



Proyecto docente

Asignatura	Técnicas de Aprendizaje Automático Escalables		
Materia	Ciencia de Datos		
Titulación	Máster Universitario en Inteligencia de Negocio y Big Data en Entornos Seguros		
Plan	621 (UVA)	Código	54549 (UVA) 8103#90º (UBU)
Periodo de impartición	S2	Tipo/Carácter	Obligatoria
Nivel/Ciclo	Máster	Curso	1
Créditos ECTS	3		
Lengua en que se imparte	Castellano		
Profesor/es responsable/s	Carlos J. Alonso González, J. Belarmino Pulido Junquera		
Datos de contacto (e-mail, teléfono...)	calonso@infor.uva.es ; (Tel. 983185602) belar@infor.uva.es (Tel. 983185606)		
Horario de tutorías	Consultar página web de www.uva.es		
Coordinador	Carlos J. Alonso González		
Departamento	Informática (ATC, CCIA y LSI)		
Web			
Descripción General			



1. Situación / Sentido de la asignatura

1.1 Contextualización

La asignatura “Técnicas de Aprendizaje Automático Escalable” introduce los elementos necesarios para aplicar técnicas de Aprendizaje Automático a grandes volúmenes de datos como lo son los procedentes de aplicaciones web o móviles, la Internet de las Cosas y las redes de sensores, así como procedentes de servicios financieros, sanidad u otros campos científicos.

El conjunto de datos que se puede usar en estos campos es enorme y el conjunto de técnicas de aprendizaje a aplicar muy variado. Estos datos puede ser propiedad de una organización o pueden proceder de múltiples fuentes, pero en todos los casos su volumen puede ser tan grande que no se puedan procesar en un único ordenador, por lo cual será necesario recurrir posiblemente a un almacenamiento distribuido, a un procesamiento distribuido o a ambos.

Además, la gran cantidad de datos a procesar hará necesario analizar con cuidado el tipo de técnicas o algoritmos aplicables, ya que los requisitos de memoria pueden hacer inviables la utilización de técnicas o aplicaciones más convencionales.

Esta asignatura se centra en los conceptos básicos del aprendizaje automático, que presenta en el contexto de Big Data, así como en dos tipos de técnicas de aprendizaje supervisado: los clasificadores y los recomendadores.

1.2 Relación con otras asignaturas

La asignatura se imparte en el segundo semestre del curso y forma parte de la Materia “Ciencia de Datos” que se centra en el procesamiento escalable de datos. Al igual que para el almacenamiento, la aplicación de técnicas de análisis en grandes volúmenes de datos, se tiene que apoyar en tecnologías adecuadas a la forma de almacenar, el tipo y el volumen de los datos con los que se está tratando.

La asignatura forma parte de esta Materia junto con las asignaturas “Aprendizaje sobre flujo de datos” y “Aprendizaje no supervisado”.

La asignatura proporciona conocimientos y habilidades necesarias para abordar los contenidos de las otras dos asignaturas de la Materia “Ciencia de Datos”, pero principalmente para las técnicas de “Aprendizaje sobre flujo de Datos”, aunque los conocimientos básicos de procesados y limpieza de datos, así como la validación de los modelos es general para la Materia.

Además, los conocimientos y habilidades adquiridas en esta asignatura pueden utilizarse dentro del contexto de la Materia “Inteligencia de Negocio” y las dos asignaturas que allí se imparten: “Inteligencia de negocio aplicada” I y II.

1.3 Prerrequisitos

Aunque la asignatura será autocontenida se recomienda haber cursado y superado las asignaturas de la Materia “Tecnologías Informáticas para el Big Data”, especialmente las de “Modelos de programación para el Big Data” y “Arquitecturas Big Data”. Sería recomendable también conocer los contenidos de la asignatura “Almacenamiento Escalable”.



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.
- CDS2. Capacidad de dirigir proyectos para la extracción de conocimiento basados en métodos eficientes de análisis de datos.
- 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:

- Comprender y aplicar métodos de inducción de clasificadores escalables a grandes conjuntos de datos.
- Comprender y aplicar métodos de generación de recomendaciones escalables a grandes conjuntos de datos.
- Comprender y ser capaz de aplicar las metodologías experimentales para la selección de modelos y evaluación de clasificadores.
- Comprender y ser capaz de aplicar la metodología para realizar un proyecto de aprendizaje automático en el contexto de Big Data.
- Conocer y utilizar algunas de las plataformas tecnológicas que permiten desarrollar proyectos de aprendizaje automático en Big Data



4. Contenido / Programa de la asignatura

4.1 Unidades docentes (bloques de contenidos)

- Introducción a las plataformas tecnológicas (Apache Spark, Scala u otras)
- Metodología de análisis Big Data. Preparación de datos.
- Clasificadores escalables básicos
- Metodología experimental de evaluación y selección de modelos
- Recomendadores
- Clasificadores escalables avanzados

4.2 Bibliografía

- Nick Pentreath. Machine Learning with Spark. Packt Publishing. 2015. ISBN: 9781783288519. <http://www.packtpub.com/>
- Petar Zečević y Marko Bonaći. Spark in Action. Manning Publications. 2016. ISBN: 9781617292606. <https://www.manning.com/books/>
- Mohamed Guller. Big Data Analytics with Spark. Apress. 2015.
- Ian H. Witten, Eibe Frank y Mark A. Hall. Data Mining: practical machine learning tools and techniques (third Edition). Morgan Kaufmann, 2011.

Bibliografía complementaria

- Apache Organization. Apache Spark. <http://spark.apache.org/>
- Apache Organization. Apache MLlib. <http://spark.apache.org/mllib/>
- Kaggle. Kaggle in class. <https://inclass.kaggle.com/>
- Rishi Yadav. Spark Cookbook. Packt Publishing 2015.
- C. Bishop. Pattern Recognition and Machine Learning. Springer, N.Y., 2005
- Jure Leskovek, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Second edition. Cambridge University Press, 2014.
- L. Kuncheva, Combining pattern classifiers, Second edition. Wiley, 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, CDS2, CDS3	45	0
Pruebas de seguimiento y exposición de trabajos	CG1, CG3	10	50
Tutoría individual, participación en foros y otros medios colaborativos	CG1, CG3, CDS1	8	0



6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Introducción a las plataformas tecnológicas (Apache Spark, Scala u otras)	0,7	<i>Semanas 1-2</i>
Metodología de análisis Big Data. Preparación de datos.	0,6	<i>Semanas 2-3</i>
Clasificadores escalables básicos	0,6	<i>Semanas 3-4</i>
Metodología experimental de evaluación y selección de modelos	0,3	<i>Semana 5</i>
Recomendadores	0,4	<i>Semana 6</i>
Clasificadores escalables avanzados	0,4	<i>Semana 7</i>



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	50%	50%
Participación en foros y otros medios participativos	10%	10%

Crterios / Comentarios a la evaluación

- **Primera convocatoria:** Será necesario realizar todos los trabajos asignados, obteniendo un mínimo de 4 puntos en todos ellos, para optar a la nota promediada.
- **Segunda convocatoria:** Será necesario realizar todos los trabajos asignados, que serán diferentes de los de la primera convocatoria, obteniendo un mínimo de 4 en todos ellos, para optar a la nota promediada.
- La calificación de participación en foros y otros medios se obtendrá durante el período de la primera o de segunda convocatoria.

Participación en los foros.

Los foros de la asignatura serán moderados a posteriori por los profesores responsables de la misma, eliminando cualquier mensaje que no se considere adecuado.

Por supuesto, el lenguaje será correcto en todo momento, no admitiéndose ningún tipo de descalificaciones o comentarios que puedan resultar ofensivo para cualquier participante.

El foro tiene una naturaleza científico-técnica. No se admitirá ningún comentario que no se ciña a los aspectos científico-técnicos relacionado con la asignatura.

No se admitirá ningún tipo de opinión personal más allá de las opiniones que con base científica-técnica se pueden presentar como conclusiones de los trabajos realizados.

La moderación de los comentarios tendrá la siguiente penalización sobre los alumnos autores del comentario:

Tras el primer comentario moderado se enviará un aviso a los alumnos implicados.

Tras el segundo comentario moderado se restará un punto de la nota final de los alumnos implicados.

Tras el tercer comentario moderado se suspenderá el acceso al foro a los alumnos implicados y se les restarán 3 puntos adicionales de la nota final.



8. Recursos de aprendizaje y apoyo tutorial del curso online

- Transparencias.
- Enunciados de ejercicios.
- Cuestionarios de autoevaluación.
- Páginas Webs relacionadas
- Bibliografía disponible en la Biblioteca
- Tutorías individualizadas o en grupo a demanda de los alumnos.



9. Consideraciones / Comentarios adicionales