ESMA 4016 Data Mining and Machine Learning

Clase 1

Dr. Edgar Acuna
Departamento de Ciencias Matematicas
Universidad de Puerto Rico-Mayaguez

E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com

Website: academic.uprm.edu/eacuna

Objetivos del curso

- Entender los conceptos fundamentales para llevara a cabo minería de datos y descubrimiento de conocimiento en base de datos usando metodos de Machine Learning.
- Experimentar algunos algoritmos más usados en minería de datos y Machine Learning en conjuntos de datos reales.

- Horario del curso: M y J de 3.00pm a 4.15pm en F323.

- Prerequistos del curso: Haber tomado un curso donde se hayan visto conceptos estadísticos.Tener algún conocimiento de matrices, sistemas de bases de datos y de algún programa de computación.

Oficina: OP307.
Horas de oficina: M y J de 9.00am a 10.30am y M 12.30pm a 3.30pm
Extension: x5872
Correo electronico del Profesor:
edgar.acuna@upr.edu ,
eacunaf@gmail.com

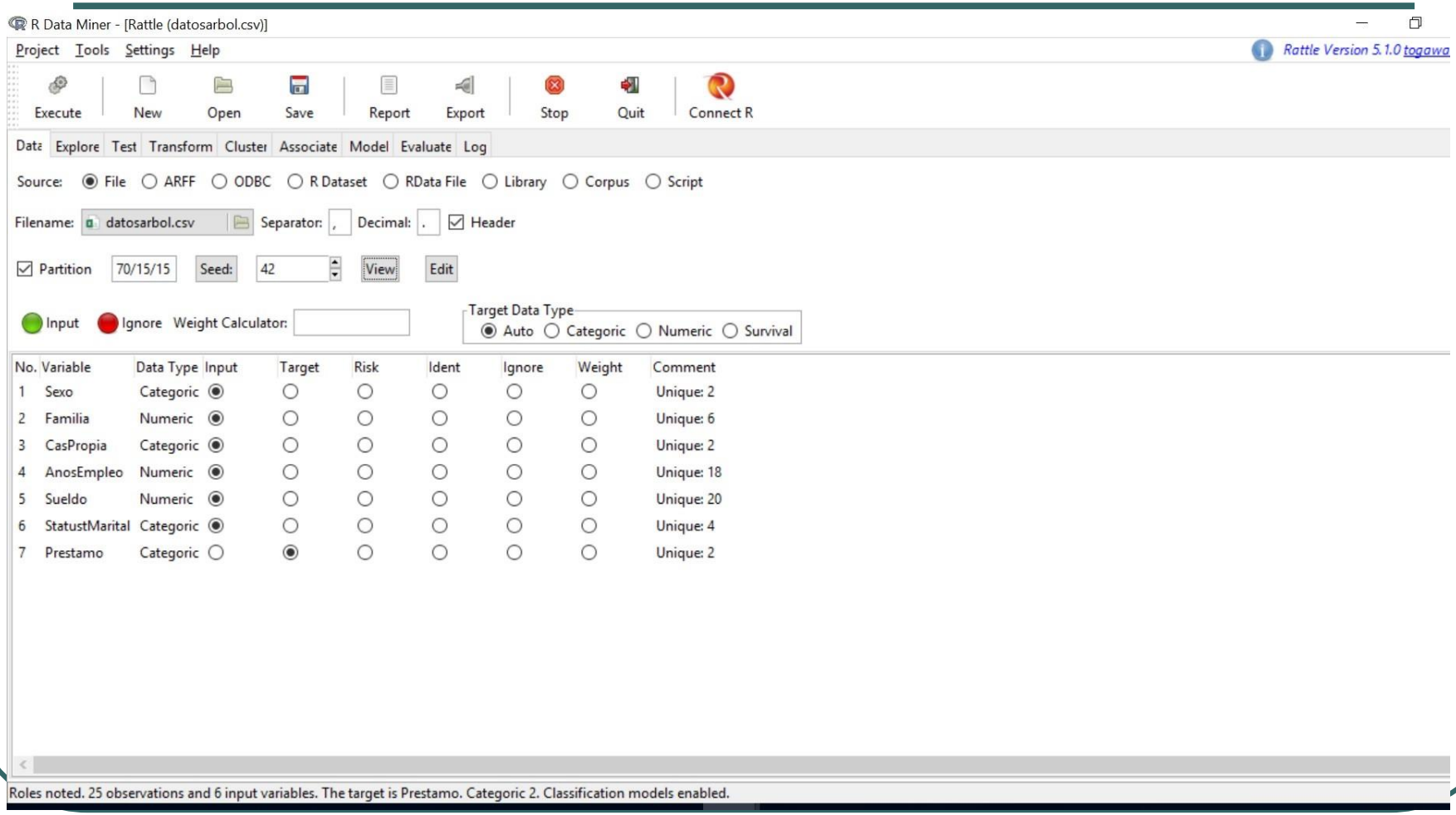
Textos

- James, Witten, Tibshirani & Hastie. Introduction to Statistical Learning Springer, 2014
- Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
- Torgo, Luis, Data Mining with R: Learning cases studies. CRC Press, 2010.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, 2005.
- Alpaydin, E. Introduction to Machine Learning. Third Edition. MiT Press, 2014
- Murphy, K. Machine Learning: a Probabilistic Perspective, 2012.

Software

- **Gratuitos:**
- R (cran.r-project.org). Inclinado a la estadistica (52.1% de usuarios segun kdnuggets.com)
- Python (python.org 52.6% de usuarios)
- Rapidminer (rapidminer.com) (32.8% de usuarios)
- **Comerciales:** Microsoft SQL (34.9%), (Excel (28.1%), KNIME (19.1%) , SAS Enterprise Miner (5.6%), IBM Watson(4.2%).

Rattle


R Data Miner - [Rattle (datosarbol.csv)]
Project Tools Settings Help
Rattle Version 5.1.0 togawa
Execute New Open Save Report Export Stop Quit Connect R
Data Explore Test Transform Cluster Associate Model Evaluate Log
Source: File ARFF ODBC R Dataset RData File Library Corpus Script
Filename: datosarbol.csv Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
No. Variable Data Type Input Target Risk Ident Ignore Weight Comment
1 Sexo Categoric Unique: 2
2 Familia Numeric Unique: 6
3 CasPropia Categoric Unique: 2
4 AnosEmpleo Numeric Unique: 18
5 Sueldo Numeric Unique: 20
6 StatustMarital Categoric Unique: 4
7 Prestamo Categoric Unique: 2
Roles noted. 25 observations and 6 input variables. The target is Prestamo. Categoric 2. Classification models enabled.

Paquetes en Python para Data Mining

Pandas: Para leer base de datos y hacer analisis estadistico basico.
Numpy: Metodos de Analisis Numericos
Matplotlib: Graficas estadisticas
Statmodels: Para hacer regression y series de tiempo
Scikit-learn: Algoritmos de Machine learning
Scipy: Libreria para hacer computacion cientifica entre ellas computos estadisticos.

Evaluacion

- Tareas (3) …… 40%
- Un examen Parcial 30%
- Proyecto ……. 30%

Motivacion

Los mecanismos para coleccion automatica de datos y el desarrollo de la tecnologia de bases de datos ha generado que se puedan almacenar grandes cantidades de datos en bases de datos, almacenes de datos y otros depositarios de informacion.

Hay la necesidad de convertir esos datos en conocimiento e informacion.

Tamano de datasets (en Bytes)

Description	Size	Storage Media
Very small	10^2	Piece of paper
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disk
Large	10^9(gigabite)	USB/Hard Disk
Massive	10^{12}(Terabyte)	Hard disk/USB
Super-massive	10^{15}(Petabyte)	File of distributed data
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

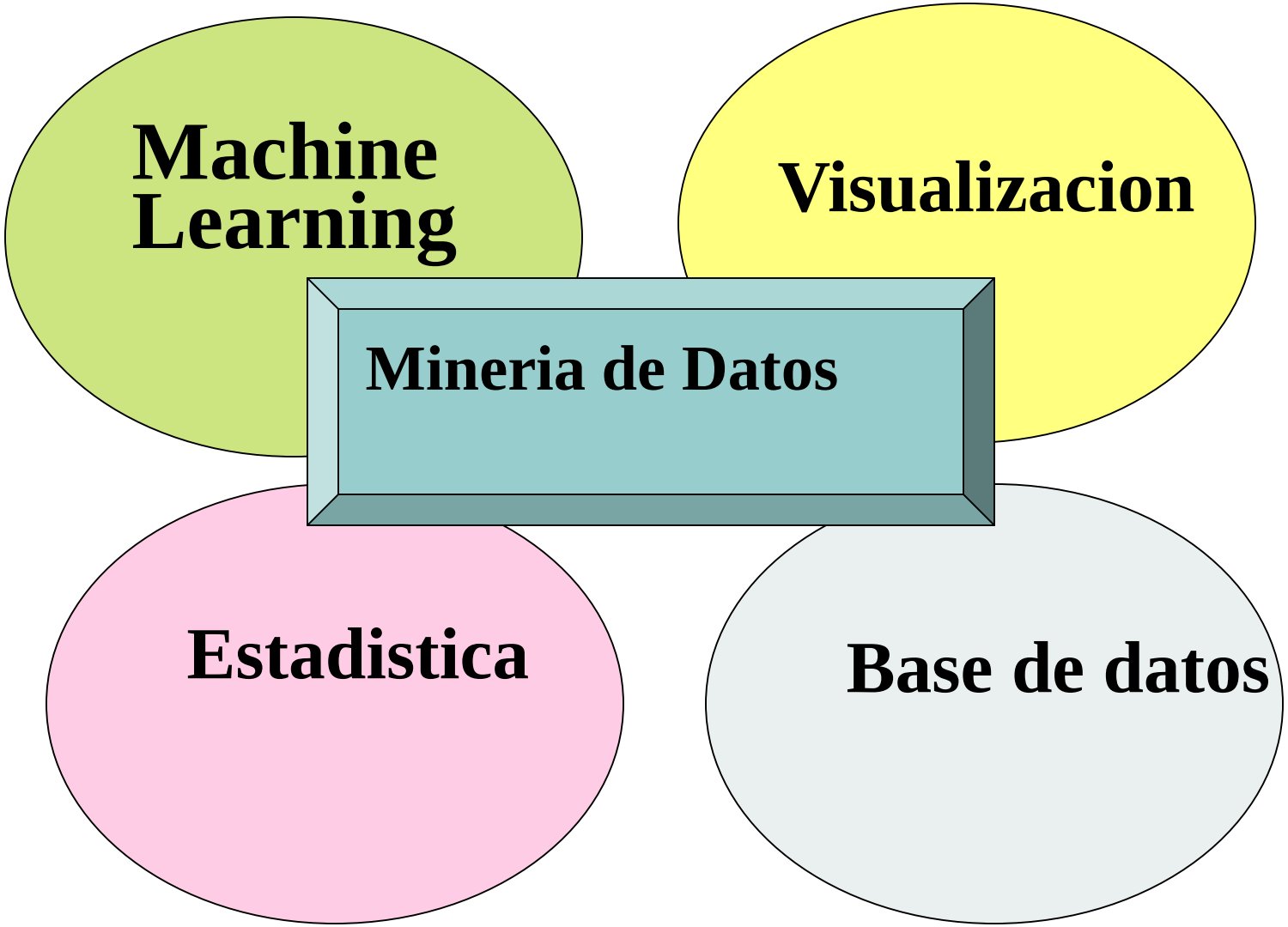
Ejemplos de grandes bases de datos hasta el 2016

- Hasta 2016, conjunto de datos de demora en los vuelos era de approx 25 Gigabytes.
- Amazon.com 45 TB de informacion de 60 millones de clients,
- En 2010, la base de datos de llamadas de ATT era de 323 Terabytes.
- Hasta el 2016,Google busca en mas de 130 trillones de paginas, que representa mas de 390 Petabytes.
- El telescopio Large Hadron Collider (LCH) almacena al año cerca de 600 Petabytes de datos de sensores.
- El 2013, se anuncio que el centro de datos de la NSA seria capaz de almacenar 5 zettabytes (1,000 exabytes).

Que es Mineria de Datos?

- Es el descubrimiento de conocimiento en un conjunto de datos enormemente grande. El conocimiento que se obtiene viene dado en forma de caracteristicas(patrones) que no son triviales, que son previamente desconocidas y que tienen bastante posibilidades de ser utiles.
- Otros nombres: Descubrimiento de conocimiento en bases de datos (KDD), extraccion de conocimiento, analisis inteligente de datos.

Areas relacionadas



Estadistica, Machine Learning

- Estadistica (~35% de DM)
  - Se basa mas en teoria. Asume propiedades distribucionales de las variables que estan siendo consideradas.
  - Se enfoca mas en probar hipotesis y en estimacion de parametros.
  - Se consideran eficientes estrategias de recolectar datos.
  - Estimacion de modelos.
- Machine learning (~30 % de DM)
  - Parte de Inteligencia Artificial. Machine es equivalente a un modelo en estadistica.
  - Mas heuristica que Estadistica.
  - Se enfoca en mejorar el rendimiento de un clasificador basado en sus experiencias pasadas.
  - Tambien considera el tiempo que dura el proceso de aprendizaje.
  - Incluye a: Redes Neurales, arboles de decision, Support Vector Machines, algoritmos geneticos.

Visualizacion, base de datos, etc

- Base de datos relacionales (~25% de DM)
  - Una base de datos relacional es un conjunto de tablas conteniendo datos de una categoria predeterminada. Cada una de las tablas (llamada relacion) contiene un o mas columnas de datos las cuales representan ciertos attributos. Cada una de las filas de la tabla contiene datos de las categorias definidas en las columnas.
  - Fue introducida por E. F. Codd de IBM en 1970.
  - El interface entre el usuario y la base de datos relacional mas usado es SQL(structured query laguage).
  - Una base de datos relacional puede ser agrandada facilmente
- Visualizacion (~5 % de DM)
  - Se explora la estructura del conjunto de datos en forma visual.
  - Puede ser usado en la etapa de pre o post procesamiento del KDD.
- Otras Areas (~ 5%): Pattern recognition, expert systems, High Performance Computing.

Data Mining no es …

- Buscar un numero en una guia telefonica
- Buscar una definicion en Google.
- Generar histogramas de salarios por grupos de edad.
- Hacer una consulta en SQL y leer la respuesta de la consulta.

Data mining es ...

- Hallar grupos de personas que padecen las mismas enfermedades.
- Determinar las caracteristicas de personas a las que se puede hacer un prestamo bancario.
- Detectar intrusos (casos anomalos) en un sistema.
- Determinar las caracteristicas de los clientes de un banco que pueden cometer fraude.
- Recommendar productos a un cliente basado en su historial de compras.
- Determinar las caracteristicas de los clientes que abandonan la subscripcion a un servicio.

Aplicaciones de DM

Administracion de negocios: Investigacion de mercados, relacion de los clientes con la gerencia, deteccion de Fraudes, Telecomunicaciones, etc.

Gobierno: deteccion de evasores de impuestos, terrorismo.

Ciencias: Astronomia, Bioinformatica (Genomics, Proteonomics, Metabolomics), decubrimiento de medicinas.

Text Mining: Extraer informacion previamente desconocida de diversas fuentes escritas (e-mails). Sentiment Analysis.

Recommendation Systems.

Web mining: E-comerce (Amazon.com)

Tipos de tareas en data mining

- Descriptivas: Se encuentra las propiedades generales de la base de datos. Se descubre las caracteristicas mas importantes de la base de datos.
- Predictivas: Se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. Nunca es 100% precisa y lo que mas importa es el rendimiento del modelo cuando es aplicado a nuevos datos.

Tareas en data mining

- Regresion (Predictiva)
- Classificacion (Predictiva)
- Classificacion No supervisada –Clustering (descriptiva)
- Reglas de Asociacion (descriptiva)
- Deteccion de Outliers (descriptiva)
- Visualizacion (descriptiva)
- Sistemas de Recomendacion (predictiva)
- Analisis de sentimiento (descriptiva)

Regresion

- Se predice el valor de una variable de respuesta continua basado en los valores de otras variables (predictoras) asumiendo que hay una relacion funcional entre ellas.
- Se puede usar modelos estadisticos, arboles de decision o redes neurales.
- Ejemplo: ventas de carros basados en las experiencia de los vendedores, publicidad, tipo de carros, etc.

Regresion[2]

- Regresion Lineal $Y=b_o+b_1X_1+.....b_pX_p$
- Regresion No-Lineal, $Y=g(X_1,\ldots,X_p)$, donde g es una funcion no lineal. Por ejemplo, $g(X_1,\ldots X_p)=X_1\ldots X_p e^{X1+\ldots Xp}$
- Regresion No-parametrica $Y=g(X_1,\ldots,X_p)$, donde g es estimada usando los datos disponibles.

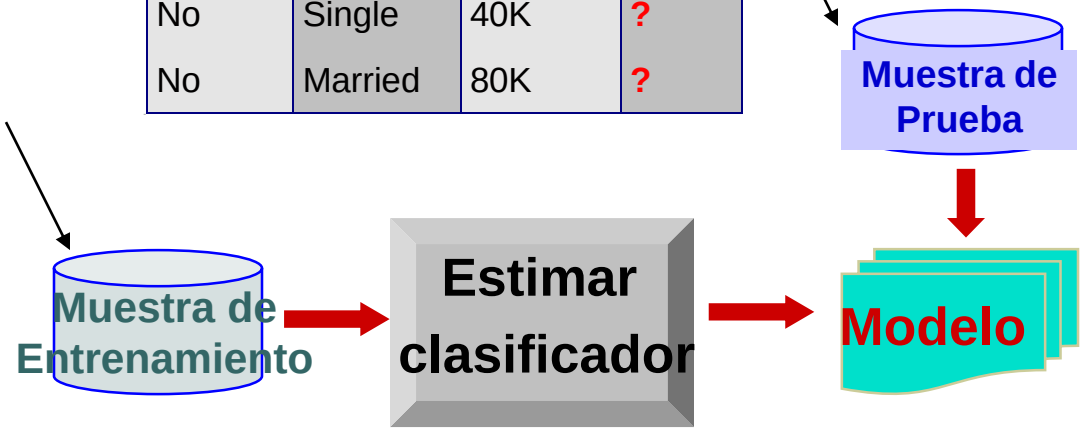
Clasificacion Supervisada

- Dado un conjunto de registros (records), llamado el conjunto de entrenamiento, cada registro contiene un conjunto de atributos y usualmente el ultimo atributo es la clase, debemos encontrar un modelo para el atributo clase en funcion de los valores de los otros atributos.
- *Objetivo: Asignar records que no se habian visto previamente (muestra de prueba) a una clase de la manera mas precisa posible.*
- Usualmente el conjunto dado es dividido en muestra de entrenamiento (70%) y muestra de prueba (30%). La primera es usada para construir el modelo y la segunda es usada para validarlo. La precision del modelo es determinada en la muestra de prueba.

Ejemplo de Clasificacion

	categorica	categorica	continua	clase
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Clasificacion Supervisada[2]

- Clasificacion supervisada puede ser considerada como un proceso de decision y la regla de decision es llamada un clasificador.
- Ejemplos de clasificadores: Analisis de discriminante Lineal (LDA), regresion logistica, k-vecinos mas cercanos, estimadores de densidad, naïve Bayes, arboles de decision, redes neurales, support vector machines.

Clasification No-supervisada (Clustering)

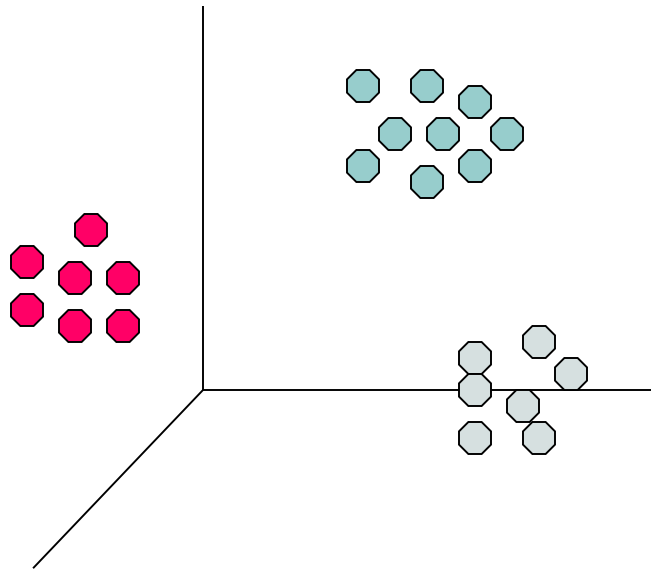
- Determinar grupos de objetos (clusters) de tal manera que los objetos dentro del mismo cluster sean bastante similar entre si mientras que objetos en grupos distintos no sean tan similares entre si.
- Se necesita usar una medida de similaridad para establecer si dos objetos pertenecen a un mismo cluster o a clusters distintos.
- Ejemplos de medidas de similaridad: Distancia Euclideana, distancia Manhattan, correlacion, distancia Hamming, etc.
- Problemas: Eleccion de la medida de similaridad, eleccion del numero de clusters, validacion de clusters.

Clustering[2]

- Clustering tri-dimensional basado en distancia euclideana.

Las distancias Intracluster son minimizadas

Las distancias Intercluster son maximizadas



Algoritmos de Clustering

- Algoritmos de Particionamiento: K-means, PAM, SOM.
- Algoritmos Jerarquicos: Aglomerativo, Divisivo.

Deteccion de “outliers”

- Los objetos que se comportan diferente o que son inconsistentes con la mayor parte de los datos son llamados “outliers”.
- Outliers pueden ser causados por un error de medicion o de ejecucion. Ellos pueden representar algun tipo de actividad fraudulenta.
- El objetivo de la deteccion de “outliers” es detectar las instancias que tienen un comportamiento fuera de lo comun.

Deteccion de “outliers”[2]

- Metodos:
  - Metodos basados en Estadisticos
  - Metodos basados en distancia
  - Metodos basados en densidad local.
- Aplicacion: Deteccion de fraude en tarjeta de creditos, Network intrusion

Reglas de asociacion

- Dado un conjunto de registros cada uno de los cuales contiene algun numero de items de una coleccion dada. El objetivo es encontar reglas de dependencia que permitan predecir la ocurrencia de un item basado en ocurrencia de otros items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Reglas descubiertas:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Reglas de Asociacion[2]

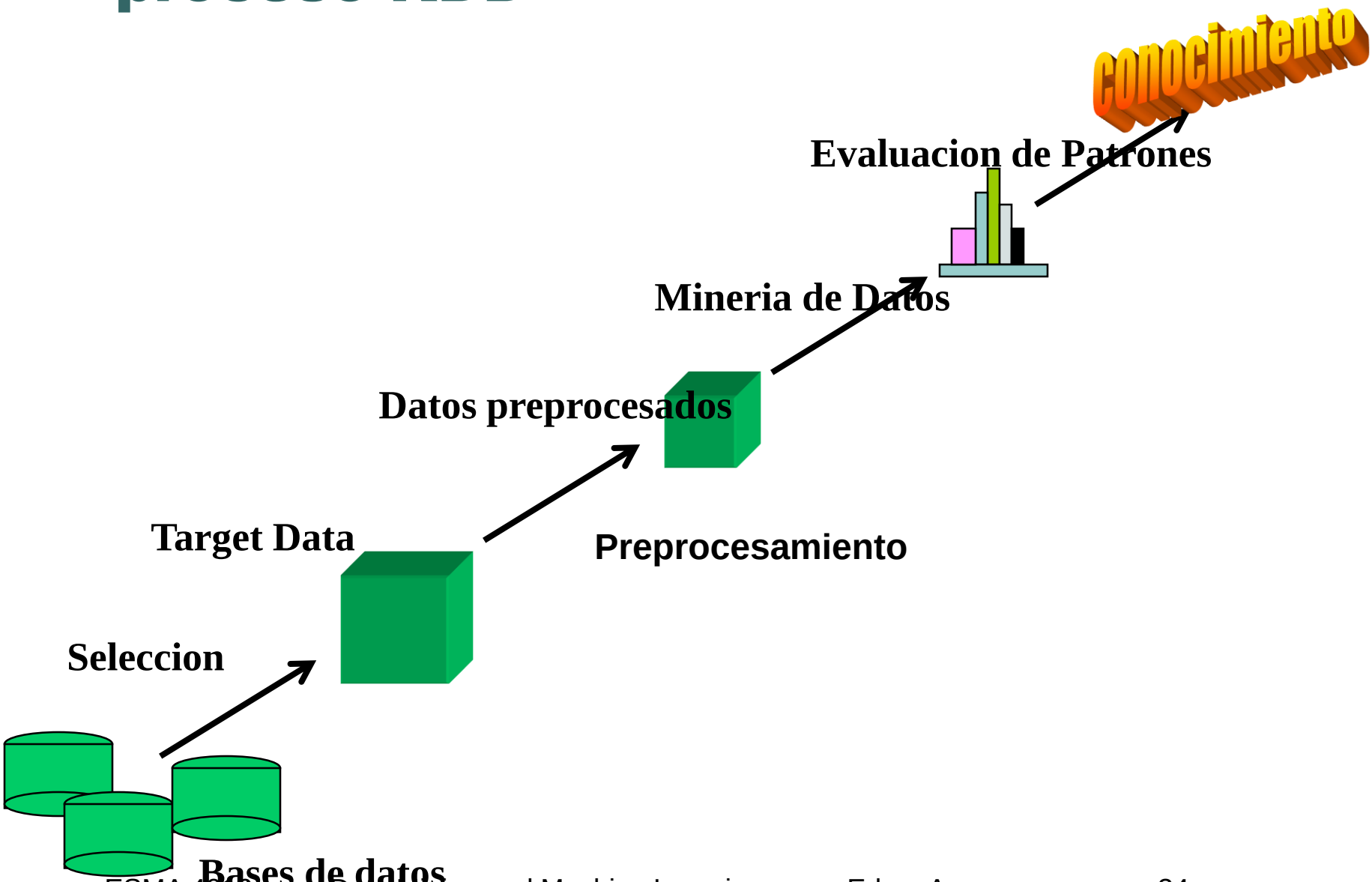
- Las reglas (X->Y) deben satisfacer un soporte minimo y una confianza impuesta por el usuario. X es llamado el antecedente Y es llamado el consecuente.
- Soporte=(# registros conteniendo X y Y)/(# registros)
- Confianza=(# registros conteniendo X y Y/(# de registros conteniendo X)

Ejemplo: El soporte de la Regla 1 es .6 y de la regla 2 es .4

La confianza de la Regla 1 es .75 y de la regla 2 es .67

Aplicacion: Mercadeo y Promocion de ventas

Mineria de Datos como un paso del proceso KDD



Steps of a KDD Process

- Conocer el dominio de la aplicacion. Sus antecedentes y objetivos.
- Determinar un target data set.
- Data cleaning and pre-procesamiento (puede requerir entre 60-80% del proceso total)
- Data reduction and transformation. Hallar variables importantes, reducir la dimensionalidad.
- Escoger la tarea de data mining que se va a usar: Sumarizacion, Classificacion, Regresion, Asociacion, clustering.
- Escoger el algoritmo de data mining que se va usar.
- Buscar los patrones mas interesantes
- Evaluacion de Patrones y representacion del conocimiento.

Retos de Data Mining

- Escalabilidad
- Dimensionalidad
- Datos complejos y Heterogeneos.
- Calidad de datos
- Propiedad y distribucion de datos
- Preservacion de privacidad