MINERIA DE DATOS USANDO SISTEMAS INTELIGENTES

Año 2018

Carrera/Plan:
Licenciatura en Sistemas
Licenciatura en Informática

Área: Algoritmos y Lenguajes
Año: 4º o 5º año
Régimen de Cursada: *Semestral*
Carácter: Optativa
Correlativas: Algoritmos y Estructuras de Datos – Computabilidad y Complejidad (Informática) / Fundamentos de Teoría de la Computación (Sistemas)

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.

Este curso tiene como eje central la resolución de problemas concretos utilizando estas técnicas. Se espera que el alumno adquiera los conceptos necesarios para poder analizar un problema y presentar los resultados obtenidos de una manera útil para la toma de decisiones.

OBJETIVOS GENERALES

Introducir al alumno en las técnicas de Minería de Datos. Se analizarán modelos basados en regresión, árboles, reglas, redes neuronales y técnicas de agrupamiento. Se cubrirán las distintas etapas del proceso de Extracción de Conocimiento como herramienta de ayuda a la toma de decisiones. El énfasis está puesto en la resolución de problemas de clasificación y predicción.

CONTENIDOS MINIMOS

Introducción a la Minería de Datos.
Técnicas
- Modelo de regresión y métodos bayesianos.
- Reglas de clasificación y asociación
- Arboles de decisión y sistemas de reglas
- Redes Neuronales
- Técnicas de Agrupamiento

PROGRAMA ANALÍTICO

- Introducción. Obtención de conocimiento a partir de los datos. El concepto de patrón. El proceso KDD. Fases del proceso de extracción del conocimiento. La Minería de Datos como fase del proceso KDD. Relación con otras disciplinas.
- Recuperación de información vs recuperación de datos. Proceso de recuperación de información.
- Preparación de Datos. Metadatos. Análisis de la información de entrada. Construcción y análisis de representaciones gráficas. Limpieza y transformación. Transformación y creación de atributos. Discretización y Numerización, Normlización de rango, escalado y centrado. Exploración mediante visualización y selección de datos.
- Técnicas de Minería de Datos. Extracción de Patrones. Introducción. Tareas y Métodos. Tareas predictivas y descriptivas. Aprendizaje supervisado y aprendizaje no supervisado. La Minería de Datos y el aprendizaje inductivo. Comparación de las técnicas de Minería de Datos.
- Arboles de decisión. Métricas de selección de atributos. Entropía. Ganancia de Información. Tasa de Ganancia. Indice Gini. Poda y Sobreajuste. Algoritmos Id3, C4.5 y Random Forest. Construcción de árboles para grandes volúmenes de datos.
- Reglas de clasificación. Partición vs cobertura. Métodos ZeroR, OneR, PRISM y PART. Métricas de una regla: soporte, cobertura, confianza, interés y convicción.
- Reglas de asociación. Calidad de las reglas. Algoritmo A priori. Concepto de ítem frecuente. Mejoras del algortimo a priori: FP-Growth y FP-Tree.
- Técnicas de Agrupamiento. Métricas de calidad del agrupamiento. Tipos de agrupamiento: Jerárquico, partitivo y probabilista. Medidas de distancia y de conectividad. Proceso de agrupamiento. Clustering partitivo. Algoritmo k-medias. Algoritmos de clustering jerárquicos aglomerativos y divisivos. Dendrogramas. Algoritmo probabilista EM (Expectation - Maximization)
- Redes Neuronales Feedforward. Descripción de la arquitectura. 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. Aprendizaje profundo.
- Redes Neuronales Competitivas. Técnicas de Agrupamiento partitivas. Agrupamiento utilizando redes neuronales. Red CPN y red SOM. Similitudes y diferencias con el agrupamiento producido por k-medias.
- Análisis y difusión del modelo obtenido. Evaluación de modelos. Comparación de técnicas de aprendizaje. Evaluación y mejora del modelo obtenido. Performance del modelo. Matriz de confusión. Sensibilidad, especificidad, precisión y recall. F-measure. Visualización utilizando curvas ROC.

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 COMUNICACION

Todo el material del curso estará disponible a través de la plataforma de educación a distancia *Ideas*. Se utilizará únicamente la cartelera disponible en *Ideas* 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.

ACTIVIDADES PRACTICAS

Durante el desarrollo del curso los alumnos resolverán en clase problemas concretos asociados a los temas vistos en la teoría. Además se publicarán autoevaluaciones, en forma periódica, con el objetivo de que los alumnos comprueben los conocimientos adquiridos en forma voluntaria.

También se definirá un trabajo integrador que se comenzará a desarrollar a partir de la cuarta semana de clase y que deberá ser entregado al finalizar el curso.

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 al final del curso la resolución del trabajo práctico integrador. Este trabajo podrá ser consultado en los horarios de práctica.
- Aprobar el examen que se tomará al finalizar el curso. Este examen cuenta con dos recuperatorios.
- Para promocionar la materia deberá obtener una calificación mayor o igual a 6 (seis) puntos.

Modalidad semi-presencial

- Realizar el 70% de las autoevaluaciones teórico-prácticas (5 autoevaluaciones de un total de 8).
- Entregar al final del curso la resolución del trabajo práctico integrador. De ser necesario se fijarán dos encuentros presenciales para consulta por parte del alumno.
- Aprobar el examen que se tomará al finalizar el curso. Este examen cuenta con dos recuperatorios. Para promocionar la materia deberá obtener una calificación mayor o igual a 6 (seis) puntos.

b) Régimen convencional

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

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.

BIBLIOGRAFIA BASICA

- Hernández Orallo, Ramírez Quintana, Ferri Ramírez. *Introducción a la Minería de Datos*. Prentice Hall. 2004. ISBN 84-205-4091-9.
- Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques, (Fourth Edition). Morgan Kaufmann. 2017. ISBN 978-0-12-804291-5.
- Nong Ye . *Data Mining: Theories, Algorithms, and Examples.* CRC Press. 2013. ISBN 9781439808382

BBLIOGRAFIA COMPLEMENTARIA

- 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.
- 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.* Peason Education. 2004. ISBN 0-321-11617-8

CRONOGRAMA DE CLASES Y EVALUACIONES

Semana	Fecha	Detalle
1	07/03/18	Introducción. Proceso KDD. Minería de Datos. Aplicaciones relacionadas. RapidMiner.
2	14/03/18	Metadatos.. Representación gráfica de atributos.
3	21/03/18	Transformación de atributos
4	28/03/18	Análisis de atributos. Correlación y regresión lineal
5	04/04/18	Arboles de Clasificación y de Regresión
6	11/04/18	Reglas de clasificación
7	18/04/18	Reglas de asociación.
8	25/04/18	Técnicas de Agrupamiento
9	02/05/18	Minería de Textos
10	09/05/18	Redes Neuronales. Perceptrón.
11	16/05/18	Multiperceptrón
12	23/05/18	Agrupamiento con Redes Neuronales Competitivas.
13	30/05/18	Consultas referidas al trabajo integrador y a la 1era. Fecha de parcial
14	06/06/18	1ra. Fecha de parcial
15	13/06/18	Muestra de exámenes de la 1ra. Fecha. Consultas referidas al trabajo integrador y a la 2da. Fecha de parcial
16	22/06/18	2da. Fecha de parcial
17	29/06/18	Muestra de exámenes de la 2da. Fecha. Consultas referidas al trabajo integrador y a la 3era. Fecha de parcial
18	06/07/18	3ra. Fecha de parcial
19	13/07/18	Muestra de exámenes de la 3da. fecha

Evaluaciones previstas	Fecha
1ra. Fecha de Parcial	06/06/18
2da. Fecha de Parcial	22/06/18
3ra. Fecha de Parcial	06/07/18
Entrega del Trabajo Integrador a través del EVEA IDEAS	Hasta el 15/07/18

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 el entorno virtual de enseñanza y aprendizaje **IDEAS**

Dra. Laura Lanzarini