MINERIA DE DATOS USANDO SISTEMAS INTELIGENTES

Año 2014

Carrera/Plan:
Licenciatura en Sistemas Plan 2003/07
Licenciatura en Informática Plan 2003/07

Área: Algoritmos y Lenguajes
Año: 4º o 5º año
Régimen de Cursada: *Semestral*
Carácter: Optativa
Correlativas:
Computabilidad y Complejidad (LI)
Algoritmos y estructuras de datos
Fundamentos de teoria de la computación (LS)

Profesor: *Laura Lanzarini*
Hs semanales: 6 hs

FUNDAMENTACIÓN

La Minería de Datos forma parte del proceso de Extracción de Conocimiento y consiste de técnicas que, a partir de datos almacenados en grades bases de datos, poseen la capacidad de adquirir conocimiento nuevo, novedoso y potencialmente útil.

El resultado de la aplicación de estas técnicas es un *modelo* de la información disponible que, expresado en forma de un conjunto de reglas, un árbol o una red neuronal, permite resumir las relaciones existentes entre los datos.

Habitualmente, ante la presencia de grandes volúmenes de información, lo que se hace es contrastar una hipótesis predeterminada, por ejemplo, a través de consultas SQL. En Minería de Datos, el proceso es totalmente inverso llegando a obtener relaciones entre los datos sin tener ninguna hipótesis preestablecida.

El curso es introductorio y presenta no sólo las técnicas más utilizadas en Minería de Datos sino que agrega otro enfoque menos tradicional basado en Redes Neuronales y Técnicas de optimización.

OBJETIVOS GENERALES

Introducir al alumno en las técnicas de Minería de Datos. Se analizarán modelos basados en regresión, árboles y reglas. Se presentarán los principios básicos de Redes Neuronales y Técnicas de Optimización aplicados a la Minería de Datos. El énfasis está puesto en la resolución de problemas de clasificación y predicción.

CONTENIDOS MINIMOS

Introducción a la Minería de Datos.
Técnicas
- Reglas de asociación
- Arboles de decisión y sistemas de reglas
- Redes Neuronales
- Extracción de conocimiento utilizando técnicas de optimización.

3

PROGRAMA ANALÍTICO

Minería de Datos

- Introducción. Fases del proceso de extracción del conocimiento. Relación con otras disciplinas.
- Recuperación de información vs recuperación de datos. Proceso de recuperación de información.
- Preparación de los datos. Recopilación. Limpieza. Exploración y Selección.
- Técnicas de Minería de Datos. Métodos basados en una medida de similitud. Agrupamiento y Clasificación. Arboles de decisión y Sistemas de Reglas de asociación

Redes Neuronales

- Introducción a las Redes Neuronales Artificiales. Los primeros modelos computacionales. La neurona artificial. Estructura básica de la red. Aprendizaje. Tipos de problemas que puede resolver. Perceptrón. Descripción del modelo. Resolución de problemas de clasificación. Adaline: Entrenamiento de un combinador lineal a través de la regla delta. Diferencias con el Perceptrón. Neurona no lineal. Funciones de transferencia sigmoides. Relación de los resultados obtenidos con los dos modelos anteriores.
- Perceptrón multicapa. Descripción de la arquitectura feedforward. Regla delta generalizada. Algoritmo de entrenamiento backpropagation. Incorporación del término de momento. Capacidad de generalización de la red. Resolución de problemas de clasificación y predicción.
- Redes Neuronales Competitivas. Mapas auto-organizativos (SOM). Descripción de la arquitectura. Concepto de vecindad. Resolución de problemas de clustering. Mapas auto-organizativos dinámicos (GSOM). Comparación con el modelo anterior. Resolución de problemas de agrupamiento aplicables a problemas tales como detección de fraudes, identificación de perfiles de clientes, etc.

Técnicas de Optimización Aproximadas

- Paradigmas Principales. Programación Evolutiva, Estrategias Evolutivas y Algoritmos Genéticos. Comparaciones entre los distintos paradigmas. Terminología. Introducción. Conceptos Básicos. Representación. Técnicas de Selección y Cruce.
- Optimización por cúmulos de partículas (PSO – Particle Swarm Optimization). Algoritmos básicos GBest y LBest. Variaciones: peso de inercia, coeficiente de constricción, modelos de velocidad.PSO mono objetivo: convergencia, basado en subpoblaciones, PSO Binario. PSO Multiobjetivo: manejo de restricciones, entornos dinámicos, *niching*.
- Obtención de reglas de asociación y de clasificación usando Algoritmos Genéticos y PSO.

4

METODOLOGÍA DE ENSEÑANZA

El dictado de la asignatura tiene modalidad de Taller lo que permite a los alumnos aplicar las estrategias propuestas en la resolución de problemas concretos sencillos a medida que se desarrolla la teoría. Las clases son guiadas a través de la proyección de transparencias utilizando el cañón y la PC disponibles en el aula.

Muchos temas tienen una fuerte justificación matemática cuya comprensión puede facilitarse a través de representaciones gráficas o de algoritmos de aproximación. Por esta razón, la materia se dicta íntegramente en la Sala de PC.

Material del Curso y comunicación con los alumnos

Todo el material del curso estará disponible a través de la plataforma de educación a distancia WebUNLP. Tanto alumnos como docentes deberán contar con un usuario y una clave para poder acceder.

Se utilizará únicamente la cartelera disponible en WebUNLP para dar difusión a las novedades del curso.

Los alumnos podrán comunicarse con los docentes a través del servicio de mensajería provisto por la plataforma.

EVALUACIÓN

Cada alumno puede optar por una de las siguientes formas de aprobación:

a) Régimen de promoción

Se ofrecen dos modalidades para promocionar la materia. El alumno deberá optar por una de ellas al inicio del curso. Los requisitos para cada modalidad son los siguientes:

Modalidad Presencial

- Asistir al 70% de las clases teóricas y prácticas.
- Entregar un conjunto de actividades prácticas durante el desarrollo del curso siguiendo un cronograma de fechas que será publicado, a través de la plataforma, al inicio del curso. Los alumnos podrán consultar estas actividades en los horarios de práctica.
- Quienes aprueben las actividades prácticas obtendrán la cursada de la asignatura.
- Para aprobar la materia, el alumno deberá realizar una monografía breve ampliando alguna de las técnicas vistas en clase. El objetivo del trabajo es que logre aplicarla en la resolución de un problema concreto sencillo y en base a los resultados obtenidos discuta las ventajas y desventajas de la técnica elegida.

Modalidad semi-presencial

- Los alumnos que no puedan asistir a clase deberán cumplir con 1 (un) encuentro presencial por mes donde deberán exponer mediante un coloquio su avance en los temas desarrollados en el curso.
- También deberán entregar las mismas actividades prácticas que la modalidad presencial contando con la posibilidad de realizar consultas a través de la plataforma las cuales serán atendidas dentro de las 48 hs de efectuadas. Esta modalidad de atención estará sujeta a la cantidad de alumnos que opten por realizar el curso en forma semi-presencial.

- Quienes aprueben los coloquios y las actividades prácticas obtendrán la cursada de la asignatura.
- Para aprobar la materia, el alumno deberá realizar un trabajo final que habitualmente consiste en el estudio e implementación de alguna técnica de Minería de Datos, diferente a las vistas en clase. El objetivo del trabajo es que el alumno logre aplicar la técnica analizada en la resolución de un problema concreto sencillo y en base a los resultados obtenidos elabore una monografía breve donde exponga sus conclusiones y observaciones.

Para registrar la calificación correspondiente, el alumno deberá inscribirse en una mesa de final donde expondrá a través de un coloquio, el trabajo descripto en la monografía.

El régimen de promoción sólo puede ser aplicado en cursos con una reducida cantidad de alumnos ya que el seguimiento individual de cada uno de ellos implica una fuerte inversión de tiempo por parte del docente a cargo.

b) Régimen convencional

Los alumnos que opten por el régimen convencional no tendrán la obligación de cumplir con ningún requisito de asistencia.

Al finalizar el curso el alumno deberá rendir un examen escrito referido a los aspectos prácticos de la materia. Este examen cuenta con dos recuperatorios. Quienes lo aprueben con nota mayor o igual a 4 (cuatro) puntos obtendrán la cursada de la asignatura debiendo luego rendir examen final.

Bibliografía Básica

- Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques, (Third Edition). Morgan Kaufmann. 2013. ISBN-13: 978-0123748560.
- Nong Ye . *Data Mining: Theories, Algorithms, and Examples.* CRC Press. 2013. ISBN 9781439808382
- Andries Engelbrecht. *Compuattional Intelligence. An introduction.*John Wiley & Sons. 2011. ISBN 978-0-470-03561-0

Bibliografía complementaria

- Hernández Orallo, Ramírez Quintana, Ferri Ramírez. *Introducción a la Minería de Datos.* Prentice Hall. 2004. ISBN 84-205-4091-9.
- Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques, (Third Edition). Morgan Kaufmann. 2013. ISBN-13: 978-0123814791.
- Freeman y Skapura *Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación.* Addison-Wesley/Diaz de Santos. 1993.
- David Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.Addison Wesley. 1998. ISBN 0-201-15767-5*
- Kohonen, T. *Self-Organizing Maps.* 2nd Edition. Springer. ISSN 0720-678X. 1997.
- Karray and De Silva. *Soft Computing and Intelligent Systems Design Theory, tools and Applications.* Pearson Education. 2004. ISBN 0-321-11617-8

CRONOGRAMA DE CLASES Y EVALUACIONES

Semana	Contenidos	Práctica	Actividades
1	Proceso de KDD.	Rapid Miner. Operadores. Clases Java	
2	Preprocesamiento de datos. Vista minable.		
3	Redes Neuronales. Introducción. Perceptrón	TP1 – RN formadas por una única neurona artificial	Entrega 1 RN aplicadas a predicción y clasificación
4	Combinador Lineal. Neurona no lineal.		
5	Multiperceptrón. Entrenamiento utilizando backpropagation	TP2 = RN con entrenamiento supervisado y no spervisado	
6	Técnicas de clustering. Métodos basados en árboles y en centroides. Redes Neuronales competitivas. CPN y SOM		
7	Técnicas de Optimización. Introducción	TP3 – AG aplicados a optimización de funciones, adaptación de pesos de RN y obtención de reglas.	Entrega 2 Técnicas de optimización aplicables a selección de atributos y obtención de reglas
8	Estrategias evolutivas. Algoritmos genéticos.		
9	Optimización utilizando PSO	TP4 – PSO continuo y discreto aplicado en la obtención de reglas.	
10	Obtención de reglas utilizando estrategias adaptativas		
11	Minería de Datos. Introducción. Fases del proceso KDD. Preparación de los datos.	TP5 – Arboles utilizando atributos nominales y numéricos	Entrega 3 Arboles y Reglas de clasificación
12	Arboles. Método ID3 y C4.5		
13	Reglas de asociación y de clasificación	TP6 – Métodos PRISM, PART, A Priori (y sus mejoras)	
14	Métodos alternativos de construcción de reglas		
15	Consulta y 1ra. Fecha de parcial		
16	Muestra de exámenes de la 1ra. fecha		
17	Consulta y 2ra. Fecha de parcial		
18	Muestra de exámenes de la 2da. fecha		
19	Consulta y 3ra. Fecha de parcial		
20	Muestra de exámenes de la 3da. fecha		

Contacto de la cátedra (mail, página, plataforma virtual de gestión de cursos):

La cátedra cuenta con una página web de acceso público en la siguiente dirección

http://weblidi.info.unlp.edu.ar/catedras/md_si/

Allí se indica la manera de contacto con la cátedra y la forma de acceder al material publicado en **WebUNLP.unlp.edu.ar**

Firmas del/los profesores responsables: