Minería de datos

Unidad 2. El proceso KDD

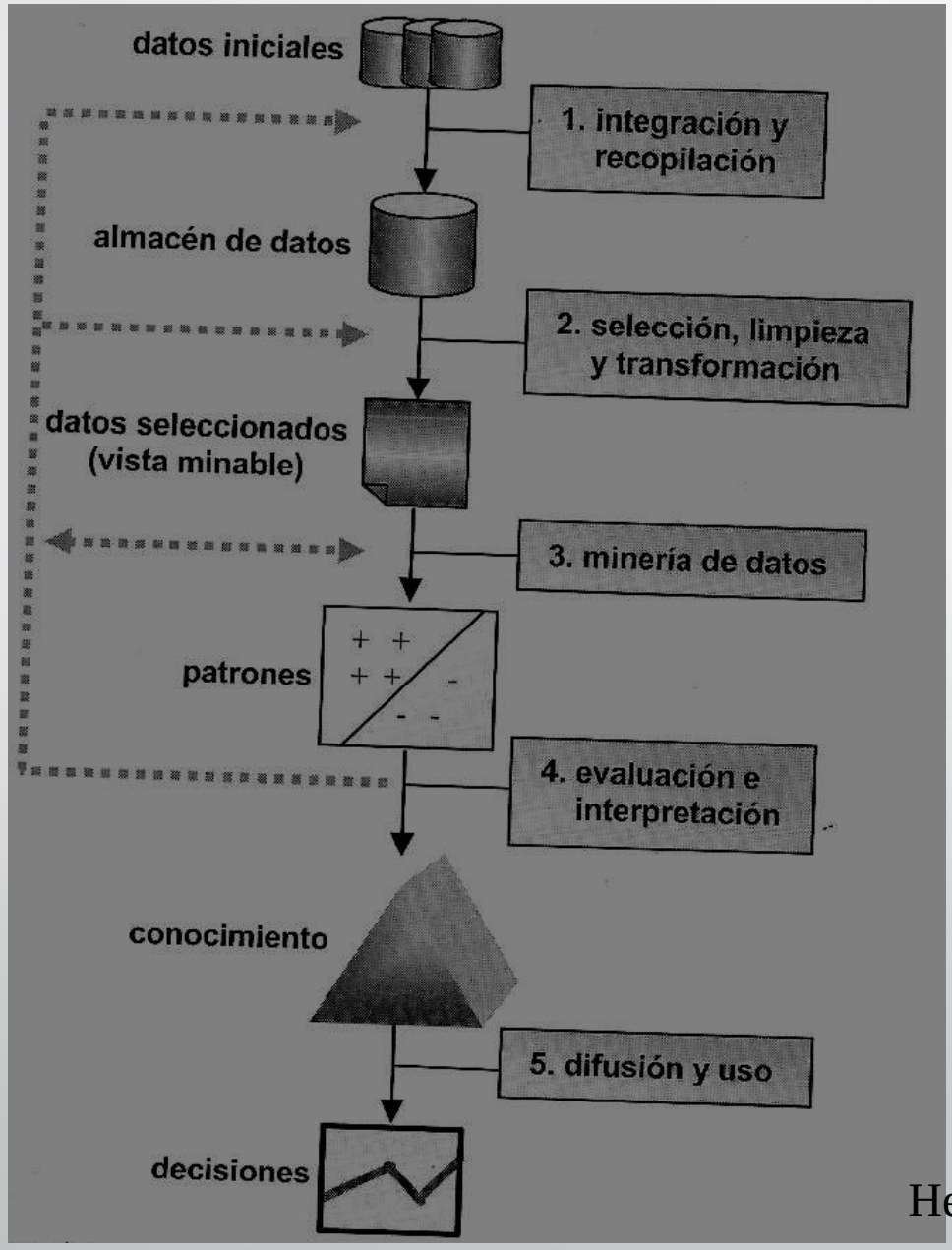
M en I Sara Vera Noguez

2. La minería de datos en el proceso de KDD

- Etapas de proceso de KDD:

1) Integración y recopilación
2) Selección, Limpieza (también llamada preprocesamiento),Transformación
3) **Minería de Datos**
4) Evaluación e Interpretación
5) Difusión y uso.

El proceso KDD



Hernández, 2008 p.20

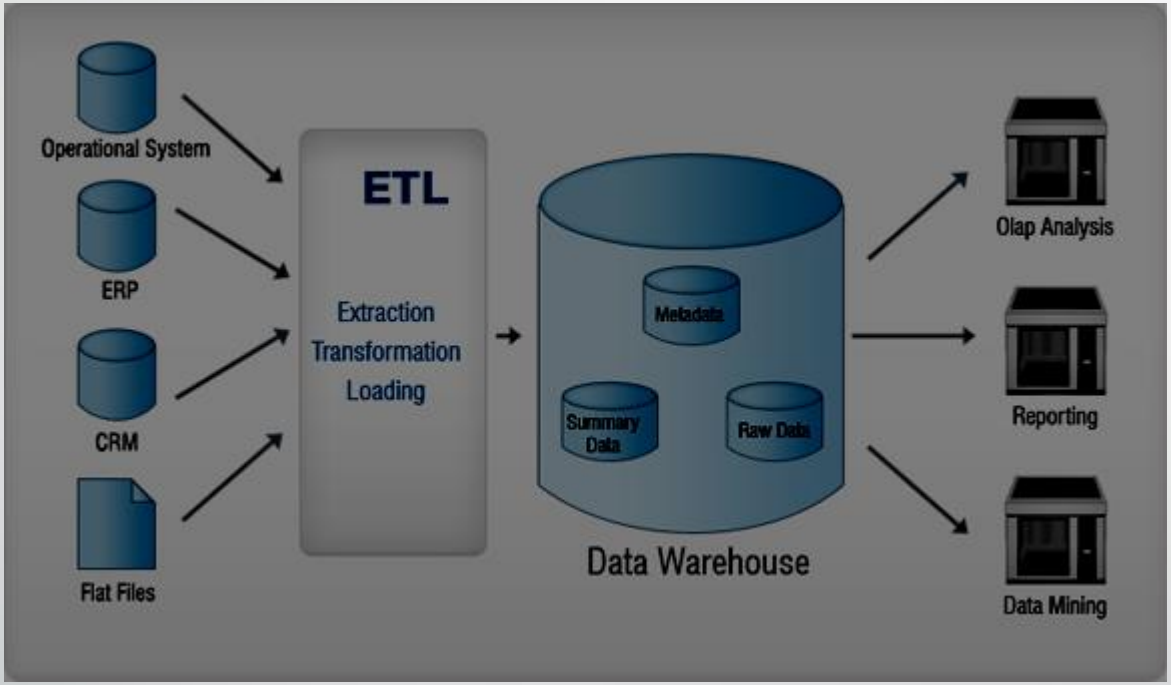
Integración y recopilación

- Generalmente la información proviene de BD OLTP (On-Line Transaction Processing)
- Pero en ocasiones es necesario extraer información desde bases de datos publicas (censos, datos del clima, etc) o base de datos privadas (como datos de otras compañías, sin infringir la ley)

Integración y recopilación

- Se deben identificar los datos necesarios, sus fuentes e integrarlos.
- La integración conlleva solucionar conflictos de tipos de datos, niveles de agregación, llaves primarias y foráneas, codificación, etc.
- Esta integración da lugar a data *warehouse* (almacenes de datos)

Integración y recopilación



https://visualrsoftware.com/data-warehouse-a-primary-introduction/

Algunos procedimientos para la integración

- Hacer una copia de las BD integrantes eliminando inconsistencias. Esto da lugar a varias BD, no se tiene un esquema integrado
- Aplicar Data *warehouse* (almacén de datos). Implica agregar y cruzar información, generando una DB *multidimensional* generando un *OLAP*

OLAP

- Permiten el análisis multidimensional de los datos
- Utiliza conocimiento previo para permitir su representación a diferentes niveles de abstracción
- Permite obtener reportes sofisticado en tiempo real de información multidimensional para comprobar o rechazar patrones hipotéticos.

OLAP vs minería de datos

- OLAP se asocia con un proceso deductivo.
- Minería de datos es un proceso inductivo
- Son herramientas complementarias
- OLAP al inicio de KDD
- Minería más adelante en KDD

Fase de selección, limpieza y transformación

- La calidad del insumo influye en la calidad del producto

Selección, limpieza y transformación

- Ya que el resultado no solo depende del algoritmo de minería, sino también de los datos, resulta importante la calidad de los datos.
- Para ello después de integrar los datos se deben seleccionar, limpiar y transformar, para generar el conjunto o vista *minable*

Selección, limpieza y transformación

- Se deben identificar los datos necesario y los irrelevantes.
- Identificar y eliminar* datos anómalos fuera de lo general (*outliers*)
- Si bien algunos algoritmos omiten los *outliers*, otros se ven fuertemente afectados por ellos.
- En ciertos casos no deben eliminarse p.e. en detección de fraudes

Selección, limpieza y transformación

- Se deben distinguir los datos necesario y los irrelevantes.
- Identificar y eliminar* datos anómalos fuera de lo general (*outliers*)
- Si bien algunos algoritmos omiten los *outliers*, otros se ven fuertemente afectados por ellos.
- En ciertos casos no deben eliminarse p.e. en detección de fraudes

Selección, limpieza y transformación

- Identificar si se tienen datos faltantes o perdidos (*missing values*)
- Identificar la causa de la ausencia de estos datos

Selección, limpieza y transformación

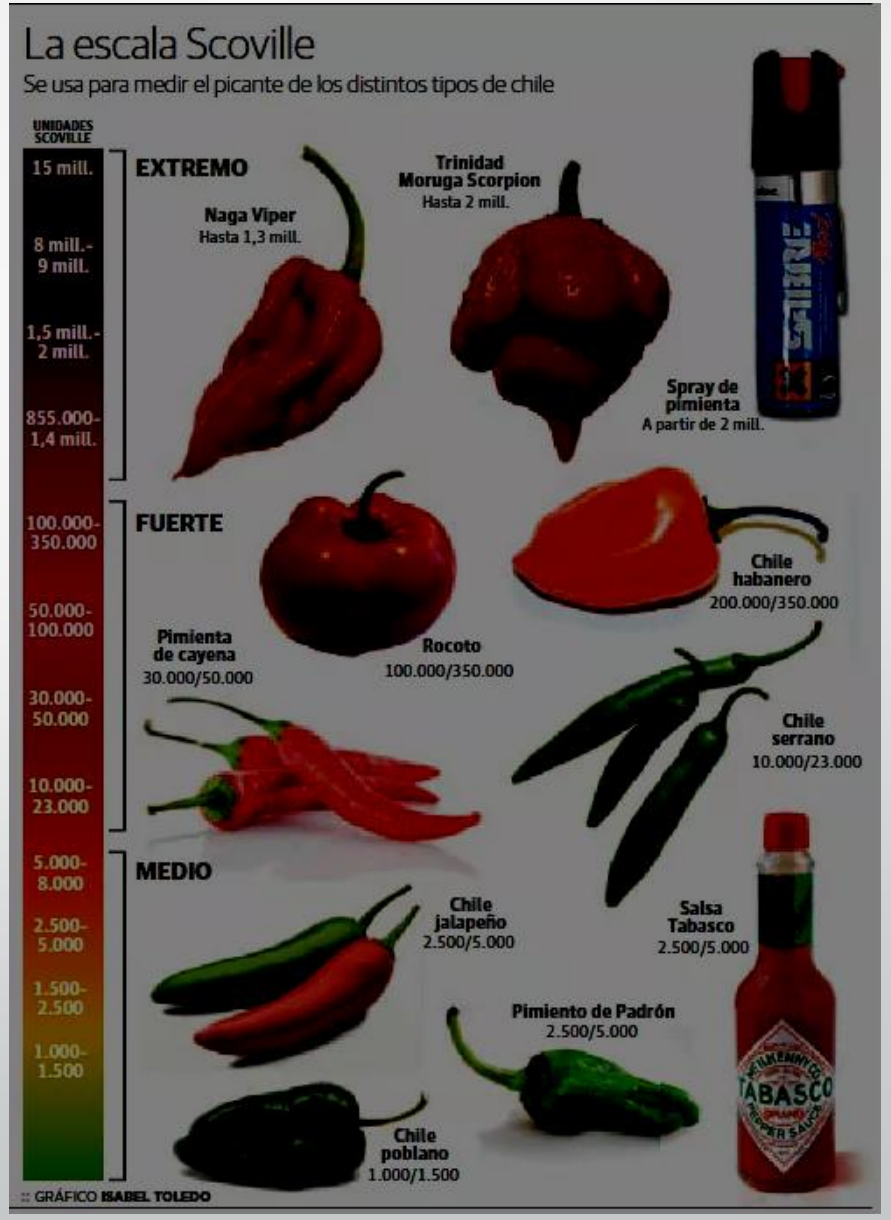
- Selección de atributos relevantes (columnas)
- Seleccionar una muestra aleatoria (renglones)

Selección, limpieza y transformación

- Construcción de atributos: que pueden ser necesario y se pueden generar a partir de otros datos, ya se mediante operaciones sobre un atributo o un conjunto de atributos, o bien mediante el cambio de tipo de un atributo
- Numerizar atributos categóricos (cambiar de tipo)
- Discretizar atributos continuos

Selección, limpieza y transformación

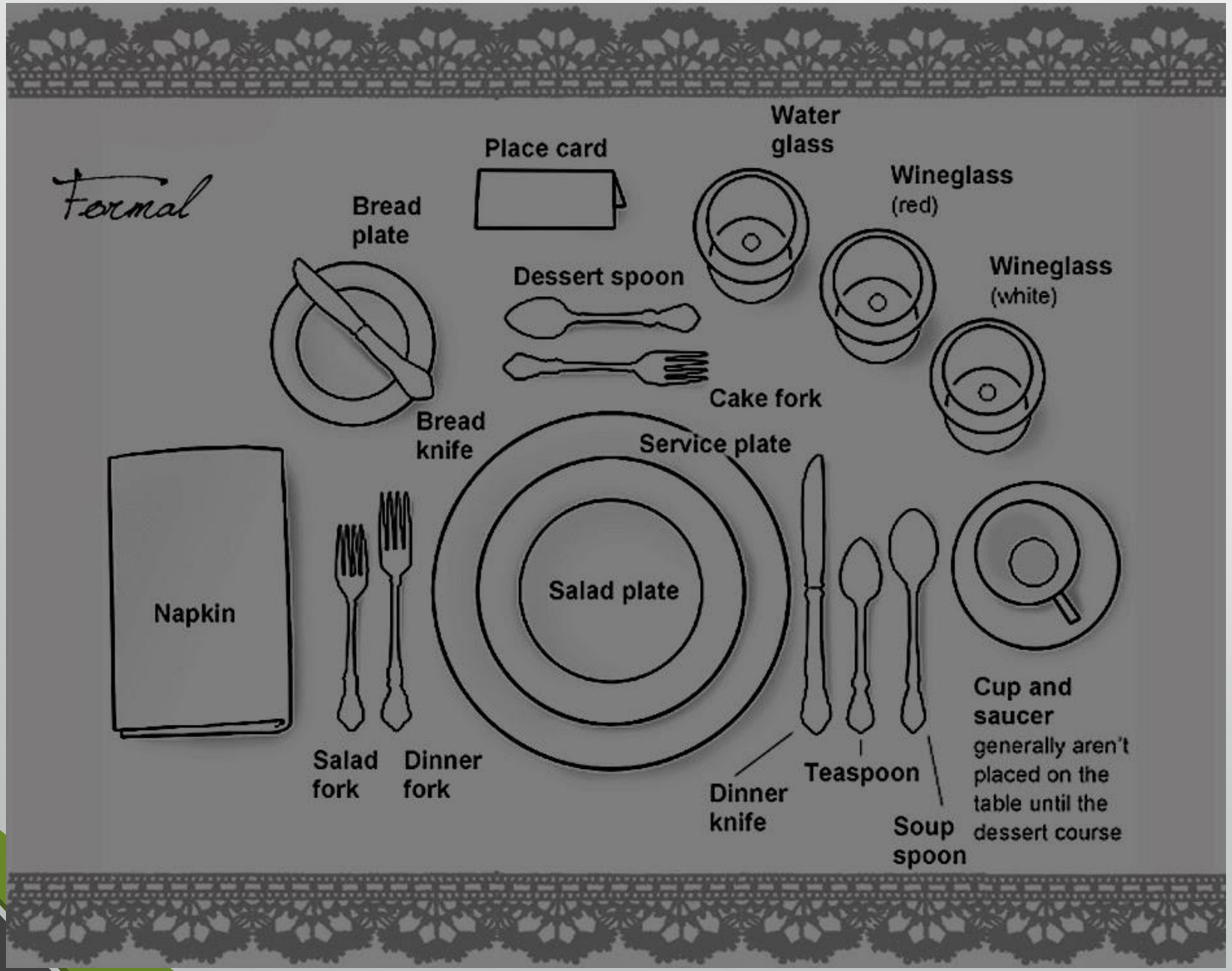
- Numerizar y Discretizar



Fase de minería de datos

- Su objetivo es extraer conocimiento para el usuarios a partir de los datos
- Implica la realización de un modelo basado en los datos, para obtener las respuestas buscadas

Fase de minería de datos



Aspectos importantes

- Determinar el tipo de tarea de minería más adecuada
- Elegir el tipo de modelo mas convenientes
- Elegir el algoritmo

Tareas de minería de datos

- Se relacionan con los problemas que se busca resolver.
- Cada tarea se desprende de requisitos específicos
- Cada tarea se asocia con ciertos tipos de datos

Tareas de minería de datos

- **Clasificación**: se ve reflejado como un valor discreto a un atributo, que determina la clase de la instancia. Los atributos usados para la clasificación reciben el nombre de atributos relevantes. El objetivo es **predecir** la clase de nuevas instancias de las que no se conoce este dato, maximizando la razón de precisión del clasificador (r=predicciones correctas/total de predicciones)

Tareas de minería de datos

- Como una variante de la clasificación se tiene el **Ranking**, como aprendizaje de preferencias, estimadores de probabilidad, etc.

Tareas de minería de datos

- **Regresión**: Se busca aprender una función real; su objetivo es predecir el valor correspondiente para nuevas instancias.
- La diferencia entre la clasificación y la regresión es que el dato obtenido es real o discreto.
- El objetivo de la regresión es minimizar el error, comúnmente error cuadrático medio.

Tareas de minería de datos

- Agrupamiento (*clustering*) es una tarea descriptiva, que consiste en obtener grupos a partir de los datos.
- La diferencia entre grupos y clases es que en las clases se tiene datos etiquetados con las clases; mientras que en los grupos las instancias se agrupan buscando maximizar la similitud entre instancias del grupo y minimizar la similitud entre los distintos grupos.
- Cada grupo se considera como un resumen de los elementos que lo conforman, lo que es de utilidad para describir los datos de forma concisa.

Tareas de minería de datos

- **Correlaciones**: son una tarea descriptiva para analizar el grado de similitud de los valores de dos variables numéricas. Si el coeficiente de correlación es o indica que no hoy correlación entre las variables; valores cercanos a uno indican una correlación positiva, mientras que valores próximos a -1 son señal de una correlación negativa

Tareas de minería de datos

- **Reglas de asociación**: son tareas descriptiva para analizar el grado de relación entre de los valores de dos variables categóricas. Pueden ser de la forma si *X* toma el valor *a*, entonces *Y* toma el valor b.
- No implican una relación causa efecto

Tareas de minería de datos

- **Reglas de asociación secuenciales**: son un tipo especial de reglas de asociación, son una descriptiva para determinar patrones secuenciales de datos, basados en secuencias temporales de acciones.
- Difieren de otras reglas de asociación en que las relaciones de los datos dependen del tiempo
- No implican una relación causa efecto

Actividad de procesamiento

- Elaborar una tabla en la que se plasmen las tareas de minería de datos clasificadas por su finalidad (descriptivas o predicativas) y el tipo de variables que analiza (continuas, discretas o temporales)

Técnicas para minería de datos

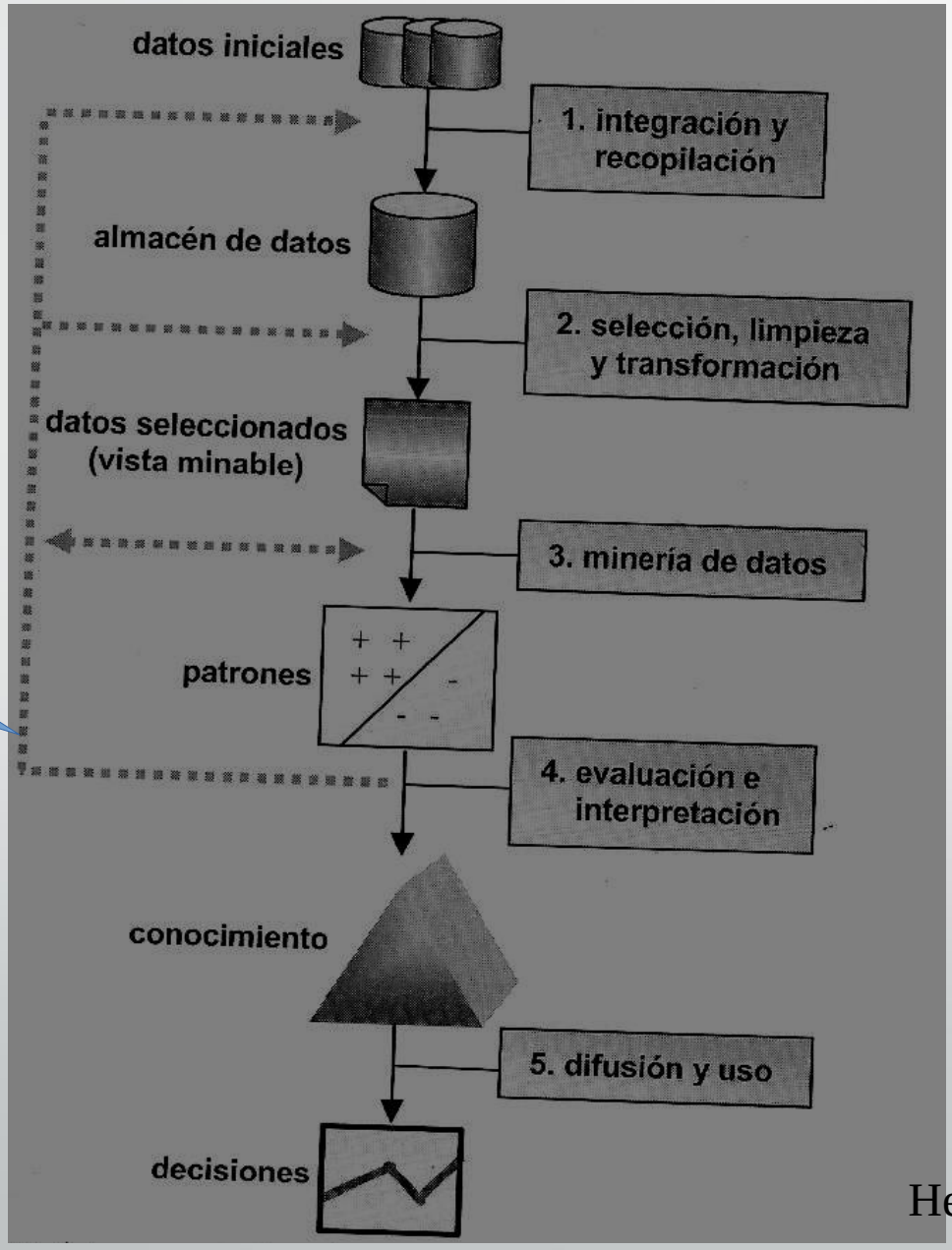
- En equipos de dos integrantes leer las técnicas de minería de datos mostradas en el capítulo 2 del libro de Hernández y realizar un cuadro de de resumen

Actividad para contrastar

- En equipos de cuatro integrantes, formado por un miembro de cada equipo conjuntar en un cuadro las actividades y técnicas relacionadas.

El proceso KDD

Se tiene un modelo iterativo



Hernández, 2008 p.20

La parte iterativa

- Una vez obtenido el modelo se debe evaluar
- Si satisface las necesidades, no es necesario hacer iteraciones; en otro caso si debe hacer otra iteración.
- Las iteraciones pueden ser para modificar el algoritmo, la técnica, afinar los datos o incluso los requerimientos.

Fase de evaluación e interpretación

- Se busca que los patrones descubiertos tengan las siguientes cualidades:
- Ser precisos
- Ser comprensibles
- Ser interesantes

Técnicas de evaluación

- Validación con datos conocidos, pero distintos a los usados en el entrenamiento.
- **Validación simple**: dividir el conjunto de datos en dos: un conjunto de entrenamiento y uno pequeño de prueba, los de prueba no se deben usar en el entrenamiento.

Técnicas de evaluación

- **Validación cruzada**: cuando se tiene pocos datos, el conjunto se divide en dos: A y B; la división es aleatoria y los conjuntos son del mismo tamaño, primero se entrena el modelo con los datos de A y se validan con B, se calcula el error; se entrenan los datos con B, se validan con A y se calcula el error, finalmente se entrenan con ambos, y se usa el modelo con el menor error

Técnicas de evaluación

- **Validación cruzada con n pliegues**: Es una variación de la cruzada, y consiste en dividir el conjunto en n grupos; se usa uno para el entrenamiento y n-1 para pruebas, se calcula el error y se repite el proceso n veces cambiando el conjunto de entrenamientos.

¿y como medimos si funciona?

- Dependiendo del problema.
- Para cada problema tenemos una medida de evaluación.

Medidas de evaluación

- Dependiendo del problema, **para clasificación** se mide la **precisión predictiva**, $p=c/n$
- c es el número de instancias del conjunto de prueba clasificadas correctas
- n es el número de instancias totales
- Se busca maximizar el valor de p tanto para el conjunto de entrenamiento como **para el de prueba**

Medidas de evaluación

- **Para reglas de asociación**, se evalúa de forma separada cada regla, considerando:
- **Cobertura** o soporte = #de instancias a las que la regla aplica y predice correctamente
- **Confianza**: proporción de instancias que la regla predice correctamente,= cobertura/# de reglas a las que se puede aplicar

Medidas de evaluación

- **Para regresión**, se evalúa el **error cuadrático medio** calculado como el promedio de los cuadrados de los valores esperados- valores calculados

Medidas de evaluación

- **Para agrupamiento**, la medida aplicada dependerá del método utilizado para su implementación, que de forma general miden la cohesión de los grupos, como por ejemplo la distancia media al centro, la **distancia** media entre grupos; o bien la **densidad**

Relación tarea -medida

- En equipos de dos integrante complemente la tabla de tareas de minería especificando la medida a evaluar en cada tarea.

Interpretación y contextualización

- Las medidas de evaluación debe ser llevadas al contexto específico, ya que este influye en la evaluación del modelo e interpretación de los resultados
- p.e. para un problema de clasificación con distribuciones de clases no balanceas un valor alto de precisiones no es sinónimo de eficiencia, (si una clase acumula el 90% de la población, y la precisión es .9, pude no saber clasificar a los elementos de clase minoritaria)

Interpretación y contextualización

- Para contemplar otros aspectos, se pueden usar herramientas como la **matriz de costos** de clasificación que es un tipos especial de **matriz de confusión**.
- Cuando no es viable estimar los costos de los errores se pude aplicar el **análisis ROC** (*Receiver Operating Characteristic*)

Fase de difusión y uso

- El uso de los modelos (generados y validados) se enfoca en dos escenarios:
- Para recomendar acciones con base en los resultados del modelo (ya se usado por una persona o un sistema)
- Para aplicar el modelo en distintos conjuntos de datos.

Fase de difusión y uso

- Es necesaria la difusión de los resultados obtenidos entre los posibles usuarios.
- Y dar seguimiento para la evolución del modelo, evaluándolo periódicamente ya que los patrones pueden evolucionar.
- Se pude requerir un ajuste o reconstrucción del modelo

Bibliografía básica

- Hernández Orallo, J., M. J. Ramírez Quintana, et al. (2004). Introducción a la Mineria de Datos. España, Pearson Educación SA.
- Han, D. J. (2007). Principles of Data Mining, MIT Press.
- Maimon, O. Z. and L. Rokach (2005). Data mining and knowledge discovery handbook. USA, Springer.
- Pérez López, C. and D. Santín Gonzalez (2006). Data Mining-Soluciones Con Enterprise Miner. México, Alfaomega, Ra-Ma.
- Sumathi, S. and S. N. Sivanandam (2006). Introduction to data mining and its applications. Berlín, Germany, Springer-Verlag New York Inc.
- Tan, P. N., M. Steinbach, et al. (2005). Introduction to data mining, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Bibliografía complementaria

- Everitt, B.S. (1994). A Handbook of Statistical Analyses using S-Plus. Chapman and Hall.
- Inmon, W.H. (1996). Building the Datawarehouse. J.Wiley & Sons.
- Han, J. and M. Kamber (2006). Data mining: concepts and techniques, Morgan Kaufmann.
- Kimball, R (1996). The Data Warehouse Toolkit. John Wiley & Sons.
- Hastie, T., R. Tibshirani, et al. (2005). The elements of statistical learning: data mining, inference and prediction, Springer
- Dunham. H. Margaret (2003). Data Mining. Introductory and Advanced Topics, Prentice Hall.
- Pyle, D. (1999), "Data Preparation for Data Mining", Morgan Kaufmann, San Francisco, CA.
- Hand, David; Mannila, Heikki; Smyth, Padhraic (2001), Principles of Data Mining, A Bradford Book. The MIT Press.
- Ian Witten and Eibe Frank (2002), Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.