

**FICHA IDENTIFICATIVA****Datos de la Asignatura**

Código	36427
Nombre	Agrupamiento y Variedades
Ciclo	Grado
Créditos ECTS	6.0
Curso académico	2020 - 2021

Titulación(es)

Titulación	Centro	Curso	Periodo
1400 - Grado de Ingeniería Informática	Escuela Técnica Superior de Ingeniería	4	Primer cuatrimestre
1406 - Grado en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	3	Primer cuatrimestre

Materias

Titulación	Materia	Caracter
1400 - Grado de Ingeniería Informática	16 - Materia Optativa	Optativa
1406 - Grado en Ciencia de Datos	9 - Aprendizaje automático y minería de datos	Obligatoria

Coordinación

Nombre	Departamento
MARTINEZ GIL, FRANCISCO	240 - Informática

RESUMEN

La asignatura 'Agrupamiento y Variedades' es el complemento natural de la asignatura 'Aprendizaje Máquina' que se imparte también en el primer cuatrimestre del tercer curso.

Se revisan las técnicas más importantes para encontrar estructuras y patrones en conjuntos de datos que no están etiquetados. La asignatura se divide en dos partes. En la primera se tratan las técnicas más extendidas de aprendizaje no supervisado, conocidas también como técnicas de agrupamiento (clustering). Se revisan el agrupamiento jerárquico, el agrupamiento por particiones en sus variantes 'Hard' (K-Means) y 'Soft' (Fuzzy-C-Means), y modelos de agrupamientos no compactos como el agrupamiento basado en densidad (DBSCAN) y el basado en grafos (agrupamiento espectral).



En la segunda parte se amplían los conocimientos acerca de reducción de la dimensionalidad en datos ya introducidos en la asignatura de segundo curso 'Modelos Lineales'. Mientras que en esta asignatura se introduce PCA como el modelo lineal más importante para la reducción de dimensionalidad, en 'Agrupamiento y variedades' extenderemos esta problemática a modelos no lineales, es decir, a organizaciones espaciales de datos que no pueden ser modeladas mediante hiperplanos. Comienza esta segunda parte con un modelo sencillo, los Mapas Autoorganizativos. Posteriormente se revisarán algunas técnicas dentro del grupo conocido como 'Manifold Learning', específicamente ISOMAP y Locally Linear Embedding. Por último se revisará la técnica para la visualización de datos conocida como t-SNE, lo cual nos conecta con la asignatura 'Visualización de datos' de segundo curso.

Aparte de las diferentes técnicas presentadas, el alumno/a adquirirá /revisará conocimientos acerca de conceptos esenciales en aprendizaje máquina, como son los conceptos de similitud, métricas, espacio de características, partición, o el problema de la explosión combinatoria de la dimensionalidad.

Debido a los numerosos temas que trata de abordar esta asignatura cuatrimestral, la presentación de los temas será necesariamente superficial, presentando las ideas básicas de cada técnica, los tipos de problemas que pretende resolver y exponiendo los algoritmos básicos.

Los conocimientos impartidos en esta asignatura son la base para comprender técnicas de aprendizaje presentadas en las asignaturas 'Procesado del Lenguaje Natural' y las diferentes optativas de análisis de datos geográficos, de audio y voz, de salud y Web y redes sociales.

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

Se recomienda haber superado las asignaturas: Álgebra, Tratamiento de Datos y Modelos Lineales.

COMPETENCIAS

1400 - Grado de Ingeniería Informática

- SI3 - Capacidad para participar activamente en la especificación, diseño, implementación y mantenimiento de los sistemas de información y comunicación.
- C1 - Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación.



- C2 - Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.
- C3 - 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.

1406 - Grado en Ciencia de Datos

- 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.
- Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.
- (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.
- (CE13) Saber diseñar, aplicar y evaluar algoritmos de Ciencia de Datos para la resolución de problemas complejos.

RESULTADOS DE APRENDIZAJE

Conocer el concepto de clustering y algoritmos principales (Jerárquico s, particionales) (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

Conocer los principales algoritmos de variedades (manifolds) que existen. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

Conocer los problemas de los algoritmos planteados y posibles soluciones. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)



DESCRIPCIÓN DE CONTENIDOS

1. Introducción

- 1.1 Aprendizaje no supervisado
- 1.2 Noción de similaridad. Concepto de métrica. Tipos de métricas
- 1.3 Conceptos básicos: espacio de características, vector de características, partición, matriz de proximidad.
- 1.4 Concepto de agrupamiento. Taxonomía

2. Agrupamiento Jerárquico

- 2.1 Ideas básicas. Agrupamientos aglomerativo y divisivo.
- 2.2 Tipos de Linkage
- 2.3 Algoritmos básicos
- 2.4 Dendogramas. Interpretación. Coeficiente de correlación

3. Agrupamiento por particiones

- 3.1 Ideas básicas. Agrupamientos Soft y Hard
- 3.2 Agrupamiento 'hard'. Algoritmo de K-Medias
- 3.3 Problemas asociados. Inicialización. Elección del número de clusters.
- 3.4 Agrupamiento 'soft'. Algoritmo Fuzzy-C-Medias
- 3.5 Problemas asociados.

4. Agrupamiento basado en grafos

- 4.1 Agrupamiento basado en grafos. Agrupamiento espectral
- 4.2 Ideas básicas. Representación de datos en grafos.
- 4.3 Laplaciana del grafo. Descomposición espectral.
- 4.4 Algoritmos. Implementaciones con librerías.

5. Agrupamiento basado en densidad

- 5.1 Ideas básicas. Conceptos de densidad
- 5.2 Algoritmo DBSCAN
- 5.3 Implementaciones en librerías y ejemplos



6. Evaluación y validación de agrupamientos

6.1 Técnicas para la evaluación de agrupamientos

6.1.1 Pureza. Índice aleatorio (Rand Index). Medida F

6.2 Técnicas para la validación de clusters

6.2.1 Regularidad. Compacidad y aislamiento. Estadísticos de Hopkins. SSE

7. Introducción a las técnicas de reducción de dimensionalidad

7.1 El problema de la maldición de la dimensionalidad

7.2 Concepto de Manifold y problemática de los agrupamientos no lineales.

7.3 Concepto de dimensionalidad intrínseca

8. Mapas Auto-organizativos

8.1 Concepto de aprendizaje competitivo. Concepto de SOM.

8.2 Algoritmo. Puntos fuertes y débiles de la aproximación. Implementaciones en librerías.

8.3 Ejemplos.

9. Aprendizaje sobre Variedades

9.1 Introducción. Ideas básicas acerca de las técnicas de Manifold Learning

9.2 Algoritmo ISOMAP. Puntos fuertes y débiles. Parámetros.

9.3 Algoritmo Locally Linear Embedding (LLE). Puntos fuertes y débiles. Parámetros.

9.4 Ejemplos

10. Reducción de dimensionalidad para visualización

10.1 Introducción al problema.

10.2 Algoritmo t-SNE. Puntos fuertes y débiles. Comportamiento

10.3 Implementaciones en librerías.

10.4 Ejemplos.

**VOLUMEN DE TRABAJO**

ACTIVIDAD	Horas	% Presencial
Clases de teoría	32,00	100
Prácticas en laboratorio	20,00	100
Prácticas en aula	8,00	100
Elaboración de trabajos en grupo	10,00	0
Elaboración de trabajos individuales	10,00	0
Estudio y trabajo autónomo	20,00	0
Lecturas de material complementario	10,00	0
Preparación de actividades de evaluación	10,00	0
Preparación de clases de teoría	10,00	0
Preparación de clases prácticas y de problemas	10,00	0
Resolución de casos prácticos	5,00	0
Resolución de cuestionarios on-line	5,00	0
TOTAL	150,00	

METODOLOGÍA DOCENTE

En las actividades teóricas se desarrollarán los temas exponiéndolos el professor utilizando medios audiovisuales, proporcionando una visión global e integradora de los contenidos. Se fomentará la participación del estudiante en clase a través de preguntas durante la exposición y la realización de cuestiones simples para fijar los conceptos que se presentan. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

En las clase de resolución de problemas se resolverán preferentemente por estudiantes los problemas planteados con una semana de antelación , de tal manera que el alumno/a tenga tiempo suficiente para trabajarlo en casa. Se incentivará la discusión de los problemas en el aula. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

En los laboratorios se planterán tareas de mayor complejidad y envergadura que las propuestas en las actividades del aula. Se fomentará el trabajo en grupo, por parejas, de las prácticas propuestas. También el alumno/a se familiarizará con las librerías de cálculo científico de Python así como con herramientas de (presentación de código como es Jupyter Notebook. CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)



EVALUACIÓN

Primera convocatoria:

Se realizará al menos una prueba objetiva parcial durante el cuatrimestre de impartición del curso.

Se relizará un examen final en la primera convocatoria.

El valor de las pruebas parciales podrá llegar hasta el 50% de la nota de teoría (SE1). El resto del porcentaje se le asignará a la prueba final.

El porcentaje sobre la nota final de esta parte (SE1) será 45%. (CB5, CT03, CT05, CE03, CE13)

El valor de las prácticas de laboratorio (SE2) representará un 35% de la nota total de la asignatura (CB5, CT03, CT05, CE03, CE13)

El valor de la evaluación continua (SE3) representará el 20% de la nota total (CB5, CT03, CT05, CE03, CE13)

Es necesario sacar una nota mínima de 4.5 en cada una de las partes anteriores (SE1, SE2, SE3) para poder superar la asignatura.

Segunda convocatoria

La nota de evaluación continua (SE3), por ser actividades presenciales , se considera no recuperable en la segunda convocatoria. Aunque no se impone la restricción de la nota mínima de la primera convocatoria.

Para la nota de prácticas se planteará la defensa de una o varias prácticas resumen o bien se relizará un examen práctico al terminar el examen de teoria que supondra el 100% de la nota de (SE2) en segunda convocatoria.

En cambio, la restricción de nota mínima se mantiene para la teoría. La nota de teoría será únicamente la nota del examen final de segunda convocatoria (sin considerar los parciales realizados en el curso).

La nota final se obtendrá como 50% SE1, 30% SE2, 20% SE3

REFERENCIAS



Básicas

- Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Pearson (2006)
- An introduction to Statistical Learning . Gureth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer (2013)
- Pattern Recognition. Sergios Theodoridis. AP (2009) Hay versión electrónica

Complementarias

- Scikit-Learn Users Guide (Hay versión electrónica)
- Python Data science handbook. Jacob Vanderplas. O'Reilly. (2016)

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 Dades (<https://go.uv.es/cienciadatos/ModelDocentGCD>). 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 se mantendrán, según lo establecido en esta guía