20-21

MÁSTER UNIVERSITARIO EN
TECNOLOGÍAS DEL LENGUAJE

GUÍA DE ESTUDIO PÚBLICA

MINERÍA DE DATOS

CÓDIGO 31101061

UNED

20-21

MINERÍA DE DATOS

CÓDIGO 31101061

ÍNDICE

Nombre de la asignatura	MINERÍA DE DATOS
Código	31101061
Curso académico	2020/2021
Título en que se imparte	MÁSTER UNIVERSITARIO EN TECNOLOGÍAS DEL LENGUAJE
Tipo	CONTENIDOS
Nº ETCS	6
Horas	150.0
Periodo	ANUAL
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

El presente curso pretende dar una visión panorámica de la teoría y conceptos fundamentales utilizados en Minería de Datos (MD), del conjunto de tareas abordadas por esta disciplina y del repertorio de técnicas y métodos existentes que permiten resolver cada una de estas tareas.
Ficha técnica:

•Tipo: Optativa

•Duración: Anual

•Créditos Totales y Horas: 6 / 150

•Horas de estudio teórico: 55

•Horas de trabajo práctico: 50

•Horas de actividades complementarias: 45

La asignatura Minería de Datos se imparte tanto en el Máster Universitario en Investigación en Inteligencia Artificial como en el Master Universitario en Tecnologías del Lenguaje de la ETSI Informática de la UNED, en ambos como optativa. Esta asignatura es de carácter anual con una carga lectiva de 6 ECTS.
Existen distintas asignaturas en ambos másteres relacionadas con esta asignatura. Así, "Métodos de Aprendizaje en IA" aborda, además de otras técnicas de aprendizaje, la mayoría de las técnicas que se estudiarán en este tema y que básicamente se encuadran dentro del denominado paradigma de aprendizaje inductivo. El alumno que haya cursado dicha asignatura tendrá mucho camino adelantado al abordar esta asignatura. No obstante, hay que tener en cuenta que la visión que allí se da está orientada eminentemente a la parte algorítmica y de implementación (programación) de cada técnica. Aquí, el enfoque está más orientado a su uso, independientemente de la implementación particular. Es decir, consideraremos el conjunto de técnicas como una biblioteca de componentes reutilizables, cada uno de los cuales será seleccionado de acuerdo a las características de la tarea que se requiere resolver. En otros casos, esta asignatura puede servir de introducción a otras asignaturas de ambos másteres, tales como "Descubrimiento de información en textos" o "Minería en la Web".

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

El alumno debe haber cursado las asignaturas de Fundamentos Matemáticos de la Informática y Estadística impartidas en el primer ciclo de la titulación de Informática de la UNED o asignaturas equivalentes en otras universidades.
En particular, debe haber adquirido competencias básicas en el manejo algebraico de matrices, cálculo de determinantes, inversión de matrices y diagonalización de éstas. Debe conocer el cálculo de las derivadas parciales e integrales de funciones multivariantes (Análisis Matemático). Finalmente, debe conocer conceptos básicos de Estadística como las propiedades de la distribución gaussiana multivariante o los tests estadísticos de contraste de hipótesis.

EQUIPO DOCENTE

Nombre y Apellidos	LUIS MANUEL SARRO BARO (Coordinador de asignatura)
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
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	INTELIGENCIA ARTIFICIAL

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los alumnos se llevará a cabo exclusivamente a través de la plataforma de e-learning Alf.
Los horarios de los profesores son:
Luis M. Sarro Baro
Guardia: Lunes, de 10:00 a 14:00
José Luis Aznarte
Guardia: lunes, de 16:00 a 20:00

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

MÁSTER UNIVERSITARIO EN LENGUAJES Y SISTEMAS INFORMÁTICOS

Competencias Básicas:

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Competencias Generales:

CPG1 - Adquirir capacidad de abstracción, análisis, síntesis y relación de ideas.

CPG2 - Adquirir capacidad crítica y de decisión

CPG3 - Adquirir capacidad de estudio y autoaprendizaje

CPG4 - Adquirir capacidad creativa y de investigación

CPG5 - Adquirir habilidades sociales para el trabajo en equipo

Competencias Específicas:

CE1 - Adquirir capacidad de comprender y manejar de forma básica los aspectos más importantes relacionados con los lenguajes y sistemas informáticos en general y, de manera especial, en los siguientes ámbitos: Tecnologías del lenguaje y de acceso a la información en web

CE2 - Adquirir capacidad de comprender y manejar de forma básica los aspectos más importantes relacionados con los lenguajes y sistemas informáticos en general y, de manera especial, en los siguientes ámbitos: Tecnologías de enseñanza, aprendizaje, colaboración y adaptación

CE3 - Adquirir capacidad de estudio de los sistemas y aproximaciones existentes y para distinguir las aproximaciones más efectivas.

CE4 - Adquirir capacidad para detectar carencias en el estado actual de la ciencia y la tecnología

CE5 - Adquirir capacidad para proponer nuevas aproximaciones que den solución a las carencias detectadas.

CE6 - Adquirir capacidad de especificar, diseñar, implementar y evaluar tanto cualitativa como cuantitativamente los modelos y sistemas propuestos.

CE7 - Adquirir capacidad para proponer y llevar a cabo experimentos con la metodología adecuada como para poder extraer conclusiones y determinar nuevas líneas de actuación e investigación.

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL AVANZADA: FUNDAMENTOS, MÉTODOS Y APLICACIONES

Competencias Básicas:

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Competencias Generales:

CG1 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.

CG2 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.

CG3 - Que los estudiantes sepan comunicar sus conclusiones -y los conocimientos y razones últimas que las sustentan- a públicos especializados y no especializados de un modo claro y sin ambigüedades.

CG4 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Competencias Específicas:

CE1 - Conocer los fundamentos de la Inteligencia Artificial y las fronteras actuales en investigación.

CE2 - Conocer un conjunto de métodos y técnicas tanto simbólicas como conexionistas y probabilistas, para resolver problemas propios de la Inteligencia Artificial.

CE3 - Conocer los procedimientos específicos de aplicación de estos métodos a un conjunto relevante de dominio (educación, medicina, ingeniería, sistemas de seguridad y vigilancia,

etc.), que representan las áreas más activas de investigación en IA.

RESULTADOS DE APRENDIZAJE

Destrezas y competencias

•Conocer las relaciones existentes de la MD (Minería de Datos) con otras disciplinas.
•Conocer las distintas fases implicadas en un proyecto de minería de datos y las relaciones existentes entre ellas.
•Conocer y saber aplicar algunas de las técnicas más relevantes en MD para realizar preparación de datos.
•Distinguir entre tarea, técnica y método en MD.
•Saber relacionar las distintas tareas propias de MD con las técnicas que permiten resolverlas.
•Conocer algunas de las técnicas más relevantes en MD. Dominar, tanto desde un punto de vista teórico como práctico, estas técnicas/algoritmos.
•Aplicar técnicas de evaluación adecuadas en función del tipo de modelo a evaluar.
•Conocer al menos uno de los entornos de desarrollo/lenguajes de programación más habituales en MD.
•Conocer las repercusiones de la MD en distintos campos: social, legal y ético.

CONTENIDOS

Tema 1: Introducción

El carácter introductorio de este tema tiene como principal objetivo dar una panorámica general de los distintos aspectos relacionados con la minería de datos (MD). Este objetivo global se concreta en que el alumnado debe ser capaz de:

•Conocer los distintos tipos de datos que se manejan en MD.
•Conocer los distintos tipos de modelos que se pueden aprender.
•Conocer la relación de la MD con otras disciplinas.
•Conocer los diferentes dominios de aplicación de la MD.
•Relacionar el concepto de Descubrimiento de Conocimiento a partir de Datos con el de Minería de Datos.
•Conocer las distintas fases implicadas en el proceso de descubrimiento de conocimiento a partir de datos.

Tema 2: Aprendizaje supervisado

En este tema haremos una primera aproximación a las técnicas más básicas del aprendizaje supervisado, incluyendo:

•Tipos de variables y terminología
•Dos aproximaciones simples al problema de la predicción: Mínimos cuadrados y vecinos más cercanos.
•Modelos Lineales y Mínimos Cuadrados.
•Modelos basados en los vecinos más cercanos
•Teoría (estadística) de la decisión.
•Métodos locales en espacios de alta dimensionalidad

Tema 3: Selección de características

En este tema se abordan distintas técnicas para establecer e incrementar la calidad de un conjunto de datos cualquiera. Entre otros asuntos, se abordan los siguientes:

•Conceptos básicos
•Rankings de variables
•Criterios basados en correlación
•Criterios basados en Teoría de la Información
•Clasificadores unidimensionales
•Ejemplos ilustrativos: relevancia y redundancia.
•Selección de subconjuntos de variables
•Métodos de envoltura y embebidos
•Métodos anidados
•Optimización directa
•Filtros para selección de subconjuntos
•Creación de atributos y reducción de la dimensionalidad
•Agrupamiento
•Factorización matricial
•Selección supervisada de características
•Métodos de validación

Tema 4: Evaluación y selección de modelos

Este tema estudia las diferentes maneras de evaluar y comparar técnicas y modelos de minería de datos:

•Sesgo, varianza y complejidad de un modelo.
•La descomposición sesgo-varianza
•El optimismo y la tasa de error en entrenamiento
•Estimaciones de la tasa de error en muestras.
•El número efectivo de parámetros de un modelo
•La aproximación bayesiana a la evaluación de modelos y el BIC

•Validación cruzada
•Métodos Bootstrap

Tema 5: Redes neuronales artificiales (el perceptrón multicapa)

En este tema se introduce uno de los modelos más populares dentro de la minería de datos: las redes neuronales artificiales. Entre otros, se cubren los siguientes asuntos:
•Conceptos básicos de redes neuronales. El perceptrón multicapa
•Entrenamiento de redes neuronales
•inicialización de pesos
•El sobreajuste
•Escalado de las entradas
•Elección de la arquitectura: capas ocultas y sus dimensiones
•Multimodalidad del espacio de parámetros
•Conjuntos (*ensembles*) de redes neuronales: métodos bayesianos, boosting y bagging.
•Comparación de modelos basados en redes neuronales.

Tema 6: Aprendizaje profundo (CNNs)

En este tema se presenta una de las arquitecturas más populares del llamado "aprendizaje profundo" para redes neuronales artificiales: las redes neuronales convolucionales. Este modelo extiende y amplía los conceptos introducidos con el perceptrón multicapa, y cubre, entre otros los siguientes asuntos:
•La operación de convolución
•La operación de *pooling*
•Interpretación de la convolución+*pooling* en términos de distribuciones a priori muy informativos
•Variaciones sobre la arquitectura básica
•Salidas estructuradas
•Adaptaciones a diferentes tipos de datos
•Algoritmos eficientes de convolución
•Características aleatorias o no supervisadas
•Bases neurofisiológicas de las redes convolucionales

Tema 7: Bosques aleatorios

En este tema se estudiarán los bosques aleatorios (*random forests*), sus principios y fundamentos, cómo evaluarlos y los peligros que se deben evitar al entrenar este tipo de modelos. Se estructura en los siguientes contenidos:

•Definición de un bosque aleatorio
•Muestras out-of-bag
•Importancia de las variables predictoras a partir de un bosque aleatorio
•Gráficos de proximidad
•Sobreajuste en los bosques aleatorios
•Varianza y decorrelación en los bosques aleatorios
•Sesgos
•Vecinos más cercanos adaptativos

Tema 8: Consecuencias éticas y sociales

En este tema, que no por ser el último debe ser visto como menos importante, abordaremos algunas consideraciones éticas y sociales acerca del uso de las técnicas y métodos vistos anteriormente. Los contenidos serán:
•Introducción
•Las 5 Cs: Consentimiento, claridad, consistencia, control y consecuencias
•Ética y seguridad en la creación de modelos
•Principios guía en el desarrollo de aplicaciones
•Cómo introducir la ética en una sociedad dominada por los datos
•Leyes y reglamentos

METODOLOGÍA

La metodología será la general del Máster, adaptada a las directrices del EEES, de acuerdo con el documento del IUED. Junto a las actividades y enlaces con fuentes de información externas, existe material didáctico propio preparado por el equipo docente. La asignatura no tiene clases presenciales. Los contenidos teóricos se impartirán a distancia, de acuerdo con las normas y estructuras de soporte telemático de la enseñanza en la UNED.
En particular, en la asignatura se abordarán de manera secuencial las diversas fases del proceso de descubrimiento de conocimiento desde el punto de vista algorítmico, de manera que es conveniente seguir los contenidos de manera igualmente secuencial. Algunos temas vienen acompañados de una o varias actividades cuya memoria servirá de base para la evaluación. Recomendamos leer primero los contenidos teóricos de cada tema (y especificos de cada actividad) antes de abordar las actividades.
No es necesario memorizar expresamente los contenidos del temario (no hay examen presencial de la asignatura), pero el equipo docente hará especial énfasis en la comprensión de los contenidos mostrada en las actividades. Éstas están diseñadas de manera que el/la estudiante debe realizar una tarea importante de contextualización y análisis. Si el/la estudiante se limita a generar resultados sin demostrar la comprensión de los conceptos en la discusión de dichos resultados se considerará que la práctica es insuficiente.

SISTEMA DE EVALUACIÓN

TIPO DE PRIMERA PRUEBA PRESENCIAL

Tipo de examen: No hay prueba presencial

TIPO DE SEGUNDA PRUEBA PRESENCIAL

Tipo de examen2: No hay prueba presencial

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad: No

Descripción

En esta asignatura no hay examen.

Criterios de evaluación

Ponderación de la prueba presencial y/o los trabajos en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC?: Si,PEC no presencial

Descripción

Para la mayor parte de los temas del curso, se propondrán actividades prácticas (un mínimo de 4) en las que el alumnado tendrá que demostrar que ha comprendido la teoría y que ha adquirido las destrezas básicas para poner en práctica esos conocimientos en un marco operativo.

Criterios de evaluación

Cada una de las actividades prácticas será evaluada de 0 a 10 puntos de acuerdo a una rúbrica previamente conocida por los alumnos y teniendo en cuenta particularmente si las argumentaciones que acompañen los experimentos permiten demostrar que el alumnado ha interiorizado los contenidos propuestos en cada actividad.

Ponderación de la PEC en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

OTRAS ACTIVIDADES EVALUABLES

¿Hay otra/s actividad/es evaluable/s?: Si,no presencial

Descripción

Será tenida en cuenta la participación del alumnado en los foros de la asignatura, particularmente si se trata de aportes relevantes acerca de los temas tratados y si son hechos con criterios de colaboración (se espera que el alumnado no solo exponga dudas y preguntas, sino que también participe en la indagación colectiva de las dudas y preguntas del resto).

Criterios de evaluación

Ponderación en la nota final	Podrá añadir hasta un punto en la calificación final.
Fecha aproximada de entrega	
Comentarios y observaciones	

¿CÓMO SE OBTIENE LA NOTA FINAL?

La calificación final se obtendrá como la media de las calificaciones de cada una de las actividades entregables, más hasta un punto por la participación del alumnado en los foros.

BIBLIOGRAFÍA BÁSICA

ISBN(13):9780262035613
Título:DEEP LEARNING
Autor/es:Ian Goodfellow ; Aaron Courville ; Yoshua Bengio ;
Editorial:THE MIT PRESS

ISBN(13):9780387848587
Título:THE ELEMENTS OF STATISTICAL LEARNING
Autor/es:Hastie, Trevor ; Tibshirani, Robert J. ; Friedman, Jerome ;
Editorial:Springer

El material docente del presente curso está compuesto por los dos libros indicados en la bibliografía básica (que están disponibles para su libre descarga) más el artículo *An Introduction to Variable and Feature Selection* de Isabelle Guyon y André Elisseeff, publicado en el Journal of Machine Learning Research, 3 (2003).

BIBLIOGRAFÍA COMPLEMENTARIA

Materiales y recursos de apoyo

De manera general, las prácticas se realizarán con el lenguaje R, aunque si alguien desea hacerlas con python u otro, podrá plantearlo al equipo docente.

Los ficheros con los datos de trabajo serán proporcionados por el equipo docente a través de la plataforma aLF o formarán parte de la distribución del software empleado. Si no se indica que la actividad correspondiente haya de ser realizada con un conjunto de datos particular, el alumno podrá elegir un fichero de casos del repositorio de la Universidad de California Irvine http://kdd.ics.uci.edu/ u otro.

La plataforma aLF proporcionará el adecuado interfaz de interacción entre el alumno y sus profesores. Esta plataforma colaborativa permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.

Se ofrecerán las herramientas necesarias para que, tanto el equipo docente como el alumnado, encuentren la manera de compaginar tanto el trabajo individual como el aprendizaje cooperativo.

BIbliografia complementaria de consulta

•C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

•H. Witten, E. Frank, Data mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann Publishers, 2005.

•The incredible potential and dangers of data mining health records. Matt McFarland. The Washington Post, October 1, 2014. https://www.washingtonpost.com/news/innovations/wp/2014/10/01/the-incredible-potential-and-dangers-of-data-mining-health-records/?noredirect=on&utm_term=.5d94f0759c37

RECURSOS DE APOYO Y WEBGRAFÍA

Ver la sección Comentarios y anexos de Bibliografía complementaria.

IGUALDAD DE GÉNERO

En coherencia con el valor asumido de la igualdad de género, todas las denominaciones que en esta Guía hacen referencia a órganos de gobierno unipersonales, de representación, o miembros de la comunidad universitaria y se efectúan en género masculino, cuando no se hayan sustituido por términos genéricos, se entenderán hechas indistintamente en género femenino o masculino, según el sexo del titular que los desempeñe.