Fundamentos para al Análisis de Datos y la Investigación

Tema 1 - La Ciencia de Datos

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

1 Introducción

Fundamentos para el Análisis de datos y la Investigación

- El objetivo principal de la asignatura es unificar los conocimientos básicos necesarios para el Máster
- En esta asignatura estudiaremos: qué es la ciencia de datos, análisis exploratorio de datos, fundamentos de cálculo, álgebra, probabilidad y estadística
- Todos los conceptos serán vistos haciendo uso del software R

1 Introducción

Contenidos:

- Introducción a la Ciencia de Datos
- Etapas de un proyecto de Ciencia de Datos
- Interpretación de Resultados. Comunicación
- Recursos

Contents

2 Ciencia de Datos

Definition

Ciencia de Datos (CD): Ciencia que se ocupa del estudio de los datos por medio del **método científico**. La CD combina las técnicas y metodologías de la estadística, las matematicas y la computación, con el objetivo de entender, interpretar e inferir conclusiones a partir de los datos y tomar decisiones. La *predicción* y la *clasificación* son dos aspectos importantes, pero no los únicos, dentro de la CD.

2 Disciplinas relacionadas

Existen una serie de disciplinas relacionadas con la CD:

- *Data mining*: metodologías que se ocupan de la recolección, almacenamiento y tratamiento de los datos, tras convertirla en información útil
- *Machine learning* (aprendizaje automático): metodologías que permiten que los métodos y modelos matemáticos basados en datos funcionen y mejoren de forma autónoma, corriguiendo los errores, valorando los métodos y obtenidos predicciones y clasificaciones
- *Big data*: ciencias que estudia cantidades grandes de datos con objeto de obtener informaciones útiles
- *Inteligencia artificial*: combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano

2 Etapas de un proyecto de CD

Podemos establecer las siguientes etapas:

- *Conocer en detalle los objetivos del proyecto*. Esta primera etapa incluye el reunirse con el cliente (directivos y expertos de la empresa) y discutir y conocer en detalle los objetivos el proyecto.
- *Obtener los datos y la información disponible*. Puede ser proporcionada por el cliente o bien por medio de encuestas, etc. Llegar a una comprensión clara de toda la información.
- *Establecer* las hipótesis a contrastar, modelos de trabajo, posibles teorías y metodologías para usar.
- *Estimar y validar* los modelos previamente establecidos. *Contraster* las hipótesis formuladas.
- *Obtener conclusiones previas*.
- *Discutir las conclusiones previas con el cliente y reconsiderar* los objetivos si es necesario.
- *Conclusiones definitivas y toma de decisiones*.

2 Metodología CRISP-DM

- El método CRISP-DM fue desarrollado por IBM (que incluye *IBM SPSS Statistics* y otros productos) y consiste en un conjunto de etapas, de modo que en cada etapa se puede volver a la anterior y mejorar el proceso. El procedimiento tiene en cuenta tanto la compresión y las necesidades del cliente y del negocio como la parte científica del análisis de datos.
- Las etapas son (ver Figura):
  - Comprensión del negocio
  - Comprensión de los datos
  - Preparación de los datos
  - Modelización
  - Evaluación
  - Plan de actuación

2 Metodología CRISP-DM

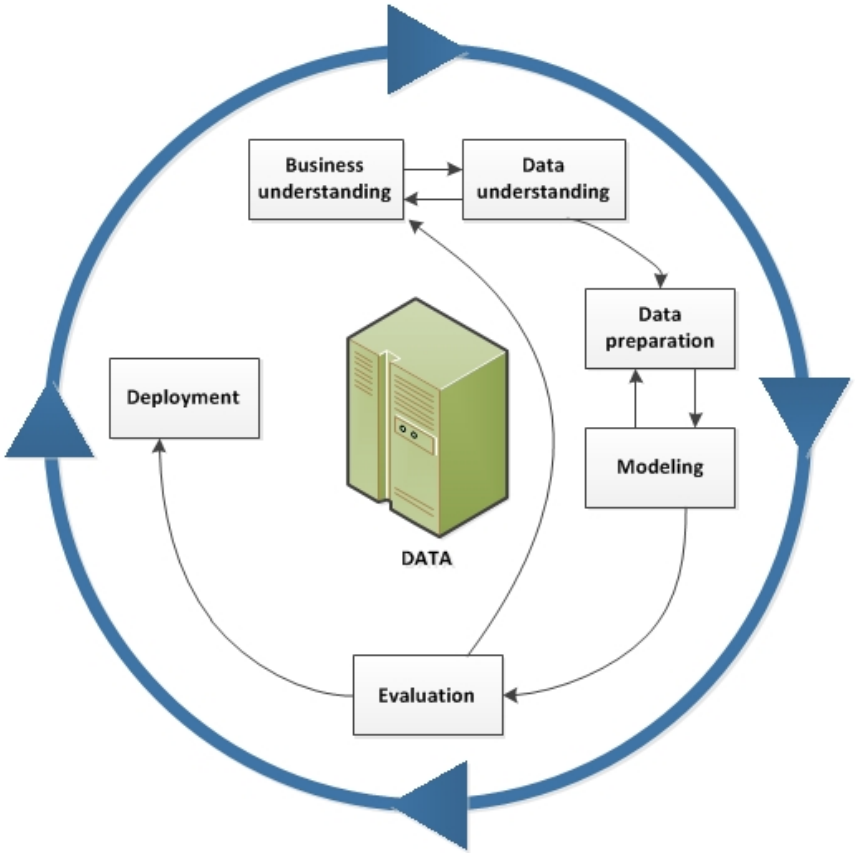

Figura: Metodología CRISP-DM. Fuente: www.ibm.com

2 Marco Metodológico. Datos y modelos

- En CD se hace uso de datos y modelos. Los datos son la materia prima del proceso de investigación.
- Los modelos constituyen representaciones de la realidad construidos mediante fórmulas matemáticas, con objeto de representar las relaciones empíricas observadas en los datos. Los modelos incluyen parámetros, restricciones etc.
- La metodología básica que usa el científico de datos, hace uso de dos disciplinas fundamentales:
  - La ciencia estadística
  - Las ciencias de la computación

2 Marco Metodológico. Datos y modelos

- El tipo de problemas que habitualmente resuelve el científico de datos tienen que ver con:
  - *Predicción*
  - *Clasificación*
- Los procedimientos de aprendizaje automático (*machine learning*) permiten elegir la mejor técnica en cada caso, usando los criterios estadísticos adecuados.
- Los métodos anteriores, junto con otros de CD, se pueden aplicar a multitud de problemas.

Contents

3 Interpretación y Comunicación

- Hoy en día, la interpretación y comunicación es un aspecto clave para el científico de datos
- La interpretación y comunicación suponen: *ser capaces de comunicar al cliente los resultados sin recurrir a tecnicismos, haciendo uso de la metodología/s más adecuada/s*
- Comunicación: ser capaces de explicar el problema y la solución en una conversación informal
- En un problema de clasificación, los *árboles de decisión* es un claro ejemplo, frente a otras de las técnicas de clasificación

Contents

4 Recursos

El número de recursos en data science ha crecido exponencialmente:

- Lenguajes de programación en data science: R, Python, SQL, Julia, etc
- `https://www.kaggle.com/`
  Empresa subsidiaria de Google LLC. Es una comunidad en línea de científicos de datos y profesionales del machine learning
- `https://github.com/`
  GitHub es una website para alojar proyectos utilizando el sistema de control de versiones Git. Se utiliza principalmente para la creación de código fuente de programas de ordenador. El software que opera GitHub fue escrito en Ruby on Rails. Desde enero de 2010, GitHub opera bajo el nombre de GitHub, Inc.
- Repositorios de datos. Por ejemplo:
  `https://archive.ics.uci.edu/`
- Más repositorios de datos:
  `https://dextutor.com/top-10-dataset-repositories/`

Fundamentos para al Análisis de Datos y la Investigación

Tema 1 - La Ciencia de Datos

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad