Máster en Ingeniería Informática
62236 - Análisis avanzado de datos

Guía docente para el curso 2015 - 2016

Curso: , Semestre: , Créditos: 3.0

Información básica

Profesores

- **Jesús Asín Lafuente** jasin@unizar.es

- **María Dolores Berrade Ursúa** berrade@unizar.es

Recomendaciones para cursar esta asignatura

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

Actividades y fechas clave de la asignatura

El calendario de clases, prácticas y exámenes, así como las fechas de entrega de trabajos de evaluación, se anunciará con suficiente antelación.

Inicio

Resultados de aprendizaje que definen la asignatura

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

1:

Es capaz de 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:

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

3:

Es capaz de 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:

Es capaz de 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.

Introducción

Breve presentación de la asignatura

Dentro de la materia optativa de Big Data, la asignatura pone el foco en las técnicas estadísticas habituales en el análisis de datos, para las que además de su uso operativo, objetivo e interpretación de resultados, se proporcionan los algoritmos de cálculo que utilizan. En particular, se introduce la estimación y los contrastes de hipótesis basados en la log-verosimilitud y con herramientas bayesianas, considerando la existencia de varias poblaciones y varias respuestas continuas o categóricas. Como técnicas multivariantes de aprendizaje supervisado se presentan modelos de tipo regresión para respuestas continuas y categóricas. También se introducen procedimientos no supervisados para el reconocimiento de patrones, con técnicas de reducción de la dimensión.

Contexto y competencias

Sentido, contexto, relevancia y objetivos generales 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.

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.

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

1:

Afrontar con éxito los siguientes desempeños transversales:

1. Utilizar 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.
2. Aplicar de los conocimientos adquiridos y resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos.
3. Adquirir conocimientos avanzados y demostrados, 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Poseer las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

2:

Afrontar con éxito los siguientes desempeños relacionados con la Ingeniería Informática:

1. Comprender y poder aplicar conocimientos avanzados de computación de altas prestaciones y métodos numéricos o computacionales a problemas de ingeniería.
2. Aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento.
3. Conceptualizar, diseñar, desarrollar y evaluar la interacción persona-ordenador de productos, sistemas, aplicaciones y servicios informáticos

Importancia de los resultados de aprendizaje que se obtienen en la asignatura:

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 inferencia estadística permite establecer la significación de los resultados y su validez para toda la población.

Evaluación

Actividades de evaluación

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

1:

Seguimiento del trabajo realizado en las sesiones prácticas. Trabajo desarrollado en el aula informática relativo al análisis de datos [30%]. Resultados de aprendizaje: 1, 2, 3 y 4

2:

Proyecto. 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, 2, 3 y 4

Actividades y recursos

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 (estudiar exámenes, 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 o con la calificación: lecturas, seminarios, jornadas, vídeos, etc.

Actividades de aprendizaje programadas (Se incluye programa)

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

1:

Contenidos a desarrollar

- 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.
- 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.
- Reconomiento de patrones no supervisado.
  - Análisis de conglomerados, método de k-medias.
  - Cluster jerárquico.

2:

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.

Planificación y calendario

Calendario de sesiones presenciales y presentación de trabajos

La organización docente prevista de las sesiones presenciales en el campus Río Ebro es la siguiente:

- Clases magistrales.
- Resolución de problemas y casos.
- Prácticas de laboratorio.

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

Los proyectos propuestos serán entregados al finalizar el cuatrimestre, en las fechas que se señalen.

Referencias bibliográficas de la bibliografía recomendada