Python y Ciencia de Datos

XVII Semana académica de Matemáticas Aplicadas y Computación

Oscar Daniel Acosta González
Licenciado en Matemáticas Aplicadas y Computación

2 de octubre de 2018

Índice

1. Introducción

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value." - Clive Humby

Algunos expertos consideran que los datos son el nuevo petroleo, la analogía encaja perfectamente, ya que el petroleo en sí, no aporta mucho valor, lo realmente valioso es la explotación, la extracción y su manufactura. Lo mismo ocurre con los datos, su verdadero valor reside en el valor que aportan a las empresas, de otro modo simplemente son recursos desaprovechados al estar almacenados en Bases de Datos, Data Warehouses o Data Lakes. La explotación de los datos se hace mediante tecnicas matemáticas, computacionales y de inteligencia artifical, con el objetivo de encontrar información relevante.

Uno de los objetivos de la ciencia es la generación de información para conocimiento del mundo y su progreso, por tanto, la ciencia de datos se enfoca en la obtención de conocimiento pero ésta lo hace a partir de los datos.

La Ciencia de datos ha emergido en el ambiente mundial de negocios debido a su naturaleza aplicada y accionable dentro de cualquier contexto organizacional siendo ésta una rama multidisciplinaria del conocimiento que involucra a las Matemáticas Aplicadas, Ciencias de la Computación, Administración y Comunicación, valiéndose de técnicas y habilidades específicas de las disciplinas mencionadas para generar valor mediante la inteligencia latente contenida en grandes volúmenes de datos que son generados y almacenados diariamente por las organizaciones.

La Ciencia de Datos se considera una evolución multidisciplinaria en los campos de análisis de negocio, ciencias de la computación, modelación matemática, estadística, analítica y minería de datos.

La ciencia de datos puede equipararse a un avión en caída libre: el panel de control es el que da los indicadores clave de la nave, esta parte podría compararse con Bussines Intelligence, mientras que la analítica es la que dice que palanca accionar de entre cientos de interruptores para evitar la caída del avión.

A continuación se presentan algunos conceptos clave que comunmente suelen confundirse:

1.1. Estadística

Ciencia que utiliza conjuntos de datos numéricos para obtener, a partir de ellos, inferencias basadas en el cálculo de probabilidades.

En resuidas cuentas, es un análisis superficial de los datos, pero que en realidad no aporta mucha información más allá de lo evidente, dado que no se realiza a profundiad o de forma multivariada.

1.2. Minería de Datos

La minería de datos o exploración de datos (es la etapa de análisis de "Knowledge Discovery in Databases.º KDD) es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

La minería de datos se enfoca a la exploración de volumenes de datos considerables, sin embargo se limita a encontrar patrones, sin darles un sentido realmente útil para el contexto.

1.3. Bussines Intelligence

Es una amplia categoría de soluciones de software informático que permite a una empresa u organización, obtener información de sus operaciones más críticas a través de aplicaciones de generación de informes y herramientas de análisis.

Desde un punto de vista más pragmático, y asociándolo directamente con las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte

a la toma de decisiones sobre el negocio.

En resumidas cuentas, BI se enfoca al reporte de la información, ya sea mediante documentos, tableros, tablas o infografías. Más que un elemento de acción, es un agente de información y monitoreo.

1.4. Analytics

Es un campo multidimensional que usa matemáticas, estadísticas, modelación predictiva y técnicas de machine learning para encontrar patrones significativos y conocimiento a partir de los datos.

Su gran diferencia con el resto de las técnicas es el aporte de información al negocio, es decir, resuelve problemas de forma informada a partir de datos reales con soluciones reales.

1.5. Big Data

Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – que inundan una empresa todos los días. Pero no es la cantidad de datos lo importante. Lo que importa es lo que las organizaciones hacen con los datos. El big data puede ser analizado para obtener insights que conlleven a mejores decisiones y acciones de negocios estratégicas.

Es un término muy común hoy en día, donde Hadoop, Spark y Scala son los nombres de software que más resuenan.

1.6. Inteligencia Artificial

La inteligencia artificial (IA) hace posible que las máquina aprendan de la experiencia, se ajusten a nuevas aportaciones y realicen tareas como hacen los humanos. La mayoría de los ejemplos de inteligencia artificial de los que usted escucha hoy día – desde computadoras que juegan ajedrez hasta automóviles que se conducen por sí solos – se sustentan mayormente en aprendizaje a fondo (deep learning) y procesamiento del lenguaje natural. Mediante el uso de estas tecnologías, las computadoras pueden ser entrenadas para realizar tareas específicas procesando grandes cantidades de datos

y reconociendo patrones en los datos.

1.7. Internet de las cosas (IoT

Internet de las Cosas es el concepto de objetos de todos los días – desde máquinas industriales hasta dispositivos de vestir (weareble devices) – mediante el uso de sensores integrados para recopilar datos y seguir una acción con esos datos a través de una red. De modo que un edificio que utiliza sensores para ajustar automáticamente la calefacción y la iluminación. O bien equipo de producción que alerta al personal de mantenimiento de un fallo inminente. Dicho de manera simple, Internet de las Cosas es el futuro de la tecnología que puede hacer nuestras vidas más eficientes.

1.8. Ciencia de datos

La ciencia de datos es el conocimiento y dominio de la mayoría de las técnicas previamente mencionadas, todas con un enfoque prático y con sentido de negocio, es decir, que resuelva una necesidad a través de estas técnicas.

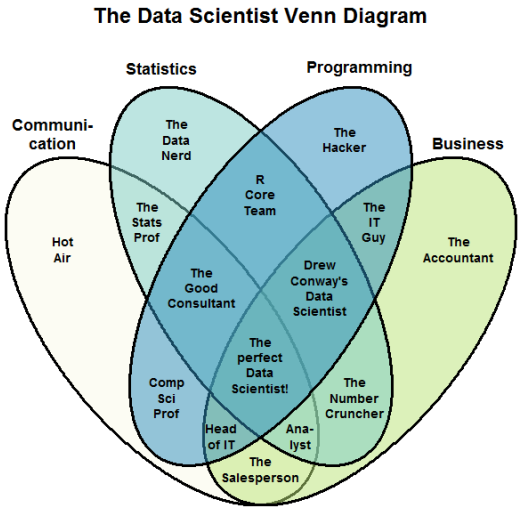
En términos generales, Data Science es el conjunto de prácticas sobre almacenamiento, gestión y análisis de conjuntos de datos lo suficientemente grandes que requieren de computación distribuida y los recursos de almacenamiento. En la actualidad la mayoría de las fuentes de datos están en internet y relacionadas con las transacciones, pero no hay que ignorar las primeras aplicaciones de la física de alta energía, la meteorología, las simulaciones militares, así como futuras aplicaciones en ciencias

1.9. Cienífico de datos

Por algunos, es considerado el trabajo más sexy del siglo 21. Requiere de una mezcla de aptitudes multidisciplinarias que abarcan la intersección de matemáticas, estadística, ciencias de la computación, comunicación y negocios. Encontrar a un científico de datos es dificl. Encontrar a una persona que entienda qué es un científico de datos es igualmente dificil.

El científico de datos es un profesionista con conocimientos en diversas áreas

que aborda los problemas desde una perspectiva de datos con un sentido muy fuerte de negocio. A continuación se presenta un diagrama de Venn que ilustra mejor el perfil:



2. Esquema del análisis predictivo

La analítica predictiva permite a las organizaciones ser un poco más proactivas, tener la vista en el futuro, anticipando resultados y comportamientos, basándose así en datos y no en una serie de especulaciones. La analítica prescriptiva va un poco más allá y sugiere acciones que podemos poner en marcha, a raíz de las predicciones y sus implicaciones.

A grandes rasgos, la modelación predictiva se divide de la siguiente manera:

2.1. Modelación Supervisada

Se refiere a aquellos modelos donde lo que se intenta predecir es un fenómeno ya conocido, esto con base en información que no depende del fe nómeno.

Bajo el esquema Matemático, la modelación supervisada se resume de la siguiente manera:

$$\widehat{y} = f(\overrightarrow{X}) \tag{1}$$

Donde $\hat{y}$ es la variable objetivo o evento a pronosticar, f es una función vectorial del estilo regresión lineal, árbol de decisión, etcetera y X es la matriz de las variables predictoras de dimensión $n.$ m con n variables y m registros.

De acuerdo con la naturaleza de la variable objetivo, la modelación supervisada puede clasificarse en dos tipos:

2.1.1. Problema de regresión

Ocurre cuando el fenómeno a pronosticar es de carácter cotinuo, es decir, el dominio de y son todos o un subconjunto de los número reales. Un ejemplo de esto sería el número de *shares* que una noticia por internet obtiene, pronósticos de ventas, etc.

La manera clásica de abordar estos problemas es con una regresión lineal, donde al final se obtiene la mejor aproximación; sin embargo, el uso de técnicas más robustas como áboles de decisión o redes neuronales ayudan a que

la precisión del modelo aumente, aunque en algunos casos en demérito de la interpretación o entendimiento.

2.1.2. Problema de clasificación

Sucede cuando la variable objetivo a predecir es de caracter discreto, es decir, se restringe a un conjunto de valores muy pequeño. Un ejemplo de esto sería sí un correo es *spam*, si un cliente va a pagar o no, o bien, el sentimiento del público hacia una publicación en redes sociales.

Frecuentemente se utilizan técnicas como Regresión Logística, para medir la probabilidad de la ocurrencia del evento; sin embargo existen otras técnicas como los árboles o incluso ensambles para ambos casos.

3. Proceso "general"de modelado

Más allá de hablar de una receta o de un proceso estandarizado o general, los pasos que a continuación se presentan son lo que en su mayoría en el campo laboral son necesarios.

3.1. Planteamiento

Se refiere al conocimiento del problema y como atacarlo. Ningún problema puede comenzar en otro punto y lo más recomendable es detenerse en este punto bastante tiempo para lograr acotar el problema, conocer con que se cuenta y que hace falta, además de el punto al que se quiere llegar. Por otro lado, es en este punto en donde se definen los cursos de acción a tomar, o incluso el descarte completo del problema dadas las condiciones actuales.

3.2. Extracción

Hace referencia a obtener los datos que describan de alguna manera el fenómeno. Si bien es la fase más corta del proceso, lo realmente importante aquí es que dicha extracción se haga de forma óptima, es decir, que solamente se utilice la información relevante, así como la muestra a considerar, reduciendo en gran medida los costos computacionales que esto ocasiona, además de que al usar algoritmos, estos tarden mucho menos en entrenarse. La reducción de la muestra a utilizar se decide en su totalidad en la etapa anterior.

3.3. Exploración y limpieza

Análisis exploratiorio y ajuste de la muestra de modelado. Se refiere principalmente a conocer los datos con los que se trabaja, ya que cada muestra es única en cuanto a sus variables. Esta fase debería ser una de las más extensas, para evitar comportamientos anómalos y terminar de acotar el problema. En esta fase se decide que variables se van a utilizar, variable objetivo, tratamiento de valores ausentes y extremos.

3.4. Modelado

Ajuste de la solución propuesta correspondiente al problema. Si las fases anteriores fueron realizadas correctamente, el modelado no debería representar gran problema, sin embargo, es en esta parte que se deben de probar diversos algoritmos que den solución a nuestro problema, para finalmente elegir aquella que mejor ajuste tenga, o bien, aquella que mejor satisfaga al usuario final.

3.5. Ajuste y monitoreo

Selección de la mejor opción, así como el monitoreo y posible ajuste. Una vez que se elige la mejor solución, esta debe de pasar por diversas pruebas que garanticen su usabilidad y robustez, ya que los modelos son efimeros y frecuentemente tienen que estar calibrandose y cambiando. Este punto inicia un proceso iterativo con el paso anterior, ya que mientras se tenga con vida al modelo, éste tiene que adaptarse a nuevas condiciones para seguir funcionando.

3.6. Comunicación

Parte fundamental del proceso, otorga la solución a los directivos. Es sin duda la cumbre de todo el proceso, ya que de nada sirve un modelo que no convenza a aquellos que lo necesitan. Se debe de generar la necesidad de la solución, demostrar que es lo suficientemente robusta y adecuada para el problema, pero sobre todo, hay que saber venderla, convencer a los directivos de que es la mejor solución.

4. Python

El lenguaje Python surgió a principios de los 90 e inicialmente fue desarrollado por Guido Van Rossum, un ingeniero holandés que trabajaba en ese momento en el CWI de Amsterdam, el Centro de Investigación de Ciencias de la Computación holandés.

Python surgió como un hobby para Guido y su nombre, Python, fue tomado del grupo cómico británico Monty Python, del que Guido era un gran fan. Desde sus comienzos, Python nació como un proyecto de software libre y posiblemente deba parte de su éxito a la decisión de hacerlo código abierto. A recientes fechas, este lenguaje se ha vuelto muy popular, dada su versatilidad y otros atributos que a continuación se presentan:

- Lenguaje multipropósito: Es utilizado como lenguaje Orientado a Objetos, Desarrollo Web, ETLs, Machine Learning, etc.
- Lenguaje interpretado: Se va ejecutando conforme pasa por el interprete, si bien no alcanza velocidades de C o C++, es bastante veloz.
- Fácil de aprender: El ser un lenguaje con una sintaxis muy intuitiva, le permite al usuario familiarizarse muy facilmente con él.
- Machine Learning: Contiene interfaces par la creación de algoritmos muy sencillas de utilizar, dejando el trabajo dificil al lenguaje y cediendo al usuario otras tareas de un nivel superior.

"You can best learn data mining and data science by doing, so start analyzing data as soon as you can! However, don't forget to learn the theory, since you need a good statistical and machine learning foundation to understand what you are doing and to find real nuggets of value in the noise of big data." - Gregory Piatetsky-Shapiro, President, KDnuggets