IBM Analytics
Documento técnico

Metodología fundamental para la ciencia de datos

En el campo de la ciencia de datos, resolver problemas y responder preguntas a través del análisis de datos es una práctica estándar. A menudo, los científicos de datos construyen un modelo para predecir resultados o descubrir patrones subyacentes, con el objetivo de obtener conocimientos. Las organizaciones pueden utilizar estos conocimientos para tomar medidas que, idealmente, mejoren los resultados futuros.

Existen numerosas técnicas de rápida evolución para el análisis de datos y la construcción de modelos. En un plazo de tiempo increíblemente breve, han progresado para dejar de ser escritorios y convertirse en depósitos paralelos a gran escala con enormes volúmenes de datos y la funcionalidad analítica incorporada en bases de datos relacionales y Apache Hadoop. Los análisis de texto sobre datos no estructurados o semiestructurados tienen cada vez mayor importancia como una manera de incorporar tendencias y otra información útil a partir del texto en modelos predictivos, lo que con frecuencia genera mejoras significativas en la calidad y la precisión del modelo.

Los enfoques analíticos emergentes buscan automatizar muchos de los pasos que intervienen en la construcción y aplicación de modelos, lo que permite que la tecnología de aprendizaje automático sea más accesible para quienes no cuentan con amplias aptitudes cuantitativas. Además, a diferencia del enfoque "descendente" en el que primero se define el problema empresarial y después se analizan los datos para buscar una solución, algunos científicos de datos pueden usar un enfoque "ascendente". Con este último, los científicos de datos analizan grandes volúmenes de datos para ver qué meta empresarial podría sugerirse según los datos, y luego abordan ese problema. Dado que la mayoría de los problemas se tratan de una manera descendente, la metodología de este documento refleja esa visión.

Una metodología para la ciencia de datos de 10 etapas que abarca tecnologías y enfoques

A medida que las capacidades de análisis de datos se vuelven más accesibles y frecuentes, los científicos de datos necesitan una metodología fundamental que sea capaz de ofrecer una estrategia orientativa, independientemente de las tecnologías, los volúmenes de datos o los enfoques que intervengan (vea la Figura 1). Esta metodología mantiene algunas similitudes con las metodologías reconocidas[1-5] para la minería de datos, pero enfatiza varias de las prácticas nuevas en ciencia de datos, como el uso de volúmenes de datos muy grandes, la incorporación de análisis de texto en el modelado predictivo y la automatización de algunos procesos.

La metodología consta de 10 etapas que conforman un proceso iterativo para la obtención de conocimientos mediante el uso de datos. Cada etapa cumple una función crucial en el contexto de la metodología general.

¿Qué es una metodología?

Una metodología es una estrategia general que orienta los procesos y las actividades dentro de un campo determinado. La metodología no depende de tecnologías o herramientas específicas, ni se trata de un conjunto de técnicas o fórmulas En cambio, una metodología proporciona al científico de datos un marco que determina la manera de proceder con los métodos, procesos y heurística que se utilizarán para obtener las respuestas o los resultados.

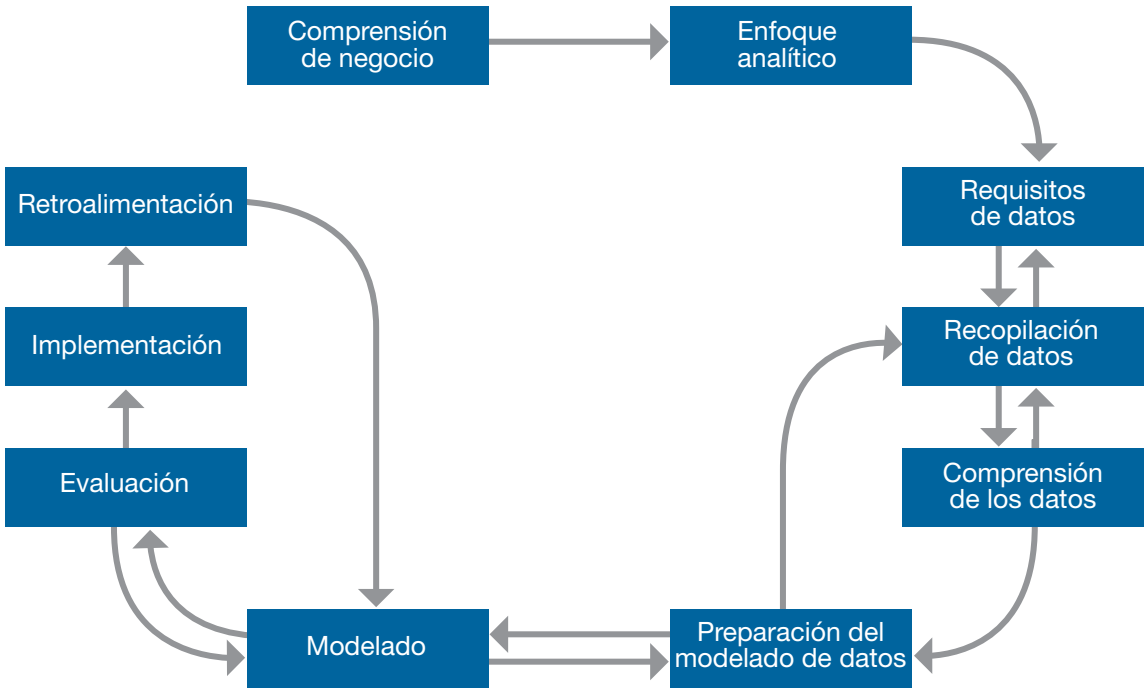

Figura 1. Metodología fundamental para la ciencia de datos.

Etapa 1: Comprensión de negocio

Cada proyecto comienza con la comprensión de negocio. Los patrocinadores empresariales que necesitan la solución analítica desempeñan la función más crítica en esta etapa, ya que definen el problema, los objetivos del proyecto y los requisitos de la solución desde una perspectiva empresarial. Esta primera etapa establece las bases para resolver con éxito el problema empresarial. Para ayudar a garantizar el éxito del proyecto, los patrocinadores deben participar durante todo el proyecto para brindar conocimientos especializados, revisar los resultados intermedios y asegurar que el trabajo siga por buen camino para generar la solución prevista.

Etapa 2: Enfoque analítico

Una vez que se haya establecido claramente el problema empresarial, el científico de datos puede definir el enfoque analítico para resolver el problema. Esta etapa implica expresar el problema en el contexto de técnicas estadísticas y de aprendizaje automático, de manera que la organización pueda identificar las más adecuadas para el resultado deseado. Por ejemplo, si el objetivo es predecir una respuesta como "sí" o "no", el enfoque analítico podría definirse como la creación, la prueba y la implementación de un modelo de clasificación.

Etapa 3: Requisitos de datos

El enfoque analítico elegido determina los requisitos de datos. Específicamente, los métodos analíticos que se usarán requieren determinados formatos, contenido y representaciones de datos, guiados por los conocimientos especializados.

Etapa 4: Recopilación de datos

En la etapa inicial de recopilación de datos, los científicos de datos identifican y recopilan los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) relevantes para la naturaleza del problema. Generalmente, deben elegir si desean realizar inversiones adicionales para obtener elementos de datos menos accesibles. La mejor opción podría ser postergar la decisión de inversión hasta obtener más información sobre los datos y el modelo. En caso de brechas en la recopilación de datos, es posible que el científico de datos deba revisar los requisitos de datos en consecuencia y recopilar datos nuevos o adicionales.

Si bien el muestreo y la subdivisión de datos continúan siendo importantes, las plataformas de alto rendimiento y la funcionalidad analítica incorporada en las bases de datos de la actualidad permiten a los científicos de datos utilizar conjuntos de datos mucho más grandes que contengan gran parte de los datos disponibles (o incluso todos). Al incorporar más datos, los modelos predictivos pueden representar mejorar los eventos inusuales, como la incidencia de enfermedades o la falla del sistema.

Etapa 5: Comprensión de los datos

Después de la recopilación de datos original, los científicos de datos generalmente utilizan estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar la calidad de los datos y revelar los conocimientos iniciales sobre los datos. Podría ser necesario recopilar datos adicionales para llenar las brechas.

Etapa 6: Preparación de datos

Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la etapa de modelado posterior. Las actividades de preparación de datos incluyen la limpieza de datos (abordar valores faltantes o no válidos, eliminar duplicados, dar el formato correcto), la combinación de datos provenientes de varias fuentes (archivos, tablas, plataformas) y la transformación de los datos en variables más útiles.

En un proceso denominado *ingeniería de factores*, los científicos de datos pueden crear variables explicativas adicionales, también denominadas *indicadores* o *características*, a través de la combinación de conocimiento especializado y las variables estructuradas existentes. Cuando están disponibles datos de texto, como registros del centro de atención al cliente o las anotaciones de los médicos en formularios no estructurados o semiestructurados, los análisis de texto son útiles para producir nuevas variables estructuradas, enriquecer el conjunto de indicadores y mejorar la precisión del modelo.

La preparación de los datos generalmente es el paso más lento de un proyecto de ciencia de datos. En muchos campos, algunos pasos de la preparación de datos son comunes en problemas de diversa índole. Automatizar determinados pasos de la preparación de datos por adelantado puede acelerar el proceso, al minimizar el tiempo de preparación ad hoc. Con los sistemas paralelos a gran escala y la funcionalidad analítica de alto rendimiento de la actualidad incorporados en el sitio donde se almacenan los datos, los científicos de datos pueden preparar los datos con mayor facilidad y rapidez mediante el uso de conjuntos de datos muy grandes.

Etapa 7: Modelado

Comenzando por la primera versión del conjunto de datos preparado, la etapa de modelado se centra en el desarrollo de modelos predictivos o descriptivos de acuerdo con el enfoque analítico definido previamente. Con los modelos predictivos,

los científicos de datos utilizan un conjunto de *entrenamiento* (datos históricos en los que se conoce el resultado de interés) para construir el modelo. En general, el proceso de modelado es altamente itarativo a medida que las organizaciones adquieren conocimientos intermedios, lo que permite ajustar la preparación de datos y la especificación del modelo. Para una técnica determinada, los científicos de datos pueden probar varios algoritmos con sus respectivos parámetros para encontrar el mejor modelo para las variables disponibles.

Etapa 8: Evaluación

Durante el desarrollo del modelo y antes de la implementación, el científico de datos evalúa el modelo para comprender su calidad y garantizar que aborde de manera adecuada y completa el problema empresarial. La evaluación del modelo implica calcular diversas medidas de diagnóstico y otras salidas, como tablas y gráficos, que permitan al científico de datos interpretar la calidad del modelo y su eficacia para resolver el problema. Para un modelo predictivo, los científicos de datos utilizan un conjunto de prueba, que es independiente del conjunto de entrenamiento, pero que sigue la misma distribución de probabilidad y tiene un resultado conocido. El conjunto de prueba se utiliza para evaluar el modelo y ajustarlo si es necesario. A veces, el modelo final se aplica también a un conjunto de validación para realizar una evaluación final.

Además, los científicos de datos pueden asignar al modelo pruebas de relevancia estadística como una evidencia adicional de su calidad. Esta evidencia adicional puede resultar determinante para justificar la implementación del modelo o tomar medidas cuando los riesgos son altos, como un costoso protocolo médico complementario o un sistema crítico de vuelo de avión.

Etapa 9: Implementación

Una vez que se ha desarrollado un modelo satisfactorio y se obtiene la aprobación de los patrocinadores empresariales, se implementa en el entorno de producción o en un entorno de prueba comparable. A menudo, se implementa de una manera limitada, hasta que se haya evaluado por completo su rendimiento. La implementación puede ser tan sencilla como generar un informe con recomendaciones o tan complicada como integrar el modelo en un flujo de trabajo complejo y en un proceso de calificación administrado por una aplicación personalizada. Implementar un modelo en un proceso empresarial operativo generalmente requiere grupos, aptitudes y tecnologías adicionales de la empresa. Por ejemplo, un grupo de ventas puede implementar un modelo de propensión de respuesta a través de un proceso de gestión de campañas creado por un equipo de desarrollo y administrado por un grupo de marketing.

Etapa 10: Retroalimentación

Al recopilar los resultados del modelo implementado, la organización recibe la retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que fue implementado. Por ejemplo, la retroalimentación podría tener la forma de índices de respuesta a una campaña promocional dirigida a un grupo de clientes identificados por el modelo como “clientes con alta probabilidad de respuesta”. Analizar esta retroalimentación permite a los científicos de datos ajustar el modelo para mejorar su precisión y utilidad. Pueden automatizar total o parcialmente los pasos de obtención de retroalimentación y evaluación, ajuste y reimplementación del modelo para agilizar el proceso de actualización del modelo y obtener mejores resultados.

Proporcionar valor continuo a la organización

El flujo de la metodología ilustra la naturaleza iterativa del proceso de resolución del problema. A medida que los científicos de datos obtienen más información sobre los datos y el modelado, con frecuencia regresan a una etapa anterior para realizar ajustes. Los modelos no se crean una vez, se implementan y permanecen implementados tal como están, sino que se mejoran continuamente y se adaptan a las condiciones cambiantes a través de la retroalimentación, el ajuste y la reimplementación. De esta manera, tanto el modelo como el trabajo requerido pueden proporcionar valor continuo a la organización mientras se necesite la solución.

Para obtener más información

Un curso nuevo sobre la metodología fundamental para la ciencia de datos está disponible a través de Big Data University. El curso en línea gratuito está disponible en: http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

Para conocer ejemplos prácticos de cómo se ha implementado esta metodología en casos de uso reales, visite:

- http://ibm.co/1SUhxFm
- http://ibm.co/1lazTvG

Agradecimientos

Gracias a Michael Haide, Michael Wurst, Ph.D., Brandon MacKenzie y Gregory Rodd por sus comentarios útiles y a Jo A. Ramos por su función en el desarrollo de esta metodología durante nuestros años de colaboración.

Acerca del autor

John B. Rollins, Ph.D., es un científico de datos de la organización IBM Analytics. Tiene antecedentes en ingeniería, minería de datos y econometría en numerosas industrias. Posee siete patentes y es el autor de un libro de ingeniería con récord de ventas y de numerosos documentos técnicos. Cuenta con doctorados en ingeniería del petróleo y economía de Texas A&M University.

IBM Analytics
Route 100
Somers, NY 10589

Producido en los Estados Unidos de América
Junio de 2015

[1] Brachman, R. y Anand, T., "El proceso de descubrimiento de conocimiento en bases de datos," en Fayyad, U. et al., eds., Avances en descubrimiento de conocimiento y minería de datos, AAAI Press, 1996 (pp. 37-57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, "Cross Industry Standard Process for Data Mining", http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E. y Chodagam, J.: Depósito dinámico: minería de datos más sencilla, IBM Redbook SG24-7418-00 (septiembre de 2007), pp. 9-26.

[5] Gregory Piatetsky, CRISP-DM continúa siendo la metodología principal para los proyectos de análisis, minería de datos o ciencia de datos, 28 de octubre de 2014, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Por favor, recicle

IMW14828-COES-00