Introducción a la ciencia de datos. Parte 1: Datos, estructura y segmentación de la ciencia de datos

M. Tim Jones

1 de febrero de 2018

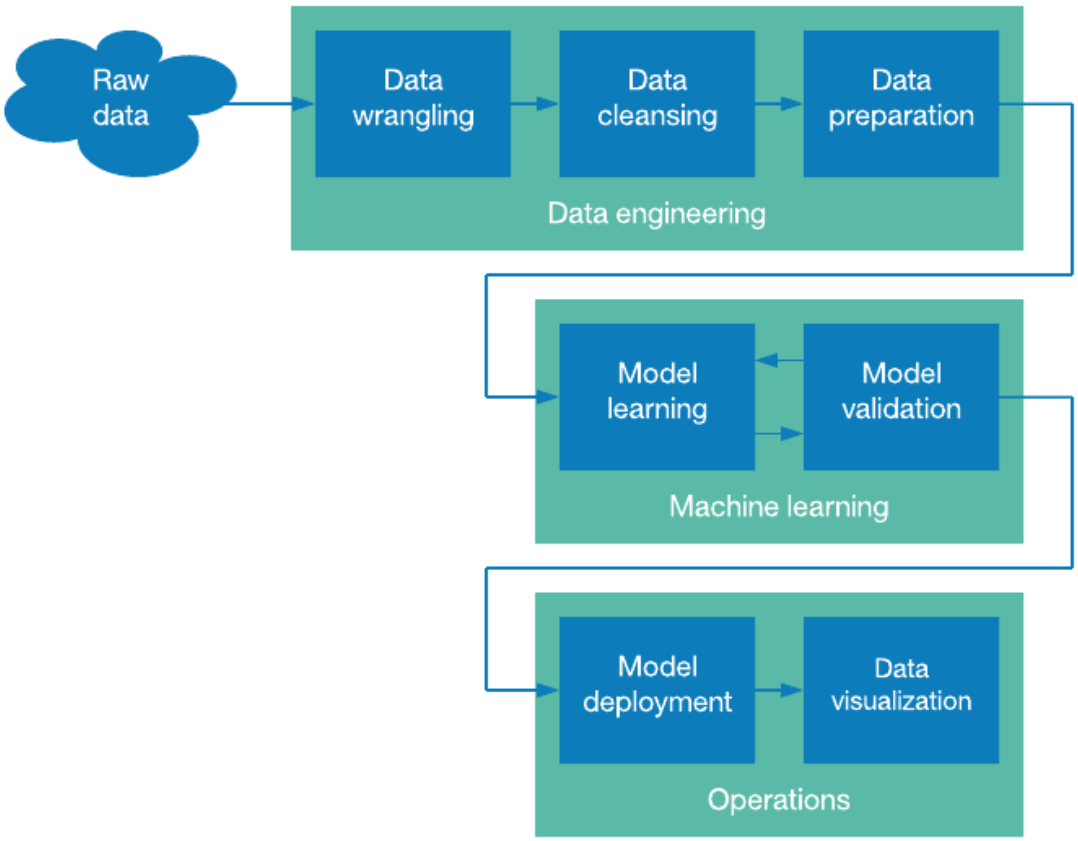
Los datos no sirven de nada si no puede procesarlos para obtener estadísticas. El campo de la ciencia de datos le brinda las herramientas y los métodos que necesita para procesar conjuntos de datos de forma eficaz y así aprovechar al máximo los datos que recopila. En este instructivo, conocerá los aspectos básicos del aprendizaje automático, incluidos la ingeniería de datos, el aprendizaje por modelos y las operaciones.

Los datos son un recurso muy valioso, pero si no cuenta con una manera de procesarlos, su valor es cuestionable. La ciencia de los datos es un campo multidisciplinario cuyo objetivo es obtener valor a partir de datos en todas sus formas. Este artículo explora el campo de la ciencia de datos a través de los datos y sus estructuras, como también el proceso de alto nivel que puede utilizar para extraer valor de los datos.

La ciencia de datos es un proceso. Esto no significa que sea un proceso mecánico y carente de creatividad. Sin embargo, si analiza las etapas del procesamiento de datos, desde la organización de las fuentes y la depuración de los datos hasta el aprendizaje automático y la posterior visualización, verá que se deben seguir pasos únicos para transformar en estadísticas los datos sin procesar.

Los pasos que siguen también pueden variar (consulte la Figura 1). En el análisis exploratorio de datos, es probable que tenga un conjunto de datos depurados que esté listo para importar a R. En este caso, puede visualizar el resultado, pero no implementar el modelo en un entorno de producción. En otro entorno, es probable que trabaje con datos reales y que, además del ajuste y la preparación de los datos, necesite un proceso de combinación y depuración de datos para poder entrenar su modelo de aprendizaje automático.

Figura 1. La segmentación de la ciencia de datos



Para comenzar, analicemos los elementos de la segmentación de la ciencia de datos a fin de comprender el proceso.

Los datos y su estructura

Los datos tienen muchas formas, pero, a gran escala, se ubican en tres categorías: estructurados, semiestructurados y no estructurados (consulte la Figura 2). Los datos estructurados son datos muy organizados que existen dentro de un repositorio, como una base de datos (o un archivo de valores separados por comas [CSV]). Los datos son de fácil acceso y su formato los vuelve apropiados para las consultas y los cómputos (mediante el uso de lenguajes como el Lenguaje de consulta estructurada [SQL] o Apache™ Hive™). Los datos no estructurados carecen de toda estructura de contenido (por ejemplo, una transmisión de audio o un texto en lenguaje natural). En el medio se encuentran los datos semiestructurados, que pueden incluir metadatos o datos que se pueden procesar con mayor facilidad que los datos no estructurados por medio del etiquetado semántico. Estos datos no están completamente estructurados porque el contenido de menor nivel probablemente aún represente datos que requieran cierto grado de procesamiento para resultar útiles.

Figura 2. Modelos de datos

Los datos estructurados son la forma de datos más útil, ya que se pueden manipular al instante. La regla general es que los datos estructurados representan únicamente el 20% de los datos totales. La mayoría de los datos del mundo (el 80% de los datos disponibles) son semiestructurados o no estructurados.

Tenga en cuenta que mucho de lo que se define como *dato no estructurado* en realidad tiene estructura (como un documento que contiene metadatos y etiquetas para el contenido), pero al contenido mismo le falta estructura y no se puede utilizar de manera inmediata. Por lo tanto, estos datos se consideran no estructurados.

Ingeniería de datos

Una encuesta de 2016 reveló que los científicos de datos dedican el 80% del tiempo a recopilar, depurar y preparar los datos para su uso en el aprendizaje automático. El 20% restante lo dedican a extraer o modelar datos por medio de algoritmos de aprendizaje automático. Aunque es la parte menos placentera del proceso, esta ingeniería de datos tiene una gran importancia y repercute en la calidad de los datos de la fase de aprendizaje automático.

Dividí la ingeniería de datos en tres partes: manejo, depuración y preparación. Dado que esta fase supone mucho trabajo, a este proceso algunos lo llaman *administración de datos*.

Manejo de datos

El *manejo de datos*, en pocas palabras, es el proceso de manipular los datos no procesados a fin de que sean útiles para el análisis de datos o para entrenar un modelo de aprendizaje automático. Esta parte de la ingeniería de datos puede incluir la obtención de los datos de uno o más conjuntos (además de la reducción del conjunto a los datos necesarios), la normalización que permite lograr la coherencia de los datos combinados a partir de varios conjuntos, y el análisis de los datos en algún tipo de estructura o almacenamiento para su posterior uso. Piense en un conjunto de datos públicos de un sitio web de datos abiertos federales. Estos datos podrían existir como un archivo de hoja de cálculo que se debe exportar en un formato más aceptable para los lenguajes de la ciencia de datos (CSV o Notación de Objetos de JavaScript). La fuente de datos también podría ser un sitio web del que una herramienta automatizada extrajo los datos. Por último, los datos podrían provenir de varias fuentes, lo que requiere que usted elija un formato común para el conjunto de datos resultante.

Este conjunto de datos resultante probablemente requeriría posprocesamiento para poder importarse a una aplicación de análisis (como el Proyecto R para la informática estadística, el Lenguaje de datos GNU

o Apache Hadoop). El *manejo de datos*, por lo tanto, es el proceso por medio del cual identifica, recopila, combina y preprocesa uno o más conjuntos de datos a fin de prepararlos para la depuración.

Depuración de datos

Una vez que haya recopilado y combinado su conjunto de datos, el siguiente paso consiste en depurarlo. Los conjuntos de datos en su estado natural suelen ser desordenados y estar repletos de errores comunes, como la escasez (o el exceso) de valores, la existencia de delimitadores erróneos o incorrectos (que segregan los datos), la inconsistencia de registros o la insuficiencia de parámetros. En algunos casos, los datos no se pueden reparar y, por ende, se deben quitar; en otros casos, se pueden corregir de forma manual o automática.

Si su conjunto de datos es sintácticamente correcto, el siguiente paso es garantizar que también lo sea semánticamente. En un conjunto de datos que contiene datos numéricos, tendrá valores atípicos que requerirán una inspección más detallada. Puede descubrir estos valores atípicos por medio del análisis estadístico. Para ello, debe observar la media, los promedios y la desviación estándar. La búsqueda de valores atípicos es un método de depuración secundario que permite garantizar que los datos sean uniformes y precisos.

Si desea obtener más información sobre la depuración de datos, consulte Cómo trabajar con datos desordenados.

Preparación de datos

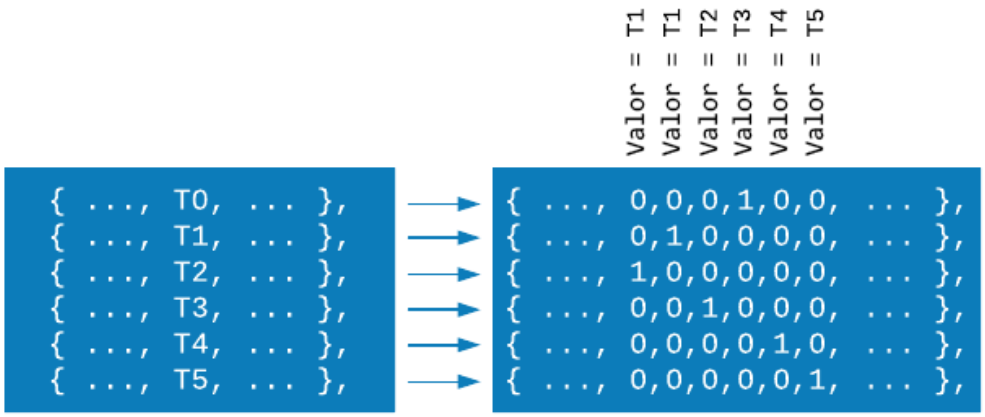
El paso final de la ingeniería de datos es la preparación (o el preprocesamiento). Este paso da por sentado que tiene un conjunto de datos depurado que podría no estar listo para que lo procese un algoritmo de aprendizaje automático. Aquí tiene algunos ejemplos en los que se puede aplicar la preparación.

En algunos casos, la normalización de datos puede resultar útil. Con la normalización, transforma una función de ingreso a fin de distribuir los datos de forma homogénea en un rango aceptable para el algoritmo de aprendizaje automático. Esta tarea puede ser tan simple como el ajuste lineal (de un rango arbitrario, dado un mínimo y un máximo de dominio de -1,0 a 1,0). También puede aplicar enfoques estadísticos más complicados. La normalización de datos puede ayudarlo a que no se quede atascado en una optimización local durante el proceso de entrenamiento (en el contexto de redes neuronales).

Otra técnica útil en la preparación de datos es la conversión de datos categóricos a valores numéricos. Piense en un conjunto de datos que incluye un grupo de símbolos que representan una función (como `{T0..T5}`). Como string, no sirve como entrada para una red neuronal, pero usted puede transformarlo por medio de un esquema one-of-*K* (también conocido como *codificación one-hot*).

En este esquema (ilustrado en la Figura 3), puede identificar la cantidad de símbolos de la función (en este caso, seis) y, luego, crear seis funciones para representar el campo original. Para cada símbolo, puede establecer una sola función, lo que da lugar a una representación adecuada de los elementos distintivos del símbolo. Como consecuencia, obtiene una mayor dimensión, pero, a su vez, proporciona un vector de función que resulta más eficaz para los algoritmos de aprendizaje automático.

Figura 3. Cómo transformar un string en un vector one-hot



Una alternativa es la codificación de enteros (mediante la cual `T0` puede ser el valor 0; T1, el valor 1, y así sucesivamente), pero este enfoque puede generar problemas de representación. Por ejemplo, en un resultado con valores reales, ¿qué representa 0,5?

Aprendizaje automático

En esta fase, debe crear y validar un modelo de aprendizaje automático. En ocasiones, el modelo de aprendizaje automático es el producto, el cual se implementa en el contexto de una aplicación para proporcionar cierta capacidad (como clasificación o predicción). En otros casos, el algoritmo de aprendizaje automático es solo un medio para un fin. En estos casos, el producto no es el algoritmo de aprendizaje automático entrenado, sino los datos que produce.

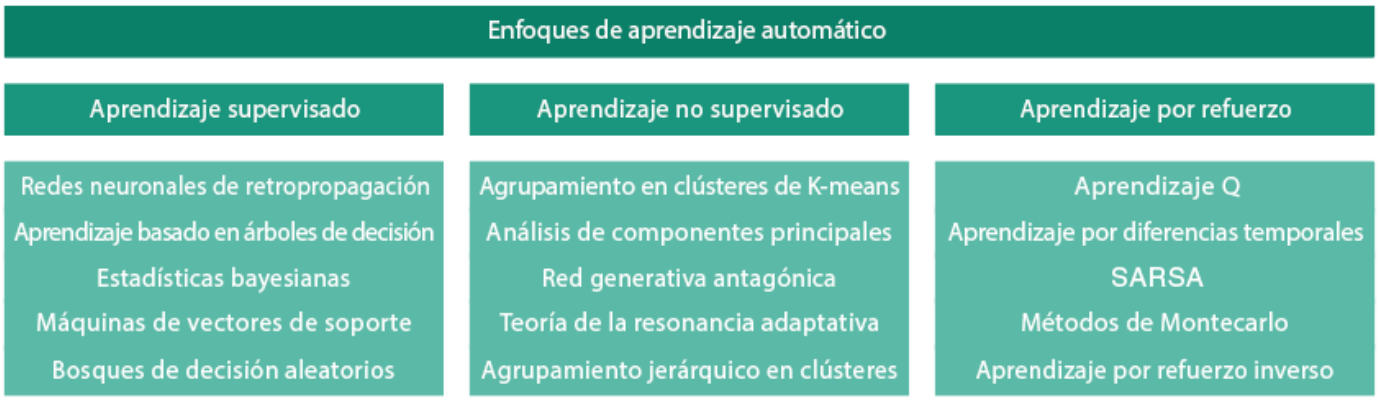
Esta sección analiza la construcción y la validación de un modelo de aprendizaje automático. Puede obtener más información sobre el aprendizaje automático a partir de datos en Cómo obtener estadísticas valiosas a partir de conjuntos de datos limpios.

Aprendizaje por modelos

El meollo de la segmentación de la ciencia de datos es el paso del procesamiento de datos. En un solo modelo, el algoritmo puede procesar los datos y generar un nuevo producto de datos. Sin embargo, en una serie de producción, el modelo de aprendizaje automático es el producto mismo, implementado para proporcionar estadísticas o agregar valor (como la implementación de una red neuronal para brindar capacidades de predicción para un mercado de seguros).

Los enfoques de aprendizaje automático son amplios y variados, tal como se ve en la Figura 4. Esta pequeña lista de algoritmos de aprendizaje automático (segregada por modelo de aprendizaje) demuestra el potencial de las capacidades que se brindan por medio del aprendizaje automático.

Figura 4. Enfoques de aprendizaje automático



El aprendizaje supervisado, como su nombre lo indica, está impulsado por un crítico que proporciona el medio para alterar el modelo en función de sus resultados. Si se tiene un conjunto de datos con una clase (por ejemplo, una variable dependiente), el algoritmo se entrena para producir la clase correcta y alterar el modelo cuando no lo logre. El modelo se entrena hasta que alcanza cierto nivel de precisión, que es el momento en que puede implementarlo para proporcionar predicción para datos no vistos.

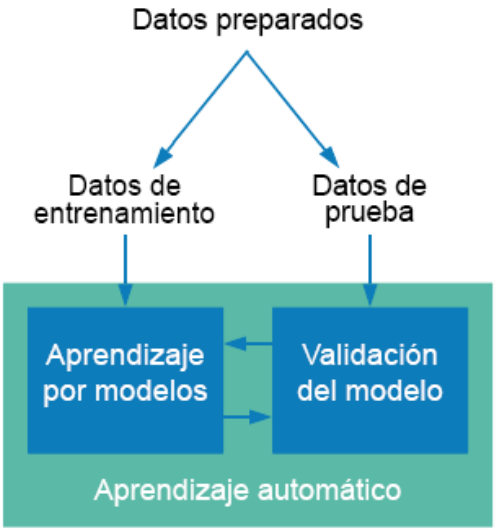
Por su parte, el aprendizaje no supervisado no tiene ninguna clase; en su lugar, inspecciona los datos y los agrupa en función de una cierta estructura que está oculta dentro de ellos. Puede aplicar estos tipos de algoritmos en los sistemas de recomendaciones. Para ello, agrupe los clientes en función del historial de visualización o compra.

Por último, el *aprendizaje por refuerzo* es un algoritmo de aprendizaje semisupervisado que ofrece un reconocimiento luego de que el modelo toma una cierta cantidad de decisiones que conducen a un resultado satisfactorio. Este tipo de modelo se utiliza para crear agentes que actúan de forma racional en cierto espacio de estado/acción (como agentes que juegan póker).

Validación del modelo

Una vez entrenado, ¿cómo se comportará un modelo en producción? Una forma de comprender su comportamiento es mediante la validación del modelo. Un enfoque habitual para la validación del modelo consiste en reservar una pequeña cantidad de los datos de entrenamiento disponibles a fin de que se prueben con respecto al modelo final (llamados *datos de prueba*). Los datos de entrenamiento se utilizan para entrenar el modelo de aprendizaje automático, mientras que los datos de prueba se usan cuando el modelo está completo con el objeto de validar la eficacia de su generalización con respecto a datos no vistos (consulte la Figura 5).

Figura 5. Datos de entrenamiento y datos de prueba para la validación del modelo



La construcción de un conjunto de datos de prueba a partir de un conjunto de datos de entrenamiento puede ser complicada. Si bien una muestra aleatoria puede funcionar, esta también puede resultar problemática. Por ejemplo, ¿la muestra aleatoria sobrerrepresenta una determinada clase o abarca bien todas las posibles clases de los datos o sus funciones? Las muestras aleatorias con distribución sobre las clases de datos pueden ser útiles para evitar el sobreajuste (es decir, el entrenamiento se ajusta demasiado a los datos de entrenamiento) o el subajuste (es decir, no se modelan los datos de entrenamiento y falta la capacidad para generalizar).

Operaciones

Operaciones hace referencia al objetivo final de la segmentación de la ciencia de datos. Este objetivo puede ser tan simple como la creación de una visualización para que su producto de datos le relate un cuento a un público determinado o responda algunas preguntas creadas antes de que el conjunto de datos se haya utilizado para entrenar un modelo. O bien, puede ser tan complejo como la implementación del modelo de aprendizaje automático en un entorno de producción para operar con datos no vistos a fin de proporcionar predicción o clasificación. Esta sección explora ambas situaciones.

Implementación del modelo

Cuando el producto de una fase de aprendizaje automático es un modelo que usted utilizará con respecto a datos futuros, el modelo se implementa en un entorno de producción a fin de aplicarlo a nuevos datos. Este modelo puede ser un sistema de predicción que tome como ingreso los datos financieros históricos (como las ventas y los ingresos mensuales) y determine si una empresa es un objetivo de adquisición razonable.

En situaciones como estas, el modelo implementado suele dejar de aprender y simplemente se le aplican datos para que realice una predicción. Existen buenos motivos para evitar el aprendizaje en producción. En el contexto del aprendizaje profundo (redes neuronales con capas profundas), se han identificado ataques adversarios que pueden alterar los resultados de una red. En una red de aprendizaje profundo para el procesamiento de imágenes, por ejemplo, la aplicación de una

imagen con una perturbación puede alterar las capacidades de predicción de la imagen de modo que en vez de “ver” un tanque, la red de aprendizaje profundo ve un “automóvil”. Los ataques adversarios crecieron con la aplicación del aprendizaje profundo, por lo que hay nuevos vectores de ataque que son objeto de investigaciones activas.

Visualización del modelo

En la ciencia de datos a menor escala, el producto buscado son los datos y no necesariamente el modelo producido en la fase de aprendizaje automático. Esta situación es la forma de operación más habitual en la segmentación de la ciencia de datos. Aquí, el modelo proporciona el medio para generar un producto de datos que responde algunas preguntas acerca del conjunto de datos original. Las opciones de visualización son amplias y se pueden producir a partir del lenguaje de programación R, gnuplot y D3.js (que puede producir argumentos interactivos muy atractivos).

Puede obtener más información sobre la visualización en el siguiente artículo de esta serie.

A continuación

Este artículo exploró una segmentación de datos genéricos para el aprendizaje automático y abarcó la ingeniería de datos, el aprendizaje por modelos y las operaciones. El próximo artículo de esta serie explorará dos modelos de aprendizaje automático para la predicción que utilizan conjuntos de datos públicos.

Temas relacionados

- Aspectos básicos de la ciencia de datos
- Introducción a R
- Aprendizaje automático con R
- Cómo identificar datos personales a partir de un texto no estructurado
- Realice un ejercicio de aprendizaje automático