COMP 6315: Mineria de Datos

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.
- Experimentar algunos algoritmos más usados en minería de datos en conjuntos de datos reales.

- Horario del curso: W de 9.00 a 11.30am en M118.

- Prerequistos del curso: Haber tomado dos cursos donde se hayan visto conceptos estadísticos, por lo menos uno de ellos a nivel graduado. Tener algún conocimiento de matrices, sistemas de bases de datos y de algún programa de computación.

Oficina: M314.
Horas de oficina: W 7.30am a 11.30am y J de 12 a 2pm
Extension: x3287
Correo electronico del Profesor:
edgar.acuna@upr.edu ,
edgar@cs.uprm.edu eacunaf@gmail.com

Textos

- Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
- Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, 2005.
- Michael Berry & Gordon Linoff, Mastering Data Mining, John Wiley & Sons, 2000.
- Torgo, Luis, Data Mining with R: Learning cases studies. CRC Press, 2010.
- **G. Shmueli, N.R. Patel y P.C. Bruce, Data Mining for Business Intelligence, John Wiley and Sons, 2007.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, 2001.
- Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, Wiley-IEEE Press, 2002.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, 2005.
- David J. Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining , MIT Press, 2000.

Software

- **Gratuitos:**
- R (cran.r-project.org). Inclinado a la estadistica (47% de usuarios, Kdnuggets 2012).
- Weka (http://www.cs.waikato.ac.nz/ml/weka/) (23%): Escrito en Java, tiene manual en espanol.
- Orange (http://www.ailab.si/orange). Bastante completo, pero requiere instalar Python y otros programas adicionales.
- **Comerciales:** SAS Enterprise Miner (31%), SPSS Clementine (30%), Statistica (22%) XLMiner(www.xlminer.com), Microsoft SQL(18%), Oracle.

Evaluacion

- Tareas (4) ……. 40%
- Un examen Parcial 30%
- Proyecto ……. 30%

Contenido del curso

I.Introduccion (3 horas)
II Pre-procesamiento (12 horas)
III. Clasificacion Supervisada (12 horas)
III. Visualizacion (5 horas)
IV. Deteccion de outliers (5 horas)
V. Clustering (6 horas)

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.

Tamanos de conjunto de datos

Descripcion	Tamano en Bytes	Modo de almacenaje
Bien pequeno	10^2	Hoja de papel
pequeno	10^4	Varias hojas
Mediano	10^6 (megabyte)	Diskette
Grande	10^9(gigabite)	Disco Duro
Masivo	10^{12}(Terabyte)	Cinta magnetica
Supermasivo	10^{15}(Petabyte)	Archivos de datos distribuidos

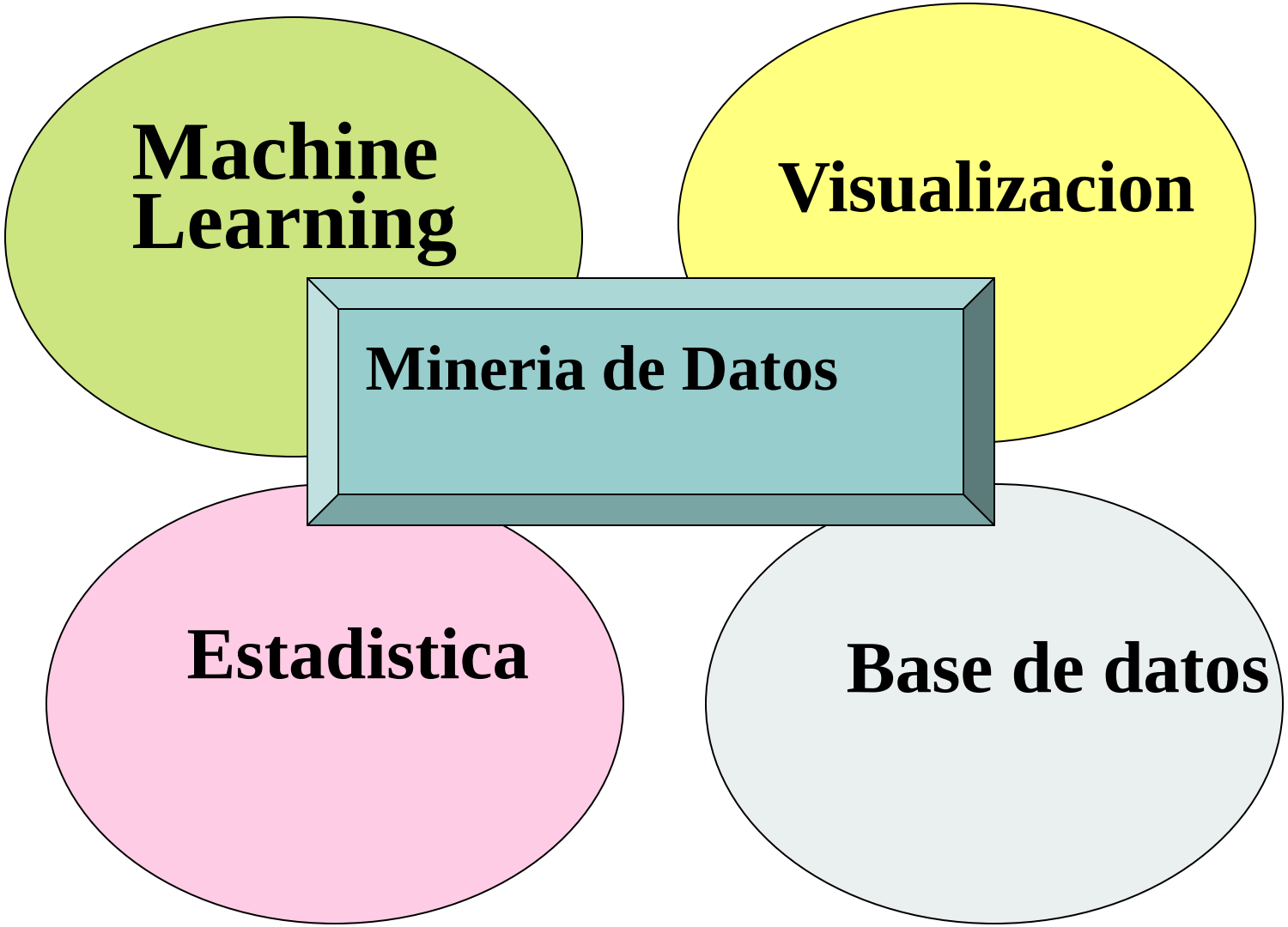
Ejemplos de grandes bases de datos

- Un telescopio puede generar hasta 1 gigabyte de datos astronomicos por segundo.
- ATT almacena hasta 26 Terabytes de informacion en llamadas telefonicas.
- El 2003, Walmart almacenaba 10 Terabytes de transacciones por dia.
- Google busca en mas de 10 billones de paginas, que representa mas de 250 TB.
- The internet archive (www.archive.org) almacena mas de 300TB de informacion.
- Se estima que en el 2002 se genero 5 exabytes(5 millones de TB) de nuevos datos.

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 (~40% 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 efiicientes estrategias de recolectar datos.
  - Estimacion de modelos.
- Machine learning (~25 % 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, algoritmos geneticos.

Visualizacion, base de datos

- Base de datos relacionales (~20% 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 (~15 % 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.

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 un SQL query y leer la respuesta del query.

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.
- 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)

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)

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

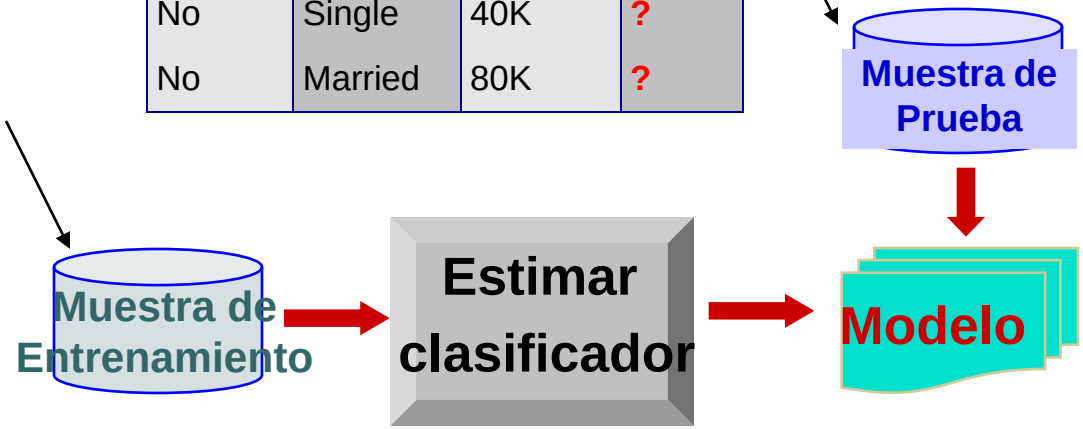
- Dado un conjunto de registros (records), llamado el conjunto de entrenamiento, cada registro contiene un conjunto de atributos y usualmente el ultimo atributo es la clase, debemos encontrar un modelo para el atributo clase en funcion de los valores de los otros atributos.
- *Objetivo: Asignar records 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)

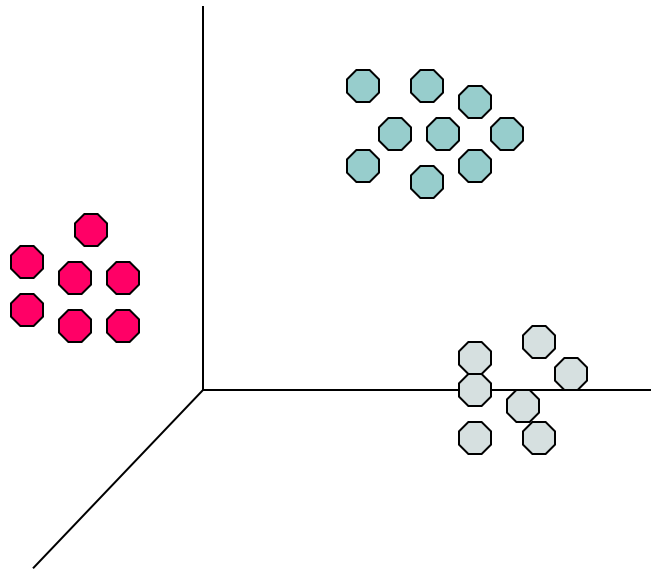
- Determinar grupos de objetos (clusters) de tal manera que los objetos dentro del mismo cluster sean bastante similar entre si mientras que objetos en grupos distintos no sean tan similares entre si.
- Se necesita usar una medida de similaridad para establecer si dos objetos pertenecen a un mismo cluster 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]

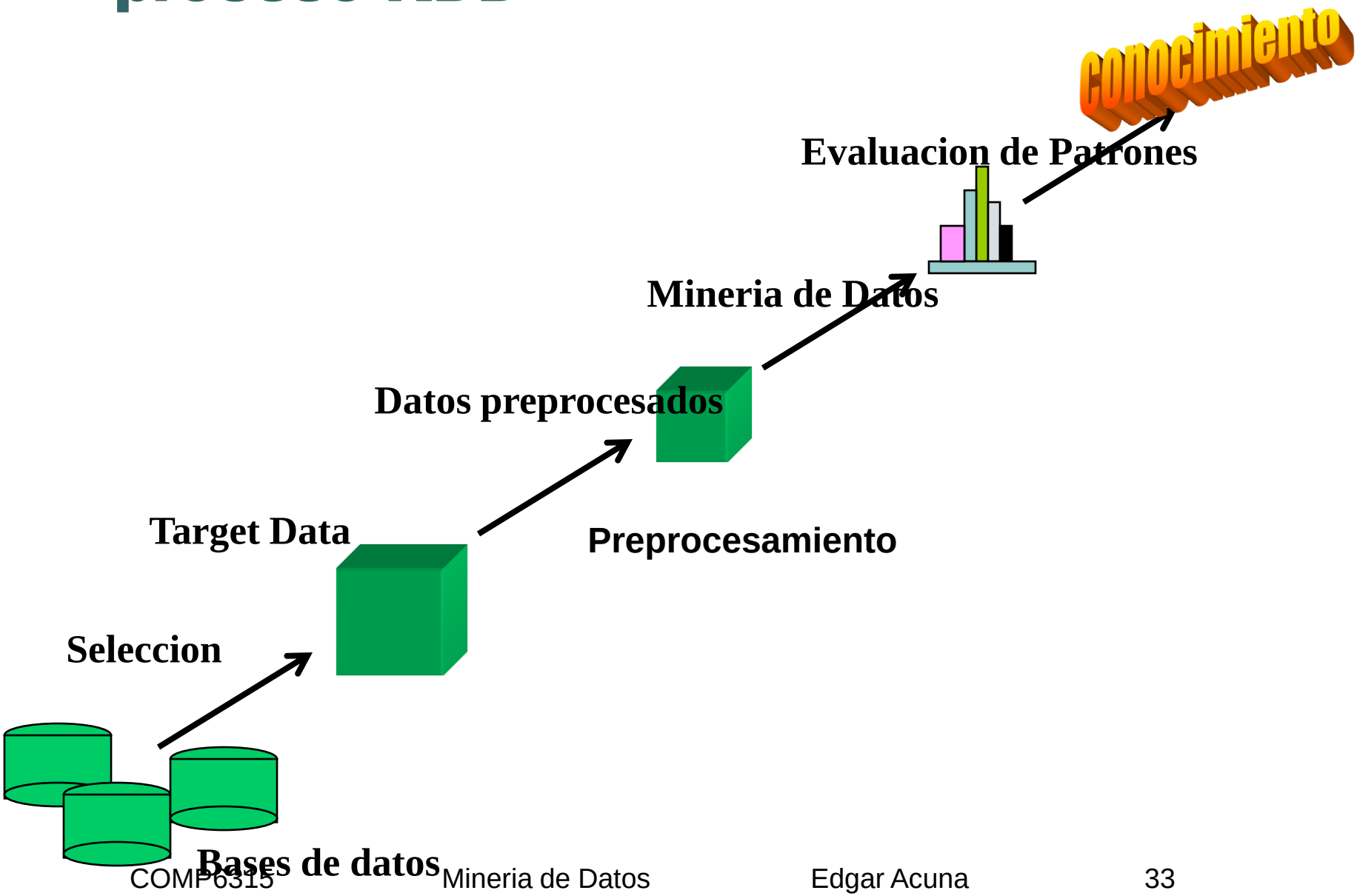
- Las reglas (X->Y) deben satisfacer un soporte minimo y una confianza impuesta por el usuario. X es llamado el antecedente Y es llamado el consecuente.
- Soporte=(# registros conteniendo X y Y)/(# registros)
- Confianza=(# registros conteniendo X y Y/(# de registros conteniendo X)

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