosotros ya unos cuantos años. Básicamente, el objetivo de este campo de la Inteligencia Artificial, es que los algoritmos, las reglas de codificación de nuestros objetivos de resolución de un problema, aprendan por sí solos. De ahí lo de "aprendizaje automático". Es decir, que los propios algoritmos generalicen conocimiento y lo induzcan a partir de los comportamientos que van observando.

Para que su aprendizaje sea bueno, preciso y efectivo, necesitan datos.

Cuantos más, mejor. De ahí que cuando irrumpe el Big Data (este nuevo paradigma de grandes cantidades de datos) el término Machine Learning irrumpe con más fuerza.

Los patrones, tendencias e interrelaciones entre las variables que el algoritmo de Machine Learning observa, se pueden ahora obtener con una mayor precisión gracias a la disponibilidad de datos.

Con el auge de las redes sociales y las grandes empresas tecnológicas que generan datos a un gran volumen, velocidad y variedad (Google, Amazon, etc.), esto se generaliza a otros sectores. Ahora, se convierten en pieza clave del día a día de muchas compañías, que ven cómo el gran volumen de datos además, les ayuda a obtener más valor de la forma de trabajar que tienen. En la siguiente ilustración que nos genera Google Trends sobre el volumen de búsqueda de ambos términos se puede observar cómo el **"Machine Learning" se destaca de nuevo cuando el Big Data entra en el "mainstream",sobretodo a partir del 2013**



¿Y por qué le ha venido tan bien al Machine Learning el Big Data? Básicamente porque como la palabra "aprendizaje" viene a ilustrar, los algoritmos necesitan de datos, primero para aprender, y segundo para obtener resultados. Cuando los datos eran limitados, corríamos el peligro de sufrir problemas de " *underfitting* ". Es decir, de entrenar poco al modelo, y que éste perdiera precisión. Y, si utilizábamos todos los datos para entrenar al modelo, nos podría pasar lo contrario, problemas de " *overfitting* ", que entonces nos generaría modelos demasiado ajustados a la muestra, y quizás, poco generalizables a otros casos.

Este problema con el Big Data desaparece. Tenemos tantos datos, que no nos debe preocupar el equilibrio entre " datos de entrenamiento " y " datos para testar y probar el modelo y su eficiencia/precisión ". La optimización del rendimiento del modelo (el "Just Right" de la gráfica anterior) ahora se puede elegir con mayor flexibilidad, dado que podemos disponer de datos para llegar a ese punto de equilibrio.

Sistemas expertos

Como se ha comentado en la definición de ML, este área debe de crear sistemas que tienen que ser capaces de aprender por ellos mismos sin ser programados de forma explícita, con la finalidad de predecir hechos futuros, realizar recomendaciones, clasificaciones de elementos, eventos, tags, etc. Por tanto; después de una fase de aprendizaje, tendremos un "sistema experto" que dada una determinada entrada nos proporcionará una salida (predicción, recomendación, clasificación, etc.), como resultado de haber aplicado una función de regresión o clasificación (que debe de aprender el sistema) sobre los datos de entrada.

Dos ejemplos de sistemas expertos creados tras aplicar alguna/s técnica/s de ML, serían los siguientes: uno, un sistema experto en predicción de quinielas (clasificación), que pasándole el nombre del equipo local y visitante, devuelve como resultado una de las tres opciones de la quiniela (1, X, 2); y otro un sistema experto en el cálculo de calorías quemadas (regresión) al hacer carrera continua (running), en el que pasándole como entrada el peso de la persona, el tiempo de carrera y la velocidad, devuelva como resultado el número de calorías quemadas.

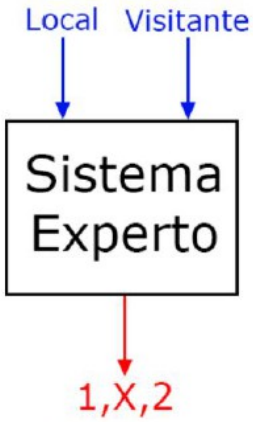


Ilustración 2: Esquema de un sistema experto en predicción de quinielas

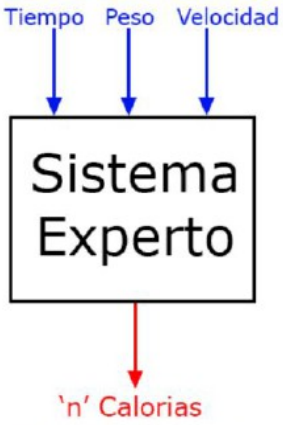


Ilustración 3: Esquema de un sistema experto en predicción de calorías quemadas

Como hemos visto, estos sistemas tienen dos formas de proporcionar un resultado: uno; la clasificación, que devuelve como salida un conjunto finito de resultados; generalmente pequeño, (y={0,1}, y={1,X,2}, y={si,no}) y otro; la regresión, que devuelve como salida un valor arbitrario (un número real, un vector de números reales, cadenas de símbolos, etc.).