

ASIGNATURA DE GRADO: MINERÍA DE DATOS (ING.TI)

UNED

Curso 2017/2018

(Código de asignatura : 71024062)

NOMBRE DE LA ASIGNATURA	MINERÍA DE DATOS (ING.TI)
CÓDIGO	71024062
CURSO ACADÉMICO	2017/2018
DEPARTAMENTO	INTELIGENCIA ARTIFICIAL
TÍTULOS EN QUE SE IMPARTE	GRADO EN INGENIERÍA EN TECNOLOGÍAS DE LA INFORMACIÓN (grado seleccionado) GRADO EN INGENIERÍA INFORMÁTICA
CURSO	CUARTO CURSO
TIPO	OPTATIVAS
Nº ECTS	6
HORAS	150
PERIODO	SEMESTRE 1
IDIOMAS EN QUE SE IMPARTE	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

Minería de Datos es sólo una de las denominaciones (la más popular, quizás, en el ámbito empres) de un área de investigación que podríamos llamar con más propiedad, Descubrimiento de Conocimiento a partir de datos. Corresponde con lo que desde antiguo se conoce como el principio de inducción en términos filosóficos.

¿En qué consiste entonces la Minería de Datos? Se trata de conseguir reproducir con computadoras, tareas genuinamente humanas relacionadas con la extracción de conocimiento a partir de datos. Esas tareas pueden ser de varios tipos. Uno de ellos agrupa tareas en las que la computadora debe aprender a partir de un conjunto de ejemplos, generalizar las relaciones entre ellos, y aplicar el modelo resultante del aprendizaje a datos nuevos. La clasificación de casos en categorías responde bien a este patrón de tareas, pero también las actividades de control, en las que la máquina debe aprender a (generar un modelo para) controlar un sistema con unos objetivos explícitos, en problemas de planificación o de asignación de recursos, o las tareas de predicción, en las que el modelo aprendido a partir de los datos nos ayuda a inferir nuevos valores de unas variables desconocidas.

En todos estos casos, vemos la importancia que desempeñan los datos en este área. Se trata de producir modelos a partir de ejemplos que condensan el conocimiento que queremos aprehender, y para los que no disponemos de un modelo de conocimiento alternativo, expresado en lenguaje natural o estructurado.

Otro tipo de tareas encuadradas en la Minería de Datos, pero que no veremos en este curso, abordan la tarea de descubrir conceptos, relaciones o reglas en conjuntos de datos no etiquetados. Mientras que en el caso anterior (tareas de clasificación o regresión) disponemos de ejemplos que expresan los modelos que deseamos inferir, en las tareas de este tipo los datos están desnudos, y nuestra tarea consiste precisamente en descubrir esquemas clasificatorios, patrones repetidos, relaciones entre ellos, agrupamientos de datos o reglas que describan la distribución

de los datos en un espacio de representación dado.

En este curso vamos a abordar los fundamentos del área. El objetivo del equipo docente ha sido, no abordar de manera extensiva pero superficial las diversas técnicas que se aplican en el área, sino proporcionar al estudiante los fundamentos que le permitan explorar en asignaturas sucesivas o por su cuenta, todas esas técnicas en las que aquí no podremos profundizar. Así pues, empezamos la casa por sus cimientos. Y los cimientos de el edificio de la Minería de Datos son principalmente matemáticos y probabilísticos.

Esta asignatura, como podréis comprobar en la memoria de la titulación, se corresponde con la materia denominada Sistemas de Información, que comparte con las asignaturas de Bases de Datos y Gestión de Bases de Datos. Para su aprendizaje no es estrictamente necesario haber cursado las anteriores, pues lo que aquí se enseña se hace de manera independiente del sistema de almacenamiento de los datos. Sin embargo, sí es muy importante haber cursado las asignaturas de Fundamentos Matemáticos y Estadística.

Los conocimientos adquiridos a través de esta asignatura son los fundamentos de un área cuya exploración continúa en el Master de Inteligencia Artificial Avanzada, en las asignaturas relacionadas con la Minería de Datos. En ellas, se aplican todo lo aprendido aquí para entender las variadas técnicas avanzadas (como Máquinas de Vectores Soporte, Procesos Gaussianos, Redes Neuronales Artificiales, etc) y para adentrarnos en el mundo de la clasificación no supervisada o agrupamiento.

Existen multitud de vías en las que los conocimientos adquiridos aquí serán de utilidad en el futuro de los estudiantes. El aprendizaje estadístico (otra de las denominaciones de la Minería de Datos) abre un sinfín de perspectivas nuevas en una nueva era en la que los datos, en muchas ocasiones, desbordan la capacidad de los humanos de procesar información. Desde lo que se conoce como el cuarto paradigma de la Ciencia (o e-Ciencia, en una expresión poco afortunada) de aplicación en áreas como las bio-tecnologías o las grandes bases de datos científicas, a las aplicaciones empreses en bancos o librerías virtuales, sin olvidar a los buscadores web. Lo que aquí aprenderemos es de aplicación general a todos esos campos, precisamente porque se trata de los fundamentos del área.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR LA ASIGNATURA

Es necesario tener conocimientos básicos de Matemáticas (Análisis y Álgebra Matricial) y Estadística, adquiridos a través de las asignaturas de Fundamentos Matemáticos y Estadística.

EQUIPO DOCENTE

Nombre y Apellidos	LUIS MANUEL SARRO BARO
Correo Electrónico	lsb@dia.uned.es
Teléfono	91398-8715
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
DEPARTAMENTO	INTELIGENCIA ARTIFICIAL
Nombre y Apellidos	JOSE LUIS AZNARTE MELLADO
Correo Electrónico	jlaznarte@dia.uned.es
Teléfono	91398-9688

HORARIO DE ATENCIÓN AL ESTUDIANTE Y TUTORIZACIÓN

1. Equipo docente (en la sede central):

Dr. D Luis Manuel Sarro Baro

Guardia: Lunes de 10 a 14 h

Horario de atención al estudiante:

Lunes 14:00 a 16:00 y Miércoles de 10:00 a 16:00h. Despacho 3.12. Tel.: 913988715. lsb@dia.uned.es

La dirección de contacto es:

ETSI Informática-UNED. Dpto. Inteligencia Artificial

c/Juan del Rosal, 16

28040 Madrid

2. Profesores tutores (en el centro asociado correspondiente). Los horarios de atención del tutor serán suministrados por los propios centros asociados al inicio de curso.

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.

RESULTADOS DE APRENDIZAJE

El estudiante, al concluir y aprobar la asignatura, dominará los conceptos básicos del área de la Minería de Datos: el análisis probabilístico del problema de reconocimiento de patrones en tareas de clasificación y regresión; su tratamiento desde la perspectiva bayesiana, con el manejo de las entidades que desempeñan papeles relevantes en ese ámbito (verosimilitud, probabilidades marginales, evidencias...); los problemas de la selección de modelos de complejidad creciente, y la maldición de la dimensionalidad. Habrá, asimismo, adquirido destreza en el manejo de probabilidades y de entidades algebraicas, principalmente matriciales, suficiente como para profundizar los contenidos de la materia contenidos en los ejercicios no resultados que propondrá el equipo docente.

En resumen, el estudiante habrá interiorizado los fundamentos del área a través de las aproximaciones más simples (los modelos lineales) y se hallará en disposición de abordar el estudio posterior de las técnicas más avanzadas (y no necesariamente a través de modelos lineales) como las Máquinas de Vectores Soporte, los Procesos Gaussianos o las Redes Neuronales (por citar sólo tres ejemplos)

CONTENIDOS

Tema 1: Introducción al Aprendizaje Estadístico

Tema 2: Modelos Lineales de Regresión

Tema 3: Modelos lineales de clasificación

METODOLOGÍA

La asignatura se cursa de una manera clásica en la educación a distancia. Al tratarse de una asignatura de fundamentos, se concede una importancia especial a los aspectos teóricos, y las actividades prácticas están supeditadas a la consolidación de los conceptos teóricos.

La interacción con el equipo docente se realizará principalmente a través de la plataforma de aprendizaje virtual de la

asignatura, donde se pretende que sean los propios alumnos los que resuelvan sus dudas (con la ayuda y supervisión en todo momento del profesor) de manera colaborativa. Todo ello, desde la convicción de que lo que se descubre se aprende mucho mejor que lo que se asimila de forma pasiva. El equipo docente valorará muy positivamente la participación en los foros con mensajes que colaboren en la resolución de problemas y dudas.

El 5% de los créditos asignados se destina a la preparación para el estudio del contenido teórico, lo que incluye la lectura de las orientaciones y una primera lectura del índice del texto base.

El segundo bloque, el más importante en cuanto a fracción del total (80%) lo constituye el estudio de los contenidos teóricos (60%) y el desarrollo de ejercicios de consolidación de lo aprendido (20%) mediante la resolución de problemas propuestos cuya respuesta estará a disposición de los alumnos (ejercicios de auto-evaluación).

Finalmente, el tercer bloque se asigna a una práctica entregable a la que corresponde el 10% de la nota (y el 15% de los créditos de la asignatura). Consistirá en la resolución de tres de los ejercicios propuestos en el texto base pero cuya solución no está disponible a través de Internet (ejercicios de descubrimiento).

SISTEMA DE EVALUACIÓN

PRUEBA PRESENCIAL

Tipo de examen	Examen de desarrollo
Preguntas desarrollo	
Duración del examen	120 (minutos)
Material permitido en el examen	El texto base de la asignatura
Criterios de evaluación	El 90% de la nota final corresponde a la nota obtenida en la prueba presencial. Sea NPP la nota de la prueba presencial calificada de 0 a 10. Los enunciados del examen se corresponderán con algunos de los ejercicios de descubrimiento propuestos a comienzo de curso por el equipo docente. Por lo tanto, no habrá preguntas que exijan la memorización de demostraciones. Si es necesario, el equipo docente proporcionará definiciones complejas que sean necesarias para la resolución de los enunciados de examen. Consideramos que lo más importante no es la memorización de fórmulas sino la comprensión de los conceptos clave del área y la adquisición de las destrezas algebraicas necesarias para manipular fórmulas.
% del examen sobre la nota final	0
Nota del examen para aprobar sin PEC	0
Nota máxima que aporta el examen a la calificación	0

final sin PEC

Nota mínima 0
en el examen
para sumar la
PEC

Comentarios y
observaciones

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

Descripción El equipo docente propondrá cada año dos subconjuntos de ejercicios tomados del texto base. El primero estará compuesto de ejercicios cuya respuesta se encuentra disponible en el sitio web del libro (auto-evaluación); el segundo contendrá ejercicios cuya respuesta no está disponible a través de Internet y de entre los que se seleccionarán los enunciados del examen (descubrimiento).

El 10% de la nota final de la asignatura corresponderá a la evaluación por parte de los tutores o equipo docente de un conjunto de 3 ejercicios de descubrimiento elegidos por el estudiante de entre los propuestos por el equipo docente. Sea NED la nota de 0 a 10 asignada a estos ejercicios de descubrimiento. Entonces, la nota combinada NC será igual a $0.9 \cdot NPP + 0.1 \cdot NED$.

**Criterios de
evaluación**

Ponderación 0
de la PEC en
la nota final

Fecha
aproximada
de entrega

Comentarios y
observaciones

OTRAS ACTIVIDADES EVALUABLES

Descripción Un último tipo de tarea incluye la realización de una práctica de experimentación numérica evaluable. Dicha práctica será evaluada por los tutores, y podrá suponer hasta 2 puntos sobre 10 en la nota final. La nota final se calculará sumando a la nota combinada NC la puntuación de la práctica (NPEN) siempre y cuando ésta última supere los 4 puntos sobre 10. Si la suma de ambas notas supera los 10 puntos, la nota evidentemente será de 10. Su objetivo (el de la práctica evaluable) es facilitar que el alumno adquiera familiaridad con los casos prácticos de experimentación numérica a los que se les aplica todo el bagaje conceptual adquirido durante el curso. El enunciado de la práctica se hará público cada año a comienzo de curso.

**Criterios de
evaluación**

Ponderación 0
en la nota
final

Fecha
aproximada
de entrega

Comentarios y
observaciones

¿Cómo se obtiene la nota final?

La nota final se obtendrá como la suma $0.9 \cdot \text{NPP} + 0.1 \cdot \text{NED} + \text{NPEN}$ o 10 en el caso de que la suma supere la nota máxima (10).

BIBLIOGRAFÍA BÁSICA

ISBN(13): 9780387310732

Título: PATTERN RECOGNITION AND MACHINE LEARNING

Autor/es: Christopher M. Bishop ;

Editorial: Springer

Buscarlo en Editorial UNED

Buscarlo en librería virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

BIBLIOGRAFÍA COMPLEMENTARIA

ISBN(13): 9780387848587

Título: THE ELEMENTS OF STATISTICAL LEARNING

Autor/es: Hastie, Trevor ; Tibshirani, Robert J. ; Friedman, Jerome ;

Editorial: Springer

Buscarlo en librería virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

Buscarlo en Catálogo del Patrimonio Bibliográfico

RECURSOS DE APOYO Y WEBGRAFÍA

El curso se desarrolla fundamentalmente a través de la plataforma aLF de la UNED. La información y el material complementario se encuentra en dicho curso, y la interacción con el equipo docente se desarrollará principalmente a través de los foros de la asignatura en dicha plataforma. Por supuesto, el equipo docente también atenderá a los alumnos a través del teléfono (913988715) o de manera presencial en el horario de guardia (Lunes de 10:00 a 14:00).

Es recomendable acordar una cita previamente. Finalmente, el equipo docente estará también disponible a través de software de videoconferencia, preferiblemente skype. De nuevo, será necesario concertar una cita con anterioridad.

El equipo docente será el responsable de responder a las dudas que surjan sobre el funcionamiento de la asignatura y sobre los contenidos teórico de ésta (siempre que sea posible, a través de los foros, pues de esta manera las respuestas quedan a disposición de otros alumnos que puedan compartirlas). En principio, y salvo circunstancias excepcionales que lo impidan, el tiempo máximo de espera para las respuestas a las preguntas del foro es de 7 días, (el tiempo entre guardia y guardia). Por regla general, nunca se alcanza ese periodo y en la medida de lo posible el equipo docente intenta responder a las cuestiones con la máxima celeridad que permiten las otras obligaciones del profesorado, entre las que cabe destacar la docencia en otras asignaturas y las tareas de investigación y administración. Como orientación, se puede decir que en periodo lectivo, fuera de épocas de examen o viajes al extranjero para reuniones o congresos, el tiempo de espera no debe rebasar las 48 horas.

Es muy importante que el alumno que solicite una respuesta directa del equipo docente lo haga constar en su mensaje al foro.