FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	36426
Nombre	Aprendizaje máquina
Ciclo	Grado
Créditos ECTS	6.0
Curso académico	2021 - 2022

Titulación(es)

Titulación	Centro	Curso	Periodo
1406 - Grado en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	3	Primer cuatrimestre

Materias

Titulación	Materia	Carácter
1406 - Grado en Ciencia de Datos	9 - Aprendizaje automático y minería de datos	Obligatoria

Coordinación

Nombre	Departamento
LAPARRA PEREZ-MUELAS, VALERO	242 - Ingeniería Electrónica
MARTIN GUERRERO, JOSE DAVID	242 - Ingeniería Electrónica
MUÑOZ MARI, JORDI	242 - Ingeniería Electrónica

RESUMEN

La asignatura “Aprendizaje Máquina” supone el primer contacto del estudiante del Grado en Ciencia de Datos con los modelos matemáticos no lineales y los correspondientes algoritmos de aprendizaje que permiten la extracción de información almacenada en bases de datos para resolver fundamentalmente problemas de los siguientes tipos:

- Clasificación
- Agrupamiento
- Regresión
- Predicción
- Planificación

Cada uno de estos problemas requiere una aproximación que puede ser diferente desde el punto de vista del aprendizaje. Por lo tanto, la asignatura empieza revisando conceptos básicos y definiciones para establecer el marco de trabajo que permitirá introducir los diferentes tipos de aprendizaje automático que se estudiarán: aprendizaje supervisado, no supervisado, semisupervisado, activo y por refuerzo. A continuación, se estudiará la manera de evaluar los resultados en estos problemas, las diferentes métricas existentes, la necesidad de hacer subdivisiones de conjuntos para garantizar un mejor funcionamiento y las posibles mejoras que se pueden plantear, como por ejemplo las técnicas de *boosting* o *bagging* para generar *ensembles*.

Una vez establecido el marco de trabajo, se estudian diferentes modelos de aprendizaje automático para resolver los problemas descritos anteriormente. De esta manera, "Aprendizaje Máquina" da un paso más respecto a asignaturas de cursos anteriores donde los alumnos se han centrado mayormente en un análisis de datos más descriptivo o en modelos basados en aproximaciones lineales.

Respecto a los modelos, se estudian los métodos mayormente utilizados y más populares hoy en día con la excepción de las redes neuronales que se describen a la asignatura "Modelos Conexionistas", en el segundo cuatrimestre. En particular, se describirán en detalle las características, funcionamiento, adecuación a diferentes problemas y la interpretación de modelos basados en máquinas de vectores soporte *(Support Vector Machines, SVM)* y árboles de decisión; en el caso de los árboles tendrá especial relevancia la generalización con *ensembles* puesto que da lugar a los modelos de bosques aleatorios *(Random Forest, RF)*, que son una de las aproximaciones más potentes que se pueden utilizar para resolver problemas de regresión y clasificación.

Las clases de teoría se impartirán en castellano y las clases prácticas y de laboratorio según consta en la ficha de la asignatura disponible en la web del grado.

CONOCIMIENTOS PREVIOS

Relación con otras asignaturas de la misma titulación

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

Otros tipos de requisitos

La asignatura se enmarca en el 3er curso del grado, donde se asume que el estudiante tiene conocimientos mínimos de los procesos implicados en un análisis inteligente de datos, como por ejemplo el filtrado de datos ruidosos, atípicos o perdidos, así como la representación e interpretación de los datos en un espacio multidimensional y la generación de visualizaciones que permitan obtener información útil del problema. No hay requisitos adicionales para poder seguir la asignatura; de los contenidos estudiant

COMPETENCIAS (RD 1393/2007) // RESULTADOS DEL APRENDIZAJE (RD 822/2021)

1406 - Grado en Ciencia de Datos

- (CG02) Capacidad de resolver problemas con iniciativa, creatividad, y de comunicar y transmitir conocimientos, habilidades y destrezas, comprendiendo la responsabilidad ética y profesional de la actividad del Científico de Datos.
- (CG03) Capacidad para la realización de modelos, cálculos, informes, planificación de tareas y otros trabajos análogos en el ámbito específico de la Ciencia de Datos.
- (CT03) Habilidad para defender su trabajo con rigor y argumentos, exponiéndolo de forma adecuada y precisa, apoyándose en los medios necesarios.
- (CT05) Capacidad para evaluar las ventajas e inconvenientes de diferentes alternativas metodológicas y/o tecnológicas en distintos ámbitos de aplicación.
- (CE03) Capacidad para resolver problemas de clasificación, modelización, segmentación y predicción a partir de un conjunto de datos.
- (CE07) Capacidad para modelar la dependencia entre una variable respuesta y varias variables explicativas, en conjuntos de datos complejos, mediante técnicas de aprendizaje máquina, interpretando los resultados obtenidos.
- (CE13) Saber diseñar, aplicar y evaluar algoritmos de Ciencia de Datos para la resolución de problemas complejos.
- (CB3) Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
- (CB4) Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.

RESULTADOS DE APRENDIZAJE (RD 1393/2007) // SIN CONTENIDO (RD 822/2021)

Los resultados de aprendizaje más importantes son:

- Conocer e implementar árboles basados en datos.
- Conocer la base de los métodos *kernel* y los diferentes *kernels* que se pueden plantear.
- Obtener reglas de asociación a partir de bases de datos *(basket analysis)*.
- Conocer las diferentes formas que se tienen de asociar sistemas expertos.

Cada uno de estos cuatro resultados permite en mayor o menor medida adquirir todas las competencias descritas anteriormente. Sin embargo, particularizando los resultados asociados a las diferentes competencias, respecto a las competencias básicas y generales, los resultados de aprendizaje más relevantes serán:

- Realización de trabajos e informes con diferente nivel de guiado para que el estudiante logre un alto grado de autonomía a finales del cuatrimestre, tanto en cuanto a la decisión de las aproximaciones más adecuadas para resolver un problema como la obtención de los mejores resultados posibles y su interpretación en un entorno multidisciplinar (CG02, CB3, CB4).
- Tratamiento de conjuntos de datos en diferentes formatos confiriendo al estudiante de la capacidad de hacer un tratamiento de datos independientemente del formato de éstos (CG02, CG03).

Las competencias transversales tendrán asociados los siguientes resultados:

- Realización de presentaciones formales para una audiencia experta y no experta (CT03, CT05).
- Defensa de los argumentos expuestos en la presentación ante las dudas que puedan plantearse (CT03, CT05).
- Capacidad de evaluación crítica del trabajo propio y del realizado por otros compañeros y colegas (CT05).

Finalmente, las competencias específicas se verán reflejadas en los siguientes resultados de aprendizaje:

- Desarrollo de modelos de clasificación, modelado, agrupamiento, predicción, regresión y planificación, entendiendo el procesamiento interno realizado por modelos y algoritmos y teniendo la capacidad de proponer variantes que puedan mejorar los resultados alcanzados (CE03, CE07, CE13).
- Capacidad para adaptar los modelos a las características peculiares de cada problema (CE03, CE07, CE13).
- Capacidad para elegir los modelos más adecuados a cada problema y evaluarlos con métricas objetivas (CE13).
- Interpretación de los resultados del modelado y utilidad de las soluciones aportadas en un entorno multidisciplinar (CE07).
- Entender los diferentes tipos de aprendizaje de los algoritmos y exprimir su capacidad tanto de manera individual como combinándolos, si procede (CE07, CE13).

DESCRIPCIÓN DE CONTENIDOS

1. Conceptos preliminares

Esta primera unidad temática introduce algunos conceptos necesarios para poder preparar los conjuntos de datos de forma que puedan ser analizados por métodos de aprendizaje automático. En particular se estudiarán los siguientes contenidos:

1. Definiciones: muestra, patrón, característica, algoritmo, dimensionalidad,...
2. Normalización y codificación
3. Selección de características

3.1. Métodos filter

3.2. Métodos wrapper

4. Extracción de características(*): Descomposición en valores singulares, Análisis de Componentes Principales, Análisis de Componentes Independientes, ...

(*) Estos métodos y otros más sofisticados se estudiarán con mucho más de detalle en la asignatura Agrupamiento y Variedades

2. Esquemas de aprendizaje y problemas asociados

La segunda unidad temática describe los diferentes tipos de aprendizaje que definen los algoritmos que hacen la tarea de extracción de información en los modelos de aprendizaje automático. Dependiendo del esquema de aprendizaje utilizado se pueden abordar diferentes problemas siempre sobre la base de que estén definidos mediante un conjunto de datos. Los diferentes apartados que se estudiarán en la unidad temática II son:

1. Aprendizaje supervisado

1.1. Problemas de clasificación

1.2. Problemas de regresión y predicción

2. Aprendizaje no supervisado: Problemas de agrupamiento y segmentación(*)
3. Aproximaciones semisupervisadas

3.1. Aprendizaje semisupervisado

3.2. Aprendizaje activo

4. Aprendizaje por refuerzo: problemas de optimización

(*)Estos métodos se estudiarán con profundidad en la asignatura Agrupamiento y Variedades

3. Evaluación de modelos

La tercera unidad temática se centra en las diferentes métricas que se pueden utilizar para evaluar el rendimiento de modelos de aprendizaje automático. Dependiendo del tipo de problema abordado y por tanto del esquema de aprendizaje, las métricas a considerar son diferentes. Además, se analizará la conveniencia de una u otra métrica en función del objetivo final que se persiga de forma que los modelos puedan sesgarse para minimizar o maximizar ciertos criterios. También se describirán algunas problemáticas que hay que tener en cuenta y diferentes técnicas que permiten mejorar el rendimiento de los modelos y hacerlos más útiles de cara a su utilización en un problema real. La tabla de contenidos de la unidad temática III es la siguiente:

1. Sobreentrenamiento y sobreajuste
2. División del conjunto de datos

2.1. Hold-out

2.2. V-fold
2.3 Leave-one-out
3. Evaluación del rendimiento de modelos de aprendizaje automático
3.1. Problemas de clasificación
3.2. Problemas de regresión
4. Mejora de modelos
4.1. Boosting
4.2. Comités de expertos: ensembles
4.3 Bagging

4. Máquinas de Vectores Soporte

Las Máquinas de Vectores Soporte (SVMs, por su nombre en inglés) son un modelo de aprendizaje inicialmente propuesto para tareas de clasificación pero que con modificaciones puede aplicarse también para regresión. Son especialmente indicadas en espacios dispersos con poca densidad de datos, empeorando bastante su funcionamiento a medida que aumenta el número de muestras del conjunto de datos. En esta unidad temática se revisará su funcionamiento y características, desglosado de la siguiente manera:

1. Introducción
2. Hiperplano de separación óptima
3. El truco del kernel
4. Regresor basado en vectores soporte (SVR)

5. Árboles de decisión

Esta última unidad temática teórica describe los modelos de aprendizaje automático que están basados en estructuras de árboles, empezando con modelos de árbol sencillo y finalizando con los modelos de bosques aleatorios (Random Forests, RFs) que utilizando bagging generan un conjunto de árboles, suponiendo un modelo que representa el estado del arte en problemas de clasificación y regresión. La tabla de contenidos de la unidad temática es:

1. Representación
2. Entropía y ganancia de información
3. Poda
4. Algoritmos principales
4.1. ID3
4.2. C4.5
4.3. CART: Classification and Regression Trees
4.4. CHAID: Chi-square Automatic Interaction Detection
5. Ensembles de árboles
5.1. Bosque aleatorio (Random Forest, RF)
5.2. Árboles extremadamente aleatorizados (Extremely Randomized Trees, ERTs)

6. Temas actuales en Aprendizaje Máquina

Después de haber estudiado los métodos más conocidos de Aprendizaje Máquina, el objetivo de esta unidad es finalizar la asignatura con la descripción de algunos de los temas que se están proponiendo e investigando en la actualidad para que el alumno conozca los desarrollos más recientes del campo. La siguiente tabla de contenidos es por tanto flexible y adaptable a la aparición de nuevas propuestas interesantes:

1. Aprendizaje profundo.
2. Aprendizaje automático cuántico.
3. Nuevas aplicaciones del aprendizaje máquina.

7. Prácticas de laboratorio

Finalmente, por su importancia en la asignatura se ha considerado conveniente incluir como una unidad temática independiente las prácticas a realizar en el laboratorio (aula informática), donde el estudiante aprenderá a implementar los modelos descritos en las clases de teoría. Muchos de los métodos analizados en la asignatura sólo cobran sentido cuando se desarrollan en un entorno de laboratorio en el que se puede observar su potencialidad, puesto que puede resultar relativamente complicado entender todas sus características de funcionamiento únicamente en base al estudio teórico y a la realización de ejercicios y problemas sencillos. Se plantean seis prácticas de laboratorio correspondiendo con los contenidos teóricos previamente descritos en las anteriores unidades temáticas:

1 Preprocesado de conjuntos de datos
1.1. Normalización
1.2. Codificación
1.3. Selección de características
1.3.1 Métodos filter
1.3.2 Métodos wrapper
2 Extracción de características
2.1. Separación de conjuntos: Hold-out, V-fold, Leave-one-out
2.2. Análisis de Componentes Principales (PCA)
3 SVMs en problemas de clasificación
3.1. Ejemplos
3.2. Aprendizaje activo con SVMs
4 SVRs en problemas de regresión
5 Modelos basados en árbol de decisión para problemas de clasificación y regresión
6 Modelos basados en ensembles de árboles
6.1. Bagging
6.2. Random Forest

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases de teoría	32,00	100
Prácticas en laboratorio	20,00	100
Prácticas en aula	8,00	100
Asistencia a eventos y actividades externas	2,00	0
Elaboración de trabajos en grupo	6,00	0
Elaboración de trabajos individuales	5,00	0
Estudio y trabajo autónomo	35,00	0
Lecturas de material complementario	4,00	0
Preparación de actividades de evaluación	20,00	0
Preparación de clases de teoría	4,00	0
Preparación de clases prácticas y de problemas	4,00	0
Resolución de casos prácticos	7,00	0
Resolución de cuestionarios on-line	3,00	0
TOTAL	**150,00**	

METODOLOGÍA DOCENTE

Las metodologías docentes utilizadas en esta asignatura son:

MD1 - Actividades teóricas (CG03, CB4, CT03, CT05, CE03, CE07, CE13): Desarrollo expositivo de la materia con la participación del estudiante en la resolución de cuestiones puntuales. Realización de cuestionarios individuales de evaluación.

MD2 - Actividades prácticas (CG02, CB3, CT05, CE03, CE07, CE13): Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio a través de los cuales se adquieren competencias sobre los diferentes aspectos de la materia.

MD4 - Trabajos en laboratorio y/o aula ordenador (CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Aprendizaje mediante la realización de actividades desarrolladas de forma individual o en grupos reducidos y llevadas a cabo en laboratorios y/o aulas de ordenador.

A continuación, se describe con más detalle el método de enseñanza-aprendizaje a utilizar en la asignatura. La metodología docente tendrá dos enfoques diferentes, uno para las clases teóricas y de problemas y otro para las clases prácticas de laboratorio. Se usará el aula Virtual y sus utilidades, especialmente en lo que corresponde a disposición de material, a las evaluaciones automáticas y en las clases remotas, si procede.

Respecto a las clases teóricas, el aprendizaje se hará en los dos sentidos, desde el profesor hasta el alumno, y desde el alumno hasta el profesor. En la parte que nace del docente, habrá dos fuentes de generación de conocimiento. Por un lado, la clase magistral en la que el profesor introducirá los conceptos nuevos que van apareciendo, relacionándolos con los conocimientos previos de los estudiantes para facilitar su comprensión. Por otro lado, el alumno dispondrá con anterioridad a su explicación en clase de material para poder preparar mínimamente las clases teóricas y así agilizar el proceso de enseñanza-aprendizaje; este mismo material contendrá también la información necesaria para que el estudiante pueda complementar y repasar la información recibida en clase.

En el proceso que fluye desde el estudiante habrá también dos aproximaciones que se corresponden con las fuentes de generación de conocimiento anteriormente comentadas. La clase magistral del profesor se verá reforzada con la resolución de ejercicios prácticos y problemas de creciente complejidad a medida que avance la unidad temática; si bien el docente realizará algún ejercicio como ejemplo, la mayoría de estos problemas tendrán que ser resueltos propiamente por los estudiantes para garantizar una comprensión total de los contenidos de la asignatura, que a pesar de tener una fuerte componente teórica, su punto fuerte es la aplicación práctica a diferentes problemas y estos ejercicios y problemas hacen que el estudiante pueda averiguar las particularidades de los diferentes algoritmos en su aplicación a problemas de diversa índole. Además, el estudiante realizará trabajos con una componente de investigación individual en contenidos que podrían considerarse como una versión sofisticada de los descritos en la asignatura; en particular, se plantea que el estudiante pueda realizar, presentar y defender trabajos sobre métodos actuales de aprendizaje automático, que al mismo tiempo, harán que mejore su base en los contenidos más básicos que se describen en detalle en la asignatura. Se contempla que algunos de estos trabajos puedan ser voluntarios y se realicen de manera individual o por parejas.

Respecto a las clases prácticas, pueden distinguirse tres metodologías docentes. Primeramente, y con anterioridad a la realización de las prácticas, los estudiantes tendrán que prepararse de manera autónoma los ejercicios prácticos a realizar, consultado las dudas que les puedan surgir con los profesores preferentemente antes de la realización de la práctica. Este aspecto se verá apoyado con la realización de un cuestionario corto y sencillo al principio de la práctica para comprobar que la preparación se ha realizado correctamente.

La segunda metodología docente que aparece en las prácticas es el propio trabajo a realizar durante la sesión de prácticas, que básicamente consistirá en la programación en Python que permita encontrar la solución correcta en los problemas planteados en la práctica. Este trabajo se realizará de manera individual o por parejas y tendrá en todo momento la supervisión del profesor; en primer lugar, porque los ejercicios serán previamente explicados y en segundo porque el alumno podrá en todo momento consultar sus dudas para poder avanzar correctamente en la realización de la práctica.

La tercera aproximación docente en prácticas vuelve a dar el protagonismo al alumno, que a la finalización de la práctica tendrá que ser capaz de hacer una discusión crítica de los resultados logrados y contestar correctamente a las preguntas formuladas por el profesor y a realizar los ejercicios planteados en la sesión. Dependiendo de las circunstancias (ocupación del aula, clase presencial o no, etc.) esta discusión podría realizarse de manera automática mediante las herramientas disponibles en Aula Virtual.

EVALUACIÓN

La calificación final de la asignatura se obtendrá como resultado de la media pesada entre las partes de teoría y de prácticas. De acuerdo con los créditos asignados a cada parte, la teoría tendrá una representación de 2/3 en la nota final y la práctica el tercio restante.

La nota de teoría correspondiente a la primera convocatoria saldrá como resultado de:

- SE1 (60%; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13): Pruebas objetivas, consistentes en uno o más exámenes de cuestiones teóricas, problemas sintéticos y problemas prácticos reales. Para superar la asignatura, se exigirá una calificación mínima de 4 (sobre 10) en esta parte.

- SE2 (30%; CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Trabajos, memorias y exposiciones orales.

- SE3 (10%; CG02, CB4, CT03, CE03, CE07, CE13): Evaluación continua de cada alumno, basada en la participación y grado de implicación del alumno en el proceso de enseñanza-aprendizaje, teniendo en cuenta la asistencia regular a las actividades presenciales previstas y la resolución de cuestiones y problemas propuestos periódicamente.

Respecto a la calificación de prácticas, el 40% de la nota corresponderá con SE2 (CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13) y el 60% con la calificación obtenida en la práctica final (SE1; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13), que tendrá lugar en la última sesión de prácticas. La práctica final será una prueba objetiva que se evaluará individualmente y que consistirá en la realización de diferentes ejercicios relacionados con una o varias prácticas anteriores. Para superar la asignatura, se exigirá una calificación mínima de 4 (sobre 10) en la práctica final. Del 40% correspondiente a la evaluación continua, el 70% corresponderá con la realización de los ejercicios propuestos en la sesión de prácticas, que podrán ser evaluados por el profesor a la finalización de la práctica. El 30% restante provendrá de la preparación previa a la sesión de prácticas y que se evaluará rápidamente al principio de cada sesión de prácticas. Las prácticas pueden realizarse de manera individual o por parejas, con la excepción de la práctica final, que obligatoriamente será individual. Además, el profesor puede optar para evaluar de manera individual las sesiones regulares de prácticas aunque éstas se hayan desarrollado por grupos de dos estudiantes.

La segunda convocatoria se evaluará igual que la primera con la excepción de que en la parte de teoría, SE1 tendrá un peso del 70% y SE3 del 0%; en la parte de prácticas el 100% corresponderá a SE1, y será necesario obtener un mínimo de 4 (sobre 10) para promediar con la parte de teoría.

En cualquier caso, el sistema de evaluación se regirá por lo establecido en el Reglamento de Evaluación y Calificación de la Universidad de Valencia para Grados y Másteres (https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639)

REFERENCIAS

Básicas

- E. Alpayidin, F. Bach (2014). Introduction to Machine Learning, Third Edition, The MIT Press (disponible com a eBook per a la Universitat de València)
- S. Theodoridis (2015). Machine Learning: a Bayesian and Optimization perspective, Elsevier (disponible com a eBook per a la Universitat de València)
- D. Haroon (2017). Python Machine Learning Case Studies: Five Case Studies for the Data Scientist, Apress (disponible com a eBook per a la Universitat de València)

Complementarias

- C. M. Bishop (2016). Pattern Recognition and Machine Learning, Springer
- K. P. Murphy (2020). Machine Learning: a probabilistic perspective, Second Edition, The MIT Press
- R. O. Duda, P. E. Hart, D. G. Stark (2016). Pattern classification, Third Edition, John Wiley & Sons Inc.
- T. Hastie, R. Tibshirani, J. Friedman (2011) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer (Series in Statistics)
- S. Raschka, V. Mirjalili (2019). Python Machine Learning. Packt Publishing
- David V. (2017). Machine Learning with Python: The Basics. CreateSpace Independent Publishing Platform

ADENDA COVID-19

Esta adenda solo se activará si la situación sanitaria lo requiere y previo acuerdo del Consejo de Gobierno

La metodología docente de la asignatura seguirá el Modelo Docente aprobado por la Comisión Académica de Título de Ciencia de Datos (https://go.uv.es/cienciadatos/ModelDocentGCD1Q). En caso de que se produzca un cierre de las instalaciones por causas sanitarias que afecte total o parcialmente a las clases de la asignatura, estas serán sustituidas por sesiones no presenciales siguiendo los horarios establecidos. Si el cierre afectara a alguna prueba de evaluación presencial de la asignatura, esta será sustituida por una prueba de naturaleza similar que se realizará en modalidad virtual a través de las herramientas informáticas soportadas por la Universitat de València. Los porcentajes de cada prueba de evaluación permanecerán invariables, según lo establecido por esta guía.