Minería de Datos

Isabel Cenamor

viu Universidad Internacional de Valencia

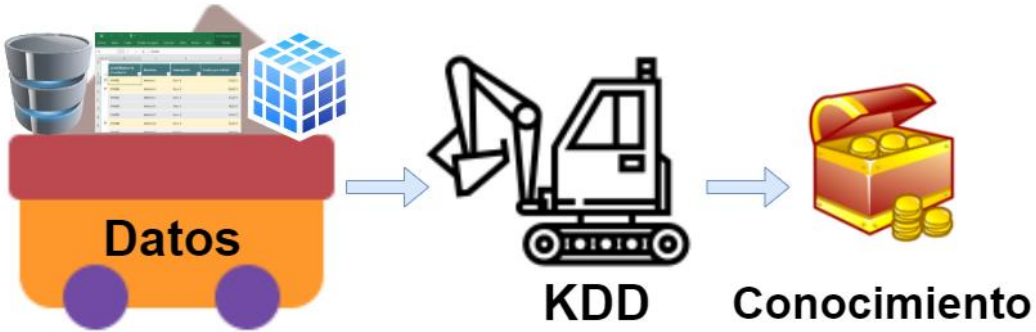
Noviembre 2022

De:
Planeta Formación y Universidades

Guia de resolución de actividades – el proceso KDD

¿Qué es el proceso KDD?

- El acrónimo KDD hace referencia a un proceso, compuesto de múltiples etapas, cuyo objetivo principal es **la extracción de conocimiento**, que ha de **resultar útil y no ser trivial,** a partir de los datos a los que se tiene acceso.



¿Qué necesitas saber?

- Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados
- KDD se puede aplicar en diferentes dominios:
  - determinar perfiles de clientes fraudulentos (evasión de impuestos)
  - descubrir relaciones implícitas existentes entre síntomas y enfermedades, entre características técnicas y diagnóstico del estado de equipos y máquinas
  - determinar perfiles de estudiantes “académicamente exitosos” en términos de sus características socioeconómicas
  - determinar patrones de compra de los clientes en sus canastas de mercado

¿Qué necesitas saber?

- ¿Qué problema quieres resolver?
- ¿Cómo puedes ayudar a tu empresa a generar conocimiento?
- ¿Qué proceso manual se esta realizando que se podría realizar de manera automática a través de los datos?

Una vez definido el objetivo

- ¿Dónde saco los datos?
- ¿Es suficiente con una fuente de datos?
- Analizar si los datos iniciales se encuentran procesados de alguna manera

Tipos de Datasets

- **Archivo**: es un fichero independiente en el que se almacena toda la información con la que se va a trabajar. Tiene como ventajas, la seguridad y rapidez para el trabajo con los datos, ya que siempre se explotan y se visualizan de manera local, sin embargo la escalabilidad y conexión con otros Datasets que no están almacenados en la misma máquina se dificulta.
- **Folder**: es la suma de diferentes Datasets almacenados en una misma carpeta, los cuales están conectados entre ellos. Estos archivos deben compartir un mismo formato como puede ser .csv, .mif o dxf.
- **Bases de datos**: este tipo de Dataset puede llegarse a confundir con el archivo, pero se diferencia por su nivel de especialidad, es decir, son bases de datos con formatos específicos diseñadas para programas puntuales. Por ejemplo las bases de datos de Oracle, las cuales solo funcionan para sus desarrollos.
- **Web**: es la compilación de datos que se almacenan dentro de un sitio web. El nombre que se le asigna por defecto a este Dataset es el correspondiente a la URL.

¿Qué es un dataset?

- Un conjunto de datos o **dataset**, es un conjunto de datos, que habitualmente están estructurados. Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.

- Una **instancia, ejemplo o registro (instance, sample, record)** es cada uno de los datos de los que se disponen para hacer un análisis. Si se quiere predecir el comportamiento de los clientes de un servicio de telefonía, cada instancia correspondería a un abonado. Cada instancia, a su vez, está compuesta de **características** que la describen, como la antigüedad del cliente en la compañía, el gasto diario en llamadas, etc. En una hoja de cálculo, las instancias serían las filas; las características, las columnas.

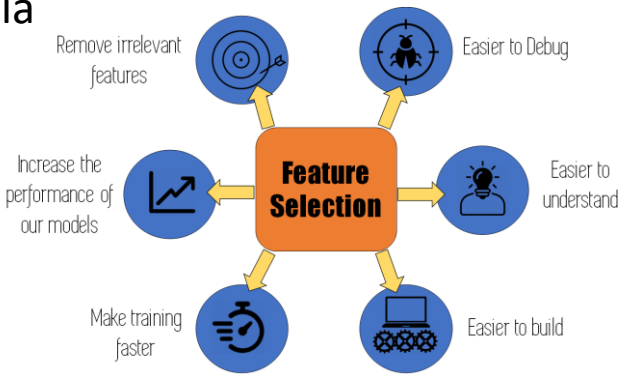
¿Qué es un dataset?

- La **característica, atributo, factor, propiedad o campo** (*feature, attribute, property, field*); son los atributos que describen cada una de las instancias del conjunto de datos. Las denominaciones se usan indistintamente en función del autor y del contexto. En el caso de una cartera de clientes, estaríamos hablando del número de compras de cada cliente, antigüedad, si es seguidor en redes sociales, si se ha dado de alta en la newsletter, qué productos comprados… En una hoja de cálculo, serían las columnas.

- El **objetivo** (*objective*) es el *atributo o factor* que queremos predecir, el objetivo de la predicción, como puede ser la probabilidad de reingreso de un paciente tras una intervención quirúrgica. También puede denominarse atributo de salida y solo existe en aprendizaje supervisado.

Revisar el / los datasets encontrados me ayudan a resolver el objetivo planteado

- ¿Cuántos registros hay?
  - ¿Están todas las filas completas o tenemos campos con valores nulos?
  - ¿Son demasiado pocos?
- ¿Qué tipos de datos tenemos?
- ¿Vas a trabajar un problema supervisado? -> Identifica la clase
- Una vez definido el objetivo
- ¿Qué características son más importantes?
- ¿Es suficiente con una fuente de datos?



Las variables estadísticas pueden ser de dos tipos:

•**Cualitativas**: son aquellas en la que los resultados posibles no son valores numéricos. Por ejemplo: color del pelo, tipo de ropa preferida, lugar de veraneo, etc.

•**Cuantitativas**: aquellas cuyo resultado es un número. A su vez, las hay de dos tipos:

- **Cuantitativas discretas**: cuando se toman valores aislados. Por ejemplo: número de amigos de tu pandilla, número de veces que vas al cine al mes, número de coches que tiene tu familia.
- **Cuantitativas continuas**: cuando, entre dos valores cualesquiera, puede haber valores intermedios. Es decir, se toman todos los valores de un determinado intervalo. Por ejemplo: peso de las personas, nivel sobre el mar en que se encuentra tu ciudad, medida del perímetro torácico.

Identificando fuente de datos potenciales

https://data.nasdaq.com/

https://opendata.socrata.com/

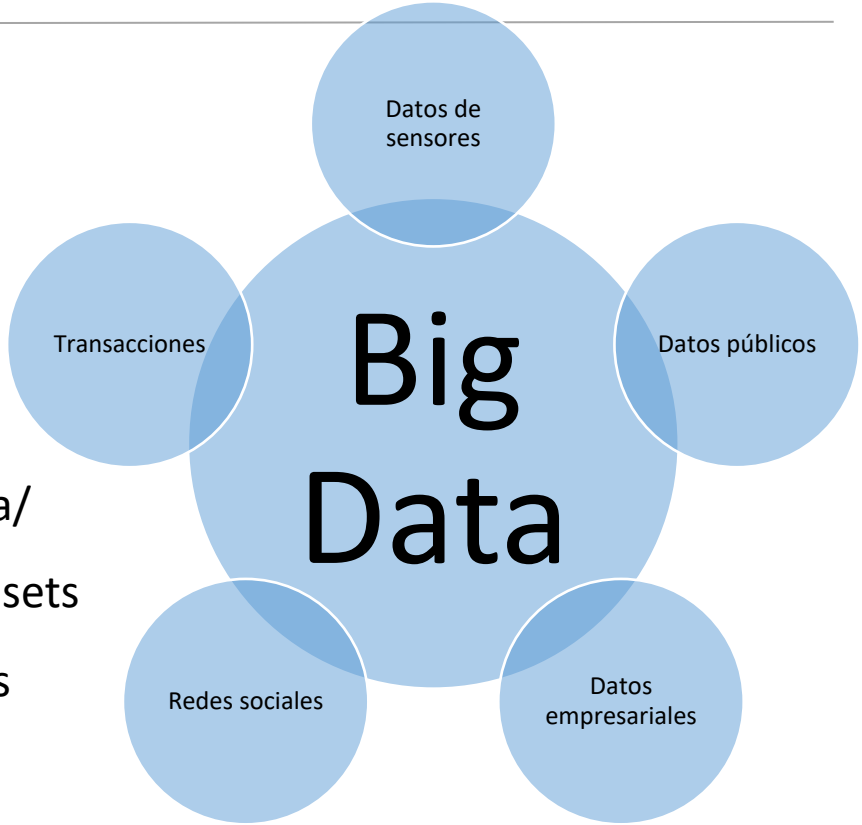
https://archive-beta.ics.uci.edu/

https://cloud.google.com/bigquery/public-data/

https://github.com/datagovsg/datagovsg-datasets

https://www.ncdc.noaa.gov/cdo-web/datasets

https://www.who.int/data/gho/gho-search



La importancia de los datos

- https://data.fivethirtyeight.com/: podrás encontrar Datasets enfocados en datos actuales de deporte, política y encuestas a nivel mundial.
- https://www.vizforsocialgood.com/: con información enfocada en el cambio social. Con temas como niños desplazados, refugiados o emprendimiento de personas discriminadas, estos Datasets ayudan a ver problemáticas sociales.
- Twitter: esta red social tiene una API para obtener datos de hashtags, tendencias y cuentas. Esta API se puede conectar con Tableau para visualizar lo que queramos. https://tableaujunkie.com/post/119681578798/creating-a-twitter-web-data-connector
- https://datasetsearch.research.google.com/: es quizás el buscador online. Están indexadas casi todas las fuentes de datos disponibles de manera pública y las webs que tienen su información bajo el marcado de datos estructurados schema.org

- Datos cuantitativas (mínimo 5-7) :
  - Discretas
  - Continuas
- Fecha:
  - Hora y minutos
  - Días, meses y años
- String:
  - Cadenas de texto sin patrón común
  - Enumerado: texto que se repite como una categoría (mínimo 2-3)
- Blob:
  - Imagines
  - Audio
  - Video

Data set Calidad del Aire: Contiene las respuestas de un dispositivo multisensor de gases desplegado sobre el terreno en una ciudad italiana. Se registran las medias de las respuestas horarias junto con las referencias de las concentraciones de gas de un analizador certificado.

Fuente: Saverio De Vito (saverio.devito '@' enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development

Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
3/10/2004	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
3/10/2004	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
3/10/2004	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
3/10/2004	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
3/10/2004	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
3/10/2004	23:00:00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
3/11/2004	0:00:00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603
3/11/2004	1:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0	0.7702
3/11/2004	2:00:00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7	0.7648
3/11/2004	3:00:00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2	0.7517
3/11/2004	4:00:00	-200	1011	14	1.3	527	21	1818	34	1197	445	10.1	60.5	0.7465
3/11/2004	5:00:00	0.7	1066	8	1.1	512	16	1918	28	1182	422	11.0	56.2	0.7366
3/11/2004	6:00:00	0.7	1052	16	1.6	553	34	1738	48	1221	472	10.5	58.1	0.7353
3/11/2004	7:00:00	1.1	1144	29	3.2	667	98	1490	82	1339	730	10.2	59.6	0.7417

https://archive.ics.uci.edu/ml/datasets/Air+quality

Ejemplo Data Set

Data set Calidad del Aire: Contiene las respuestas de un dispositivo multisensor de gases desplegado sobre el terreno en una ciudad italiana. Se registran las medias de las respuestas horarias junto con las referencias de las concentraciones de gas de un analizador certificado.

0 Fecha (DD/MM/AAAA)
1 Hora (HH.MM.SS)
2 Concentración media horaria real de CO en mg/m^3 (analizador de referencia)
3 Respuesta media horaria del sensor PT08.S1 (óxido de estaño) (nominalmente dirigido a CO)
4 Promedio horario real de la concentración global de hidrocarburos no metánicos en microg/m^3 (analizador de referencia)
5 Concentración de benceno de media horaria real en microg/m^3 (analizador de referencia)
6 Respuesta media horaria del sensor PT08.S2 (titanio) (objetivo nominal de NMHC)
7 Concentración de NOx de media horaria real en ppb (analizador de referencia)
8 PT08.S3 (óxido de tungsteno) respuesta media horaria del sensor (objetivo nominal de NOx)
9 Concentración de NO2 real media horaria en microg/m^3 (analizador de referencia)
10 PT08.S4 (óxido de tungsteno) respuesta media horaria del sensor (objetivo nominal de NO2)
11 PT08.S5 (óxido de indio) respuesta media horaria del sensor (objetivo nominal de O3)
12 Temperatura en Â°C
13 Humedad relativa (%)
14 Humedad absoluta AH

¿Qué es un dataset?

- Un conjunto de datos o **dataset**, es un conjunto de datos, que habitualmente están estructurados. Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.

- Una **instancia, ejemplo o registro (instance, sample, record)** es cada uno de los datos de los que se disponen para hacer un análisis. Si se quiere predecir el comportamiento de los clientes de un servicio de telefonía, cada instancia correspondería a un abonado. Cada instancia, a su vez, está compuesta de **características** que la describen, como la antigüedad del cliente en la compañía, el gasto diario en llamadas, etc. En una hoja de cálculo, las instancias serían las filas; las características, las columnas.

¿Qué se necesita?

1. Portada
2. Índice
3. Introducción y objetivo: descripción del problema a tratar, incluyendo las hipótesis que vosotros queréis plantear.
    1. ¿Qué problema queréis resolver?
    2. ¿Cómo os puede ayudar la minería de datos en este proceso?
    3. ¿Cuál es vuestra motivación?
    4. Definición del objetivo
4. Fuente/s de datos: donde puedes encontrar datos acorde a tu problema y la descripción de los mismos. Se debe incluir por cada atributo o conjunto de atributos: tipo de datos, rango si se conoce, descripción semántica del mismo y si es atributo de entrada o de salida.

¿Qué se necesita?

4. a Selección de datos: si de lo datos originales has filtrado alguno de ellos. Este proceso se hace porque TÚ crees que no aportan para el objetivo descrito en el apartado 3.
5. Limpieza de los datos: descripción de problemas en las variables que tiene tu dataset y cómo pretendes resolverlas.
6. Transformación de los datos: descripción de las transformaciones consideradas en función de tus datos y el objetivo que habéis propuesto.
7. Data set final: una comparación de cuantos datos te han quedado después de la fase 4ª, 5, 6. Incluyendo número total de instancias y atributos finales de entrada y salida.
8. Evaluación: plantear como queréis evaluar el modelo y como los hipotéticos resultados pueden conseguir o no los objetivos planteados.
9. Conclusiones: si ha sacado algo en claro de la propuesta

¿Qué se necesita?

4. a Selección de datos: si de lo datos originales has filtrado alguno de ellos. Este proceso se hace porque TÚ crees que no aportan para el objetivo descrito en el apartado 3.
5. Limpieza de los datos: descripción de problemas en las variables que tiene tu dataset y cómo pretendes resolverlas.
6. Transformación de los datos: descripción de las transformaciones consideradas en función de tus datos y el objetivo que habéis propuesto.
7. Data set final: una comparación de cuantos datos te han quedado después de la fase 4ª, 5, 6. Incluyendo número total de instancias y atributos finales de entrada y salida.
8. Evaluación: plantear como queréis evaluar el modelo y como los hipotéticos resultados pueden conseguir o no los objetivos planteados.
9. Conclusiones: si ha sacado algo en claro de la propuesta

Evaluación

	Suspenso (< 5)	Aprobado (>= 5)	Sobresaliente (>= 9)
Estilo (30%)	Redacción incoherente gramaticalmente y/o con más de 3 faltas ortográficas.	Redacción coherente, estilo informal o alguna falta ortográfica.	Redacción impecable, con estructura definida, estilo formal y sin faltas ortográficas.
Contenido (40%)	No se realizan todos los pasos del ciclo del proceso KDD, la visión y simplista. Solo se describe teóricamente el proceso KDD y no realiza ningún análisis especifico de los datos.	Se realizan todos los pasos del proceso KDD; pero se queda en ámbito superficial.	Se analiza en profundidad cada uno de los pasos y no se utilizan técnicas para ahorrar trabajo de limpieza de datos. Demuestra que sabe que pasos debe aplicar a su conjunto de datos. El alumno demuestra conocimiento de los datos y sabe cómo tratarlos.
Originalidad y pasos adicionales (30%)	El alumno se ha centrado en comentar o aplicar el número mínimo de pasos y no explica porque los ha realizado.	Ha realizado todos los pasos para conseguir el objetivo del proceso KDD y explica el motivo de los pasos intermedios.	Ha realizado todos los pasos para conseguir el objetivo del proceso KDD y explica el motivo de los pasos intermedios. Analiza y lanza hipótesis sobre lo que puede encontrar y que ciclos adicionales realizar. Crea variables adicionales, investiga y propone técnicas especificas para datos completos.

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades