ASIGNATURA DE GRADO:

MINERÍA DE DATOS (ING.TI)

Curso 2015/2016

(Código:71024062)

1.PRESENTACIÓN DE LA ASIGNATURA

Minería de Datos es sólo una de las denominaciones (la más popular, quizás, en el ámbito empresarial) de un área de investigación que podríamos llamar con más propiedad, Descubrimiento de Conocimiento a partir de datos. Corresponde con lo que desde antiguo se conoce como el principio de inducción en términos filosóficos.

¿En qué consiste entonces la Minería de Datos? Se trata de conseguir reproducir con computadoras, tareas genuinamente humanas relacionadas con la extracción de conocimiento a partir de datos. Esas tareas pueden ser de varios tipos. Uno de ellos agrupa tareas en las que la computadora debe aprender a partir de un conjunto de ejemplos, generalizar las relaciones entre ellos, y aplicar el modelo resultante del aprendizaje a datos nuevos. La clasificación de casos en categorías responde bien a este patrón de tareas, pero también las actividades de control, en las que la máquina debe aprender a (generar un modelo para) controlar un sistema con unos objetivos explícitos, en problemas de planificación o de asignación de recursos, o las tareas de predicción, en las que el modelo aprendido a partir de los datos nos ayuda a inferir nuevos valores de unas variables desconocidas.

En todos estos casos, vemos la importancia que desempeñan los datos en este área. Se trata de producir modelos a partir de ejemplos que condensan el conocimiento que queremos aprehender, y para los que no disponemos de un modelo de conocimiento alternativo, expresado en lenguaje natural o estructurado.

Otro tipo de tareas encuadradas en la Minería de Datos, pero que no veremos en este curso, abordan la tarea de descubrir conceptos, relaciones o reglas en conjuntos de datos no etiquetados. Mientras que en el caso anterior (tareas de clasificación o regresión) disponemos de ejemplos que expresan los modelos que deseamos inferir, en las tareas de este tipo los datos están desnudos, y nuestra tarea consiste precisamente en descubrir esquemas clasificatorios, patrones repetidos, relaciones entre ellos, agrupamientos de datos o reglas que describan la distribución de los datos en un espacio de representación dado.

En este curso vamos a abordar los fundamentos del área. El objetivo del equipo docente ha sido, no abordar de manera extensiva pero superficial las diversas técnicas que se aplican en el área, sino proporcionar al estudiante los fundamentos que le permitan explorar en asignaturas sucesivas o por su cuenta, todas esas técnicas en las que aquí no podremos profundizar. Así pues, empezamos la casa por sus cimientos. Y los cimientos de el edificio de la Minería de Datos son principalmente matemáticos y probabilísticos.

2.CONTEXTUALIZACIÓN EN EL PLAN DE ESTUDIOS

Esta asignatura, como podréis comprobar en la memoria de la titulación, se corresponde con la materia denominada Sistemas de Información, que comparte con las asignaturas de Bases de Datos y Gestión de Bases de Datos. Para su aprendizaje no es estrictamente necesario haber cursado las anteriores, pues lo que aquí se enseña se hace de manera independiente del sistema de almacenamiento de los datos. Sin embargo, sí es muy importante haber cursado las asignaturas de Fundamentos Matemáticos y Estadística.

Los conocimientos adquiridos a través de esta asignatura son los fundamentos de un área cuya exploración continúa en el Master de Inteligencia Artificial Avanzada, en las asignaturas relacionadas con la Minería de Datos. En ellas, se aplican todo lo aprendido aquí para entender las variadas técnicas avanzadas (como Máquinas de Vectores Soporte, Procesos Gaussianos, Redes Neuronales Artificiales, etc) y para adentrarnos en el mundo de la clasificación no supervisada o agrupamiento.

Existen multitud de vías en las que los conocimientos adquiridos aquí serán de utilidad en el futuro de los estudiantes. El aprendizaje estadístico (otra de las denominaciones de la Minería de Datos) abre un sinfín de perspectivas nuevas en una nueva era en la que los datos, en muchas ocasiones, desbordan la capacidad de los humanos de procesar información. Desde lo que se conoce como el cuarto paradigma de la Ciencia (o e-Ciencia, en una expresión poco afortunada) de aplicación en áreas como las bio-tecnologías o las grandes bases de datos científicas, a las aplicaciones empresariales en bancos o librerías virtuales, sin olvidar a los buscadores web. Lo que aquí aprenderemos es de aplicación general a todos esos campos, precisamente porque se trata de los fundamentos del área.

Las competencias que se trabajarán en esta asignatura son las siguientes:

(G.5) Competencias en el uso de las herramientas y recursos de la Sociedad del Conocimiento: Manejo de las TIC. Competencia en la búsqueda de información relevante. Competencia en la gestión y organización de la información. Competencia en la recolección de datos, el manejo de bases de datos y su presentación

FB.1 Capacidad para la resolución de los problemas matemáticos que puedan plantearse en la ingeniería. Aptitud para aplicar los conocimientos sobre: álgebra lineal, cálculo diferencial e integral, métodos numéricos, algorítmica numérica y estadística y optimización.

BC.15 Conocimiento y aplicación de los principios fundamentales y técnicas básicas de los sistemas inteligentes y su aplicación práctica.

BTEc.5 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 relacionados con aspectos de computación, percepción y actuando en ámbitos o entornos inteligentes.

3.REQUISITOS PREVIOS REQUERIDOS PARA CURSAR LA ASIGNATURA

Es necesario tener conocimientos básicos de Matemáticas (Análisis y Álgebra Matricial) y Estadística, adquiridos a través de las asignaturas de Fundamentos Matemáticos y Estadística.

4.RESULTADOS DE APRENDIZAJE

El estudiante, al concluir y aprobar la asignatura, dominará los conceptos básicos del área de la Minería de Datos: el análisis probabilístico del problema de reconocimiento de patrones en tareas de clasificación y regresión; su tratamiento desde la perspectiva bayesiana, con el manejo de las entidades que desempeñan papeles relevantes en ese ámbito (verosimilitud, probabilidades marginales, evidencias...); los problemas de la selección de modelos de complejidad creciente, y la maldición de la dimensionalidad. Habrá, asimismo, adquirido destreza en el manejo de probabilidades y de entidades algebraicas, principalmente matriciales, suficiente como para profundizar los contenidos de la materia según se propone en los ejercicios no resultos que propondrá el equipo docente.

En resumen, el estudiante habrá interiorizado los fundamentos del área a través de las aproximaciones más simples (los modelos lineales) y se hallará en disposición de abordar el estudio posterior de las técnicas más avanzadas (y no necesariamente a través de modelos lineales) como las Máquinas de Vectores Soporte, los Procesos Gaussianos o las Redes Neuronales (por citar sólo dos ejemplos)

5.CONTENIDOS DE LA ASIGNATURA

La asignatura consta de tres bloques básicos:

1 - Introducción. La introducción repasa los conceptos básicos de estadística e introduce los conceptos y técnicas que se manejarán durante el curso. Consta de varios sub-bloques, entre los que cabe destacar los dedicados a i) conceptos básicos

de probabilidad, ii) selección de modelos, iii) la maldición de la dimensionalidad, iv) teoría de la decisión, v) teoría de la información y vi) distribuciones de probabilidad.

2 - Modelos lineales de regresión: En el segundo bloque se abordará la obtención de modelos lineales para tareas de regresión. Para ello, haremos uso de los conceptos introducidos anteriormente. El bloque se subdivide en varios sub-bloques: regresión mediante funciones de base lineales, la descomposición sesgo-varianza, regresión lineal bayesiana, comparación bayesiana de modelos, la aproximación de la evidencia.

3 - Modelos lineales de clasificación: En el tercer bloque, cerramos las técnicas supervisadas de Minería de Datos con la aplicación de todo lo aprendido hasta ahora al caso particular de modelos en los que la variable que debemos predecir es categórica. El problema lo abordaremos desde tres perspectivas diferentes: los modelos no probabilísticos, y los probabilísticos discriminantes y generativos. El esquema del bloque es el siguiente: funciones discriminantes, modelos probabilísticos generativos, modelos probabilísticos discriminantes, la aproximación de Laplace y la regresión logística bayesiana.

6.EQUIPO DOCENTE

- LUIS MANUEL SARRO BARO

7.METODOLOGÍA Y ACTIVIDADES DE APRENDIZAJE

La asignatura se cursa de una manera clásica en la educación a distancia. Al tratarse de una asignatura de fundamentos, se concede una importancia especial a los aspectos teóricos, y las actividades prácticas están supeditadas a la consolidación de los conceptos teóricos.

La interacción con el equipo docente se realizará principalmente a través de la plataforma de aprendizaje virtual de la asignatura, donde se pretende que sean los propios alumnos los que resuelvan sus dudas (con la ayuda y supervisión en todo momento del profesor) de manera colaborativa. Todo ello, desde la convicción de que lo que se descubre se aprende mucho mejor que lo que se asimila de forma pasiva. El equipo docente valorará muy positivamente la participación en los foros con mensajes que colaboren en la resolución de problemas y dudas.

El 5% de los créditos asignados se destina a la preparación para el estudio del contenido teórico, lo que incluye la lectura de las orientaciones y una primera lectura del índice del texto base.

El segundo bloque, el más importante en cuanto a fracción del total (80%) lo constituye el estudio de los contenidos teóricos (60%) y el desarrollo de ejercicios de consolidación de lo aprendido (20%) mediante la resolución de problemas propuestos cuya respuesta estará a disposición de los alumnos (ejercicios de auto-evaluación).

Finalmente, el tercer bloque se asigna a una práctica entregable a la que corresponde el 10% de la nota (y el 15% de los créditos de la asignatura). Consistirá en la resolución de tres de los ejercicios propuestos en el texto base pero cuya solución no está disponible a través de Internet (ejercicios de descubrimiento).

8.EVALUACIÓN

El equipo docente propondrá durante el curso dos subconjuntos de ejercicios tomados del texto base. El primero estará compuesto de ejercicios cuya respuesta se encuentra disponible en el sitio web del libro; el segundo contendrá ejercicios cuya respuesta no está disponible a través de Internet (ejercicios de descubrimiento) y de entre los que se seleccionarán los enunciados del examen. El estudiante podrá auto-evaluarse al final de cada bloque mediante los ejercicios resueltos, comprobando si sus soluciones corresponden con las correctas, y mediante los ejercicios de descubrimiento intentando llegar a las soluciones sin guía.

La evaluación propiamente dicha se realizará por parte del equipo docente mediante una prueba presencial cuya nota constituirá el 90% de la calificación final, y por los tutores de la asignatura, que evaluarán

- un conjunto de tres ejercicios a elegir por el estudiante de entre los ejercicios de descubrimiento propuestos por el equipo docente (10% de la calificación final).
- una actividad práctica voluntaria propuesta por el equipo docente, y que podrá sumar hasta un 20% (2 puntos sobre 10) a la nota obtenida en conjunto (prueba presencial más nota de los 3 ejercicios), siempre que ésta supere los 4 puntos.

La prueba presencial se centrará en los aspectos teóricos y matemáticos de la asignatura, sin que sea necesario desarrollar problemas numéricos en detalle. Los enunciados se elegirán de entre los ejercicios de auto-evaluación (descubrimiento) propuestos en el texto base. Los aspectos más prácticos (de programación numérica) se evaluarán en la actividad voluntaria.

En la prueba presencial se evaluará entre otras cosas la claridad y concisión en la exposición de los conceptos clave de la asignatura, y de manera especial la rigurosidad en el empleo de dichos conceptos. Se penalizará el uso de expresiones poco claras, ambiguas o el lenguaje no técnico. Las respuestas deben expresar con claridad que el alumno ha entendido e interiorizado los conceptos y técnicas descritos en los contenidos de la asignatura, por lo que no se tendrán en cuenta respuestas que sean en lo esencial transliteraciones del texto base.

9.BIBLIOGRAFÍA BÁSICA

ISBN(13): 9780387310732
Título: PATTERN RECOGNITION AND MACHINE LEARNING
Autor/es: Christopher M. Bishop ;
Editorial: Springer

Buscarlo en Editorial UNED

Buscarlo en libreria virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

10.BIBLIOGRAFÍA COMPLEMENTARIA

ISBN(13): 9780387848587
Título: THE ELEMENTS OF STATISTICAL LEARNING
Autor/es: Hastie, Trevor ; Tibshirani, Robert J. ; Friedman, Jerome ;
Editorial: Springer

Buscarlo en libreria virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

Buscarlo en Catálogo del Patrimonio Bibliográfico

11.RECURSOS DE APOYO

El curso se desarrolla fundamentalmente a través de la plataforma aLF de la UNED. La información y el material complementario se encuentra en dicho curso, y la interacción con el equipo docente se desarrollará principalmente a través de los foros de la asignatura en dicha plataforma. Por supuesto, el equipo docente también atenderá a los alumnos a través del teléfono (913988715) o de manera presencial en el horario de guardia (Lunes de 10:30 a 14:30). Es recomendable acordar una cita previamente. Finalmente, el equipo docente estará también disponible a través de software de videoconferencia, preferiblemente skype. De nuevo, será necesario concertar una cita con anterioridad.

El equipo docente será el responsable de responder a las dudas que surjan sobre el funcionamiento de la asignatura y sobre los contenidos teórico de ésta (siempre que sea posible, a través de los foros, pues de esta manera las respuestas quedan a disposición de otros alumnos que puedan compartirlas). En principio, y salvo circunstancias excepcionales que lo impidan, el tiempo máximo de espera para las respuestas a las preguntas del foro es de 7 días, (el tiempo entre guardia y guardia). Por regla general, nunca se alcanza ese periodo y en la medida de lo posible el equipo docente intenta responder a las cuestiones con la máxima celeridad que permiten las otras obligaciones del profesorado, entre las que cabe destacar la docencia en otras asignaturas y las tareas de investigación y administración. Como orientación, se puede decir que en periodo lectivo, fuera de épocas de examen o viajes al extranjero para reuniones o congresos, el tiempo de espera no debe rebasar las 48 horas.

Es muy importante que el alumno que solicite una respuesta directa del equipo docente lo haga constar en su mensaje al foro.

12.TUTORIZACIÓN

1. Equipo docente (en la sede central):

   Dr. D Luis Manuel Sarro Baro

   Lunes 10:30 a 14:30. Despacho 3.12. Tel.: 913988715. lsb@dia.uned.es

   La dirección de contacto es:

   ETSI Informática-UNED. Dpto. Inteligencia Artificial

   c/Juan del Rosal, 16

   28040 Madrid

2. Profesores tutores (en el centro asociado correspondiente). Los horarios de atención del tutor serán suministrados por los propios centros asociados al inicio de curso.