

Introducción a la Minería de Datos

CC52A - Inteligencia Artificial

Gonzalo Ríos D.

DCC - UChile

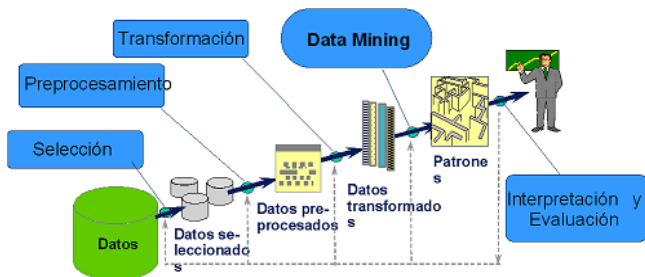
Otoño 2011

Introducción a la Minería de Datos

Knowledge Discovery in Databases

Definición

KDD es el proceso no-trivial de identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos



Introducción a la Minería de Datos

Knowledge Discovery in Databases



Copyright © 2000 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited



Copyright © 2000 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación
 - Exploración y Visualización de los Datos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación
 - Exploración y Visualización de los Datos
 - Definición de modelos y patrones a encontrar

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación
 - Exploración y Visualización de los Datos
 - Definición de modelos y patrones a encontrar
 - Diseño, desarrollo y ejecución de algoritmos

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación
 - Exploración y Visualización de los Datos
 - Definición de modelos y patrones a encontrar
 - Diseño, desarrollo y ejecución de algoritmos
- Evaluación e Interpretación de los Resultados

Introducción a la Minería de Datos

Proceso KDD

- Entendimiento del Problema
- Entendimiento de los Datos
 - Qué información y estructura de ella tenemos
 - Qué partes y subconjuntos son relevantes
 - Calidad de los datos
- Preparación de los Datos
 - Extracción
 - Transformación y Limpieza
- Modelación
 - Exploración y Visualización de los Datos
 - Definición de modelos y patrones a encontrar
 - Diseño, desarrollo y ejecución de algoritmos
- Evaluación e Interpretación de los Resultados
- Implementación del modelo

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica
- Aprendizaje

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica
- Aprendizaje
 - Machine Learning

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica
- Aprendizaje
 - Machine Learning
 - Optimización

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica
- Aprendizaje
 - Machine Learning
 - Optimización
 - Cálculo Numérico

Introducción a la Minería de Datos

Disciplinas relacionadas con Minería de Datos

- Almacenamiento y Extracción de Datos
 - Bases de Datos
- Modelación
 - Algoritmos
 - Probabilidades
 - Estadística
 - Lógica
- Aprendizaje
 - Machine Learning
 - Optimización
 - Cálculo Numérico
 - Computación Paralela

Introducción a la Minería de Datos

Tareas en Minería de Datos: Clasificación

Definición

Clasificación es una tarea predictiva, en donde cada instancia pertenece a una clase, la cual se indica mediante el valor de un atributo discreto que llamamos la clase de la instancia. El resto de los atributos de la instancia se utilizan para predecir la clase.

Ejemplo

A partir de datos de los clientes de un banco, crear un sistema de análisis de riesgo, que recomiende si dar crédito o no a la persona que lo solicita



Introducción a la Minería de Datos

Tareas en Minería de Datos: Regresión

Definición

Regresión es una tarea predictiva, que consiste en aprender una función real que asigna a cada instancia un valor real. Un caso particular son las regresiones en series de tiempo, en donde la variable va cambiando su valor a lo largo del tiempo.

Ejemplo

Una compañía financiera desea crear un modelo que prediga el valor del dólar en el futuro, dado sus valores en el pasado.



Introducción a la Minería de Datos

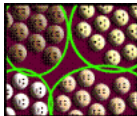
Tareas en Minería de Datos: Clustering

Definición

Clustering es una tarea descriptiva, que consiste en obtener grupos "naturales" a partir de los datos. A diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta

Ejemplo

Un sitio web, con el fin de dar un servicio más personalizado, desea encontrar grupos importantes de usuarios, a base a sus preferencia de compras.



Introducción a la Minería de Datos

Tareas en Minería de Datos: Reglas de Asociación

Definición

Reglas de Asociación es una tarea descriptiva, que tiene como objetivo identificar relaciones no explícitas entre distintos atributos, principalmente categóricos. Una regla de asociación tiene la forma de "Si $X=x$, entonces $Y=y$ ". Si los atributos son numéricos, se le llama correlación, que entrega el grado de similitud de las variables.

Ejemplo

Un supermercado desea encontrar reglas de asociación entre sus productos, con el fin de crear promociones tentadoras para los clientes, y así subir sus ventas.



Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión
- Redes Neuronales

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas
- Algoritmos Evolutivos

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesianas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas
- Algoritmos Evolutivos
- Conjuntos Difusos

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesianas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas
- Algoritmos Evolutivos
- Conjuntos Difusos
- Cadenas de Markov Ocultas

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas
- Algoritmos Evolutivos
- Conjuntos Difusos
- Cadenas de Markov Ocultas
- Análisis de Componentes Principales

Introducción a la Minería de Datos

Técnicas en Minería de Datos

- Regresión Funcional
- Máquina de Vectores Soporte
- Redes Bayesinas
- Árboles de Decisión
- Redes Neuronales
- Inducción de Reglas
- Algoritmos Evolutivos
- Conjuntos Difusos
- Cadenas de Markov Ocultas
- Análisis de Componentes Principales
- Vecinos Cercanos

Introducción a la Minería de Datos

Clasificación de Modelos en Minería de Datos

Existen diversas clasificaciones para los Modelos en Minería de Datos. Una de las clasificaciones más relevantes es:

- **Generalización del modelo**

Existen diversas clasificaciones para los Modelos en Minería de Datos. Una de las clasificaciones más relevantes es:

- **Generalización del modelo**

- Los métodos **retardados** (lazy) se caracterizan porque retrasan la decisión de la generalización del conjunto de entrenamiento, hasta el instante en que se recibe un nuevo dato a procesar. Los ejemplos de entrenamiento nunca son desechados, ya que el procesamiento de nuevos datos se obtienen a partir de estos.
Ejemplo: Vecinos Cercanos

Existen diversas clasificaciones para los Modelos en Minería de Datos. Una de las clasificaciones más relevantes es:

- **Generalización del modelo**

- Los métodos **retardados** (lazy) se caracterizan porque retrasan la decisión de la generalización del conjunto de entrenamiento, hasta el instante en que se recibe un nuevo dato a procesar. Los ejemplos de entrenamiento nunca son desechados, ya que el procesamiento de nuevos datos se obtienen a partir de estos.

Ejemplo: Vecinos Cercanos

- Los métodos **anticipativos** (eager) contruyen un modelo de generalización, antes de tener que realizar la tarea de generalización. En este caso, los ejemplos son desechados y reemplazados por el modelo.

Ejemplo: Redes Neuronales

Introducción a la Minería de Datos

Clasificación de Modelos en Minería de Datos

Otra clasificación de Modelos muy importante es:

- **Supervisión del entrenamiento**

Otra clasificación de Modelos muy importante es:

- **Supervisión del entrenamiento**

- Los métodos **supervisados** se caracterizan ya que en la fase de entrenamiento del modelo, se sabe a priori el valor que deben tener los ejemplos de aprendizaje. Esto permite verificar si el modelo es correcto o no.

Ejemplo: Regresión Funcional

Otra clasificación de Modelos muy importante es:

- **Supervisión del entrenamiento**

- Los métodos **supervisados** se caracterizan ya que en la fase de entrenamiento del modelo, se sabe a priori el valor que deben tener los ejemplos de aprendizaje. Esto permite verificar si el modelo es correcto o no.

Ejemplo: Regresión Funcional

- En los métodos **no supervisados** no se sabe a priori el valor de los ejemplos de entrenamiento, sino que luego del aprendizaje se conocen las características. Esto no permite verificar si un modelo es correcto o no, por lo que se ocupan otras nociones.

Ejemplo: Inducción de Reglas

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos
 - **Recall**: Proporción de verdaderos positivos de entre los positivos

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos
 - **Recall**: Proporción de verdaderos positivos de entre los positivos
- En la **regresión**, se busca que el modelo que minimice el error de predicción, dentro de los cuales:

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos
 - **Recall**: Proporción de verdaderos positivos de entre los positivos
- En la **regresión**, se busca que el modelo que minimice el error de predicción, dentro de los cuales:
 - **Error cuadrático**: $\sum (y_i - f(x_i))^2$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos
 - **Recall**: Proporción de verdaderos positivos de entre los positivos
- En la **regresión**, se busca que el modelo que minimice el error de predicción, dentro de los cuales:
 - **Error cuadrático**: $\sum (y_i - f(x_i))^2$
 - **Error relativo**: $\sum \left(\frac{y_i - f(x_i)}{y_i} \right)^2$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En la **clasificación**, se busca que el modelo que minimice el error de clasificación, dentro de los cuales:
 - **Falsos Positivos**: Datos clasificados en la clase objetivo de forma errónea
 - **Falsos Negativo**: Datos clasificados en las clases no objetivo de forma errónea
 - **Tasa de Error**: Proporción de datos clasificados de forma errónea
 - **Precisión**: Proporción de verdaderos de entre los predichos como positivos
 - **Recall**: Proporción de verdaderos positivos de entre los positivos
- En la **regresión**, se busca que el modelo que minimice el error de predicción, dentro de los cuales:
 - **Error cuadrático**: $\sum (y_i - f(x_i))^2$
 - **Error relativo**: $\sum \left(\frac{y_i - f(x_i)}{y_i} \right)^2$
 - **Error absoluto**: $\sum \left| \frac{y_i - f(x_i)}{y_i} \right|$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.

- **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.
 - **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
 - **Distancia valor absoluto:** $d(x, y) = \sqrt{\sum |x_i - y_i|}$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.
 - **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
 - **Distancia valor absoluto:** $d(x, y) = \sqrt{\sum |x_i - y_i|}$
 - **Distancia Mahalanobis:** $d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.
 - **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
 - **Distancia valor absoluto:** $d(x, y) = \sqrt{\sum |x_i - y_i|}$
 - **Distancia Mahalanobis:** $d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$
- En las **reglas de asociación**, se desea maximizar la precisión y el soporte:

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.
 - **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
 - **Distancia valor absoluto:** $d(x, y) = \sqrt{\sum |x_i - y_i|}$
 - **Distancia Mahalanobis:** $d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$
- En las **reglas de asociación**, se desea maximizar la precisión y el soporte:
 - **Precisión:** Porcentaje de predicción correcta, cuando se puede aplicar

Introducción a la Minería de Datos

Tareas y Función Objetivo en Minería de Datos

- En el **clustering**, se desea encontrar grupos que los miembros dentro de un grupo tengan la mínima distancia, y los grupos entre sí tengan la máxima distancia.
 - **Distancia euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
 - **Distancia valor absoluto:** $d(x, y) = \sqrt{\sum |x_i - y_i|}$
 - **Distancia Mahalanobis:** $d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$
- En las **reglas de asociación**, se desea maximizar la precisión y el soporte:
 - **Precisión:** Porcentaje de predicción correcta, cuando se puede aplicar
 - **Soporte:** Número de instancias de predicción correcta

Introducción a la Minería de Datos

Evaluación en Minería de Datos

Definición

Holdout es el método de validación, en donde se particiona el conjunto de casos en dos grupos: uno de entrenamiento y otro de testeo. El grupo de entrenamiento es usado para la inducción del modelo, y el de testeo para para estimar la tasa de error.

Definición

Random Subsampling es la realización múltiple de Holdout, con grupos independientes, y la estimación de la tasa de error es la media de las tasas de errores.

Introducción a la Minería de Datos

Evaluación en Minería de Datos

Definición

Validación Cruzada se basa en particionar los datos en K subconjuntos de, aproximadamente, el mismo tamaño. Luego, para cada uno de estos grupos, éste se usa para testear el modelo y los $K-1$ grupos restantes se usan para la inducción del modelo. Luego, la estimación de la tasa de error es la media de las tasas de errores.

Definición

Bootstrapping se basa en generar un conjunto de entrenamiento y otro de testeo, escogidos de forma aleatoria con reemplazamiento del conjunto de datos iniciales. Una variación es que el grupo de testeo son los datos no escogidos como datos de entrenamiento.