Proyecto docente

Asignatura	Knowledge Discovery/Aprendizaje no supervisado		
Materia	Ciencia de datos/Data Science		
Titulación	Máster Universitario en Inteligencia de Negocio y Big Data en Entornos Seguros		
Plan		**Código**	
Periodo de impartición	Segundo semestre	**Tipo/Carácter**	Obligatoria
Nivel/Ciclo	Máster	**Curso**	1
Créditos ECTS	3		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	César Ignacio García Osorio, José Francisco Diez Pastor		
Datos de contacto (e-mail, teléfono…)	cgosorio@ubu.es, jfdpastor@ubu.es 947259358		
Horario de tutorías	Previa solicitud		
Coordinador	César Ignacio García Osorio		
Departamento	Área de Lenguajes y Sistemas Informáticos, Ingeniería Civil, Universidad de Burgos		
Web			
Descripción General	El aprendizaje automático no supervisado es la tarea de aprendizaje automático de inferir una función para describir la estructura oculta a partir de datos «no etiquetados». Algunas de las tareas típicas de este tipo de aprendizaje son la agrupación e identificación de datos similares, o la reducción de dimensionalidad y la visualización.		

1. Situación / Sentido de la asignatura

1.1 Contextualización

Aunque es frecuente asociar el aprendizaje automático con la extracción de conocimiento de conjuntos de datos previamente etiquetados, existen varias tareas interesantes que se pueden llevar a cabo con conjuntos de datos no etiquetados, algunos ejemplos:

- La búsqueda de elementos similares, por ejemplo, mirar una colección de páginas Web y encontrar páginas casi duplicadas. Estas páginas podrían ser plagios, por ejemplo, o podrían ser réplicas que tienen casi el mismo contenido. En el contexto del *big data* se han propuesto métodos eficientes para hacer este tipo de comprobaciones utilizando el concepto de *locally sensitive hashing*.
- La agrupación o *clustering*, que es el proceso de examinar una colección de instancias o «puntos» y agrupar aquellos que forman «conglomerados» de acuerdo con alguna medida de distancia. El objetivo es que los puntos en el mismo grupo tengan una pequeña distancia el uno del otro, mientras que los puntos en diferentes grupos se encuentren a una gran distancia el uno del otro.
- La reducción de la dimensionalidad, consiste en encontrar una representación más compacta, con menos atributos, de un conjunto de datos que inicialmente está caracterizado por un gran número de atributos. La idea es que la nueva representación conserve las principales características estadísticas del conjunto original. Si es posible reducir el conjunto de características a sólo 2 o 3, entonces se hace posible visualizar el conjunto de datos de partida para el estudio de sus características. Además, la forma en la que las características de partida se han combinado para dar lugar al conjunto reducido también puede aportar información interesante sobre el conjunto de datos.
- Descubrimiento de conjuntos de elementos frecuentes. Este problema a menudo se considera como el descubrimiento de «reglas de asociación», que a su vez pueden utilizarse como una de las formas de abordar la construcción de sistemas de recomendación, en los que se pueden sugerir compra de nuevos productos utilizando el historial de comprar de los clientes.

1.2 Relación con otras asignaturas

La asignatura está relacionada principalmente con las de la materia a la que pertenece: Ciencia de Datos. Estas asignaturas son:

- Modelos de Programación para el Big Data
- Aprendizaje sobre flujos de datos

La relación con la primera está en que para los algoritmos explicados en esta asignatura existen implementaciones específicas en bibliotecas como las explicadas en esa asignatura. La relación con la segunda, está en que algunos de los algoritmos explicados en esta asignatura tienen su correspondiente versión en el contexto del análisis de flujos de datos.

1.3 Prerrequisitos

Se necesitan conocimientos de programación y álgebra. Recomendable saber utilizar un lenguaje como Python o Scala. Nivel de inglés medio que permita la lectura de documentación y el seguimiento de vídeos.

2. Competencias

2.1 Generales del título

- CG1. Adquisición de competencias teóricas y prácticas para el análisis y diseño de soluciones empresariales en Big Data (almacenamiento y procesamiento de grandes volúmenes de información heterogénea).
- CG3. Capacidad de diseñar e implementar sistemas capaces de extraer conocimiento práctico de grandes volúmenes de datos aplicado al mundo de la empresa (Inteligencia de Negocio/*Business Intelligence*).

2.2 Específicas materia

- CDS1. Capacidad de aplicar, validar y evaluar métodos de Ciencia de Datos/*Data Science* e Inteligencia Artificial sobre conjuntos y flujos de datos masivos y complejos.
- CDS3. Capacidad para el análisis, exploración y síntesis de conjuntos complejos de datos no estructurados y de diseñar soluciones que permitan extraer de los mismos información relevante y valiosa para el soporte a la toma de decisiones.

3. Resultados de aprendizaje

Al finalizar la asignatura, el alumno será capaz de:

- Buscar de forma eficiente elementos similares en grandes volúmenes de datos.
- Determinar la similitud de datos almacenados y/o agrupamiento en función de su similitud.
- Ser capaz de reducir la dimensionalidad de un conjunto de datos para su visualización.
- Ser capaz de hacer un análisis de las componentes principales.
- Identificar elementos frecuentes y descubrir de reglas de asociación.

4. Contenido / Programa de la asignatura

4.1 Unidades docentes (bloques de contenidos)

- Búsqueda de elementos similares (minhasing, Jacquard similarity, Locally Sensitive Hashing).
- Técnicas de agrupamiento (K-means, agrupamiento jerárquico).
- Visualización y reducción de la dimensionalidad (PCA, SVD).
- Problema de artículos frecuentes (modelados de cestas de la compra, reglas de asociación)

4.2 Bibliografía

Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, *Mining of Massive Datasets*, 2014

5. Metodología de enseñanza y dedicación del estudiante a la asignatura

Actividad Formativa	Competencias relacionadas	Horas	Presencialidad (%)
Clases, conferencias y técnicas expositivas	CG1, CG3, CDS1, CDS3	12	0
Actividades autónomas y en grupo (trabajos y lecturas dirigidas)	CG1, CG3, CDS1, CDS3	45	0
Pruebas de seguimiento y exposición de trabajos	CG1, CG3, CDS1, CDS3	10	50
Tutoría individual, participación en foros y otros medios colaborativos	CG1, CG3, CDS1, CDS3	8	0

6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Identificación de documentos similares	0.86	Semanas 1 y 2
Clustering	0.86	Semanas 3 y 4
Reducción de la dimensionalidad	0.86	Semanas 5 y 6
Conjuntos de ítemes frecuentes	0.42	Semana 7

Lo anterior son estimaciones que no deben interpretarse en su literalidad. En la práctica, siempre surgen imponderables que podrían obligar a hacer cambios respecto a esta planificación inicial.

7. Evaluación

Instrumento / Procedimiento	Peso primera convocatoria	Peso segunda convocatoria
Evaluación sumativa, que incluye pruebas parciales individuales y prueba final	40%	40%
Realización de trabajos, proyectos, resolución de problemas y casos	40%	40%
Participación en foros y otros medios participativos	20%	20%

Criterios / Comentarios a la evaluación
• **Convocatoria ordinaria:** La calificación final será la media ponderada al porcentaje indicado en las tablas. Para la superación de la asignatura se exigirá un mínimo de 4 puntos sobre 10 en los procedimientos de «Evaluación sumativa…» y «Realización de trabajos…». • **Convocatoria extraordinaria:** Es posible que el procedimiento «Participación en foros y otros medios participativos» no sea recuperable en su totalidad en 2ª convocatoria. La evaluación se basa en la interacción entre los alumnos y es posible que esta no pueda organizarse de forma satisfactoria por restricciones de tiempo o de número de alumnos en ese período. En ese caso, se conservará la nota obtenida en la 1ª convocatoria

8. Recursos de aprendizaje y apoyo tutorial del curso online

Transparencias.
Vídeos (en inglés).
Notas en castellano sobre los vídeos.
Cuestionarios de autoevaluación.
Cuestionarios de evaluación.
Enunciados de ejercicios.
Páginas Webs relacionadas.
Bibliografía disponible en la Biblioteca
Tutorías individualizadas o en grupo a demanda de los alumnos.

9. Consideraciones / Comentarios adicionales