Universidad Zaragoza

Curso : 2019/20

62236 - Análisis avanzado de datos

Información del Plan Docente

Año académico: 2019/20
Asignatura: 62236 - Análisis avanzado de datos
Centro académico: 110 - Escuela de Ingeniería y Arquitectura

Titulación: 534 - Máster Universitario en Ingeniería Informática

Créditos: 3.0
Curso: 2
Periodo de impartición: Primer semestre
Clase de asignatura: Optativa
Materia: ---

1.Información Básica

1.1.Objetivos de la asignatura

La asignatura y sus resultados previstos responden a los siguientes planteamientos y objetivos:

Las técnicas estadísticas son habituales en el contexto y práctica de la Ingeniería Informática, tanto más cuanto más difundida está la necesidad de analizar la información que contienen las bases de datos, frecuentemente, de gran volumen. Los procedimientos estadísticos permiten por una parte establecer las principales características comunes y diferenciadoras de la base de datos, en variables o individuos, y por otra parte, cuantificar la incertidumbre presente en los datos.

Los estudiantes de un grado de Ingeniería Informática han aprendido a reconocer las situaciones donde son útiles los procedimientos estadísticos para una o varias poblaciones así como la aplicación de técnicas paramétricas o no paramétricas.

A partir de ese conocimiento básico, en esta asignatura se pretende que los estudiantes continúen su formación con la construcción de modelos que expliquen las relaciones entre variables o individuos en estudios observacionales. En particular, esto requiere conocer las herramientas habituales en la estimación máximo-verosímil y bayesiana, lo que incluye el algoritmo EM, los métodos MCMC y los procedimientos de remuestreo.

La asignatura introduce un abanico de técnicas estadísticas de entre las cuales el estudiante ha de elegir la más adecuada para el análisis de una base de datos específica. Excede de las posibilidades de la asignatura la revisión exhaustiva de estos procedimientos, por ello se presentan los más usuales en cada ámbito. Mediante el uso de las técnicas de tipo regresión se calculan predicciones de valores de la variable respuesta y cotas del error para tales predicciones, en situaciones de aprendizaje supervisado. Para las situaciones donde no existe una variable respuesta sino un conjunto de variables que representan la realidad, los métodos de aprendizaje no supervisado ayudan a reconocer los patrones de variables y casos, con el doble objetivo de caracterizarlos y de reducir la dimensión. Así mismo es importante que el estudiante conozca para cada una de las técnicas introducidas sus posibilidades y limitaciones.

Un objetivo derivado del anterior es que los estudiantes sean capaces de realizar el análisis de una base de datos mediante el software adecuado. Para ello se introducen los conocimientos necesarios para implementar dichos métodos en un software estadístico libre, basado en el lenguaje y entorno R.

En consecuencia, el objetivo global de la asignatura es que el estudiante conozca, comprenda y sea capaz de utilizar un conjunto de herramientas estadísticas para obtener soluciones en problemas en el ámbito del análisis de grandes bases de datos.

1.2.Contexto y sentido de la asignatura en la titulación

La docencia de la asignatura Análisis avanzado de datos se centra en el estudio de herramientas estadísticas de gran utilidad en el desarrollo de la materia de Big Data. Su desarrollo está completamente centrado y orientado en las técnicas para el tratamiento de datos y la extracción de la información que contienen.

El contenido de la asignatura aborda la presentación de los algoritmos habituales de estimación de modelos y de la significación de sus elementos. Posteriormente, se usan esas herramientas para presentar técnicas de aprendizaje supervisado y no supervisado. En aprendizaje supervisado se construyen modelos estadísticos para la estimación y predicción de una variable de interés relacionada con otras cuyo valor se conoce. El aprendizaje no supervisado proporciona herramientas para detectar patrones y relaciones cuando en el problema, por su naturaleza, se proporcionan únicamente los valores de las variables de entrada. Ambas técnicas han experimentado un espectacular desarrollo en los últimos años debido a la proliferación de extensas bases de datos en múltiples ámbitos (comerciales, genéticos, médicos), pasando a ser una de las herramientas fundamentales en muchos campos de la Ingeniería (identificación de patrones complejos, obtención de soluciones aproximadas mediante simulación, ajustes de datos experimentales, etc.), que de otro modo resultarían imposibles de tratar.

1.3.Recomendaciones para cursar la asignatura

No existe ningún requisito ni recomendación especial para cursar la asignatura.

2.Competencias y resultados de aprendizaje

2.1.Competencias

Al superar la asignatura, el estudiante será más competente para...

Conseguir adquirir las siguientes competencias básicas y generales:

CG-04 - Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la Ingeniería en Informática.

CG-08 - Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos

CG-11 - Capacidad para adquirir conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.

CG-12 - Capacidad para aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.

CG-13 - Capacidad para evaluar y seleccionar la teoría científica adecuada y la metodología precisa de sus campos de estudio para formular juicios a partir de información incompleta o limitada incluyendo, cuando sea preciso y pertinente, una reflexión sobre la responsabilidad social o ética ligada a la solución que se proponga en cada caso

CG-14 - Capacidad para predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad

CG-16 - Capacidad para desarrollar la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.

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.

Conseguir adquirir las siguientes competencias específicas:

CTI-07 - Capacidad para comprender y poder aplicar conocimientos avanzados de computación de altas prestaciones y métodos numéricos o computacionales a problemas de ingeniería.

CTI-09 - Capacidad para aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento.

2.2.Resultados de aprendizaje

El estudiante, para superar esta asignatura, deberá demostrar los siguientes resultados...

Específicos de la asignatura

1. Interpretar datos observacionales o experimentales, extraer la información que contienen, identificar las relaciones entre ellos y evaluar hipótesis en presencia de incertidumbre y variabilidad, interpretando adecuadamente sus resultados.

2. Comprender los métodos de estimación, por máxima-verosimilitud y bayesianos, conocer las herramientas y algoritmos para la estimación en grandes bases de datos.

3. Aplicar procedimientos estadísticos de construcción y validación de modelos empíricos que expresan la relación entre una variable respuesta y otras variables cuyo valor se puede conocer.

4. Utilizar las técnicas más relevantes de análisis multivariante que contribuyen a explicar las relaciones entre los datos e identificar patrones cuando no hay una variable respuesta.

Específicos del bloque de optatividad

5. Aplicar técnicas de aprendizaje, minería de datos y minería de procesos para la extracción de conocimiento en entornos que manejen grandes cantidades de datos a través de la Web.

6. Aplicar técnicas matemáticas para el análisis de grandes cantidades de datos en la Web.

2.3.Importancia de los resultados de aprendizaje

En buena parte de los problemas asociados al ámbito Big Data se proponen hipótesis de trabajo cuya comprobación sólo puede establecerse a partir de resultados de carácter estadístico. En estos casos se plantean estudios basados en información recogida en bases de datos observacionales, a menudo extensas. Los métodos estadísticos son, de una parte, el procedimiento para extraer la información relevante contenida en ellas, en particular para reconocer patrones y relaciones entre variables de interés. Por otra parte, la metodología estadística permite establecer la significación de los resultados y su validez para toda la población.

3.Evaluación

3.1.Tipo de pruebas y su valor sobre la nota final y criterios de evaluación para cada prueba

El estudiante deberá demostrar que ha alcanzado los resultados de aprendizaje previstos mediante las siguientes actividades de evaluacion

Seguimiento del trabajo realizado en las sesiones prácticas. Trabajo desarrollado en el aula informática relativo al análisis de datos [20%]. Resultados de aprendizaje: 1, 2, 3 y 4

Trabajo dirigido. Un proyecto individual o en grupo en el que se podrán en práctica los conocimientos y habilidades adquiridos en la asignatura. En la evaluación del trabajo tutorado propuesto a lo largo del cuatrimestre se tendrá en cuenta tanto la memoria presentada, como la idoneidad y originalidad de la solución propuesta. [70%]. Resultados de aprendizaje: 1 a 6

Presentación de forma oral y debate sobre el desarrollo y resultados del trabajo académico (10%). Resultados de aprendizaje: 1 a 6

El estudiante que no opte por el procedimiento de evaluación descrito anteriormente, no supere dichas pruebas durante el periodo docente o que quisiera mejorar su calificación tendrá derecho a realizar una prueba global que será programada dentro del periodo de exámenes correspondiente a la primera o segunda convocatoria.

4.Metodología, actividades de aprendizaje, programa y recursos

4.1.Presentación metodológica general

El proceso de aprendizaje que se ha diseñado para esta asignatura se basa en lo siguiente:

Las actividades de enseñanza y aprendizaje presenciales se basan en:

1. **Clase presencial.** Exposición de contenidos mediante presentación o explicación por parte de un profesor (posiblemente incluyendo demostraciones).
2. **Realización de trabajos prácticos de aplicación o investigación.** El estudiante ha de desarrollar individualmente un trabajo de aplicación de técnicas estadísticas en la resolución de problemas que involucren a amplias colecciones de datos. Se ofrece la opción de que el trabajo corresponda al análisis de una base de datos de interés para el estudiante o bien el estudio crítico de un artículo de investigación publicado y que haga uso de las técnicas presentadas en clase. En ambos casos se ha de elaborar una Memoria con todos los resultados que se entrega al profesor para su evaluación
3. **Laboratorio**. Actividades desarrolladas en espacios especiales con equipamiento especializado (laboratorio, aulas informáticas).
4. **Tutoría**. Período de instrucción realizado por un tutor con el objetivo de revisar y discutir los materiales y temas presentados en las clases.
5. **Evaluación**. Conjunto de pruebas escritas, orales, prácticas, proyectos, trabajos, etc. utilizados en la evaluación del progreso del estudiante

Las actividades de enseñanza y aprendizaje no presenciales se basan en:

1. **Trabajos prácticos.** Preparación de actividades para exponer o entregar en las clases prácticas.
2. **Estudio teórico.** Estudio de contenidos relacionados con las clases teóricas: incluye cualquier actividad de estudio

que no se haya computado en el apartado anterior (trabajo en biblioteca, lecturas complementarias, hacer problemas y ejercicios, etc.)

3. **Actividades complementarias**. Son tutorías no académicas y actividades formativas voluntarias relacionadas con la asignatura, pero no con la preparación de exámenes: lecturas, seminarios, jornadas, etc.

4.2.Actividades de aprendizaje

El programa que se ofrece al estudiante para ayudarle a lograr los resultados previstos comprende las siguientes actividades...

Clases magistrales interactivas

Se desarrollan en las sesiones de trabajo en el laboratorio informático. En esas sesiones se presentan situaciones reales que promuevan el interés por las distintas técnicas, los conceptos asociados, se muestra el tratamiento adecuado con el software y se incentiva la práctica de los propios alumnos.

Se utiliza el lenguaje R y se introducen las herramientas, funciones, librerías estándar disponibles en este software para abordar las distintas técnicas estadísticas.

Desarrollo de un proyecto

Cada estudiante ha de desarrollar un proyecto consistente en el análisis estadístico de una colección de datos, de alta dimensión, aplicando las técnicas oportunas para obtener conclusiones.

Trabajo del estudiante

La asignatura consta de 3 créditos ECTS que suponen una dedicación estimada por parte del alumno de 75 horas (35 horas presenciales y 40 horas no presenciales) distribuidas del siguiente modo:

- 30 horas, aproximadamente, de actividades presenciales (clases magistrales incluyendo seminarios profesionales, resolución de problemas y casos, y prácticas de laboratorio).
- 20 horas de trabajo en grupo.
- 20 horas de trabajo y estudio individual efectivo.
- 5 horas dedicadas a distintas pruebas de evaluación.

4.3.Programa

- Introducción
    - Aprendizaje estadístico.
    - Análisis exploratorio de datos.
    - Conceptos básicos de muestreo e inferencia estadística: estimación puntual y por intervalo, contrastes de hipótesis.
    - Verosimilitud: Estimación por máxima-verosimilitud, test de cociente de verosimilitudes.
    - Teoría estadística de la decisión. Métodos bayesianos.
    - El algoritmo EM. El método MCMC.
    - Simulación estadística
- Reconocimiento de relaciones explícitas: Modelos de regresión
    - Regresión lineal simple, crítica y validación del modelo, transformación Box-Cox, predicción.
    - Modelo lineal general, covariables y factores, análisis de la varianza.
    - Procedimientos automáticos de construcción de modelos: best subset, stepwise.
    - Validación, validación cruzada, métodos bootstrap.
    - Regresión en alta dimensionalidad.
    - Modelos con respuesta no gaussiana: GLM y GAM.
- Reconocimiento de patrones asistido: Regresión logística.
    - Modelos de regresión logística binaria.
    - Modelos de regresión logística multinomial.
    - Tabla de contingencia, modelos log-lineales.
- Reconocimiento de patrones no supervisado.
    - Análisis de conglomerados, método de k-medias.
    - Cluster jerárquico.

4.4.Planificación de las actividades de aprendizaje y calendario de fechas clave

Calendario de sesiones presenciales y presentación de trabajos

La asignatura se imparte en el cuatrimestre de otoño.

La organización docente prevista en las sesiones presenciales en el campus Río Ebro es de 2 horas en un laboratorio informático.

Se imparte en su totalidad en el aula informática con un doble propósito. Motivar el uso de las técnicas estadísticas desarrolladas en la asignatura a partir del análisis de datos y, simultáneamente, utilizar el software libre R para el tratamiento estadístico de datos por parte de los estudiantes.

En el Anillo Digital Docente se dispone de toda la documentación, el horario de todas las sesiones y un cronograma para la realización de los trabajos de la asignatura.

Los horarios de todas las clases y actividades se anunciarán con suficiente antelación a través de las webs del centro y de la asignatura.

Las fechas de entrega y seguimiento de los trabajos prácticos tutorados se comunican con suficiente antelación en clase y en la página web de la asignatura en el anillo digital docente, https://moodle.unizar.es/.

Las fechas de las pruebas de evaluación global son fijadas por la Escuela de Ingeniería y Arquitectura y publicadas en la página web del máster.

4.5.Bibliografía y recursos recomendados