ASIGNATURA DE MÁSTER:

MINERÍA DE DATOS

Curso 2016/2017

(Código:31101061)

1.PRESENTACIÓN

El presente curso pretende dar una visión panorámica de la teoría y conceptos fundamentales utilizados en Minería de Datos (MD), del conjunto de tareas abordadas por esta disciplina y del repertorio de técnicas y métodos existentes que permiten resolver cada una de estas tareas.

La asignatura se enmarca dentro del Programa de Posgrado en Inteligencia Artificial y Sistemas Informáticos impartido por la Escuela Técnica Superior de Ingeniería Informática de la UNED.

Dentro de este posgrado se imparte tanto en la especialidad de "Sistemas Inteligentes de Diagnóstico, Planificación y Control (Máster en Inteligencia Artificial Avanzada)" como en el de "Tecnologías del Lenguaje en la Web (Master en Lenguajes y Sistemas Informáticos)"

Ficha técnica:

- Tipo: Optativa
- Duración: Anual
- Créditos Totales y Horas: 6 / 150
- Horas de estudio teórico: 55
- Horas de trabajo práctico: 50
- Horas de actividades complementarias: 45

2.CONTEXTUALIZACIÓN

Esta asignatura es común a los dos programas de Master de este posgrado. Así, dentro de la titulación del Master "Lenguajes y Sistemas Informáticos" se encuadra dentro del módulo denominado ESP-LSI-1: Tecnologías del lenguaje en la web. De otra lado, dentro del programa de Master "IA Avanzada. Fundamentos Métodos y Aplicaciones" pertenece al módulo denominado "ESP-IA.1: Sistemas Inteligentes de diagnóstico, planificación y control".

Existen distintas asignaturas en el resto del programa de ambos master relacionadas con esta asignatura. Así, "Métodos de Aprendizaje en IA" aborda, además de otras técnicas de aprendizaje, la mayoría de las técnicas que se estudiarán en este tema y que básicamente se encuadran dentro del denominado paradigma de aprendizaje inductivo. El alumno que haya cursado dicha asignatura tendrá mucho camino adelantado al abordar esta asignatura. No obstante, hay que tener en cuenta que la visión que allí se da está orientada eminentemente a la parte algorítmica y de implementación (programación) de cada técnica. Aquí, el enfoque está más orientado a su uso, independientemente de la implementación particular. Es decir, consideraremos el conjunto de técnicas como una biblioteca de componentes reutilizables, cada uno de los cuales será seleccionado de acuerdo a las características de la tarea que se requiere resolver. En otros casos, esta asignatura puede servir de introducción a otras asignaturas de este programa de posgrado, tales como "Descubrimiento de información en textos" o "Minería en la Web", ambas pertenecientes al módulo "ESP-LSI-1".

3.REQUISITOS PREVIOS RECOMENDABLES

El alumno debe haber cursado las asignaturas de Álgebra, Análisis Matemático y Estadística impartidas en el primer ciclo de la titulación de Informática de la UNED o asignaturas equivalentes en otras universidades.

En particular, debe haber adquirido competencias básicas en el manejo algebraico de matrices, cálculo de determinantes , inversión de matrices y diagonalización de éstas. Debe ser capaz de calcular con soltura derivadas parciales e integrales de funciones multivariantes (Análisis Matemático). Finalmente, debe conocer conceptos básicos de Estadística como las propiedades de la distribución gaussiana multivariante o los tests estadísticos de contraste de hipótesis.

4.RESULTADOS DE APRENDIZAJE

Destrezas y competencias

- Conocer las relaciones existentes de la MD con otras disciplinas.
- Conocer las distintas fases implicadas en un proyecto de minería de datos y las relaciones existentes entre ellas.
- Conocer y saber aplicar las distintas técnicas existentes en MD para realizar preparación de datos.
- Distinguir entre tarea, técnica y método en MD.
- Saber relacionar las distintas tareas propias de MD con las técnicas que permiten resolverlas.
- Conocer que tipo de tarea es capaz de abordar cada técnica de MD.
- Conocer varios tipos de algoritmos o métodos para cada técnica de MD.
- Dominar, tanto desde un punto de vista teórico como práctico, los distintas técnicas/algoritmos utilizados en MD.
- Aplicar técnicas de evaluación adecuadas en función del tipo de modelo a evaluar.
- Practicar con algunas de las herramientas software de minería de datos.
- Afrontar la solución de un proyecto de MD siempre desde un punto de vista metodológico o ingenieril, nunca como un arte.
- Conocer y aplicar las metodologías de MD dedicadas a la creación y seguimiento de un proyecto de minería de datos.
- Saber responder a la pregunta de: ¿Cuándo implantar un proyecto de minería de datos en una organización?
- Conocer las repercusiones de la MD en distintos campos: social, legal y ético.
- Conocer los retos que plantea la MD actualmente y las tendencias futuras.

5.CONTENIDOS DE LA ASIGNATURA

2.1 Estructura y contenido teórico

1.INTRODUCCIÓN

1.1.El concepto de Minería de Datos

1.2.La minería de datos y el proceso de descubrimiento de conocimiento a partir de datos

1.3.Relación con otras disciplinas

1.4.Aplicaciones

1.5.Fases del proceso de extracción de conocimiento a partir de datos

2.PREPARACIÓN DE DATOS

2.1.Consideraciones previas generales. Los almacenes de datos.

2.2.Técnicas sencillas de preprocesado

2.2.1.Compleción (datos faltantes)

2.2.2.Limpieza de errores

2.2.3.Transformación de atributos

2.2.4.Escalado

2.2.5.Discretización

2.2.6.Numerización

2.3.Técnicas de reducción de la dimensionalidad I: Análisis de Componentes Principales.

2.4.Técnicas de reducción de la dimensionalidad II: Métodos de Filtrado y Envoltura

3.TAREAS Y TÉCNICAS DE MINERÍA DE DATOS

3.1.Tareas en minería de datos.

3.2.Correspondencia entre métodos y tareas.

3.3.Caracterización de las técnicas de minería de datos.

3.4.Técnicas de Minería de Datos

3.4.1.Métodos estadísticos.

3.4.2.Reglas de asociación y dependencia.

3.4.3.Métodos Bayesianos.

3.4.4.Árboles de Decisión y sistemas de reglas.

3.4.5.Redes Neuronales Artificiales.

3.4.6.Máquinas de vectores soporte.

3.4.7.Extracción de conocimiento con algoritmos evolutivos y reglas difusas.

3.4.8.Métodos basados en casos y vecindad.

4.EVALUACIÓN

4.1.Consideraciones generales.

4.2.Técnicas básicas de evaluación de clasificadores

4.2.1.Medidas de la calidad de un clasificador: la tasa de errores

4.2.2.La descomposición del error en sesgo y varianza: el concepto de generalización

4.2.3.El sobreentrenamiento

4.2.4.Repetibilidad estadística: la validación cruzada.

4.3.Aspectos específicos de la evaluación de los diferentes clasificadores estudiados

4.4.Técnicas estadísticas de comparación de clasificadores

4.5.Medidas de calidad de agrupamiento

4.6.Interpretación, difusión y uso de modelos

5.IMPLANTACIÓN E IMPACTO DE LA MINERÍA DE DATOS

5.1.Implantación de un Programa de Minería de Datos (PMD) en una organización

5.1.1.Cuándo implantar un PMD: Necesidades y objetivos

5.1.2.Fases de un PMD: Estándar CRISP-DM

5.1.3.Integración de un PMD dentro de una organización

5.1.4.Recursos necesarios

5.2.Repercusiones y retos de la minería de datos

5.2.1.Impacto social

5.2.2.Cuestiones éticas y legales

5.2.3.Problemas y soluciones: Tendencias futuras

6.EQUIPO DOCENTE

- LUIS MANUEL SARRO BARO
- JOSE LUIS AZNARTE MELLADO

7.METODOLOGÍA

La general del programa de postgrado adaptada a las directrices del EEES, de acuerdo con el documento del IUED. Junto a las actividades y enlaces con fuentes de información externas, existe material didáctico propio preparado por el equipo docente. La asignatura no tiene clases presenciales. Los contenidos teóricos se impartirán a distancia, de acuerdo con las normas y estructuras de soporte telemático de la enseñanza en la UNED.

En particular, en la asignatura se abordarán de manera secuencial las diversas fases del proceso de descubrimiento de conocimiento desde el punto de vista algorítmico, de manera que es conveniente seguir los contenidos de manera igualmente secuencial. Cada tema (excepto el primero) viene acompañado de una o varias actividades cuya memoria servirá de base para la evaluación. Recomendamos leer primero los contenidos teóricos de cada tema (y especificos de cada actividad) antes de abordar las actividades.

No es necesario memorizar expresamente los contenidos del temario (no hay examen presencial de la asignatura), pero el quipo docente hará especial énfasis en la comprensión de los contenidos mostrada en las actividades. Éstas están diseñadas de manera que el/la estudiante debe realizar una tarea importante de contextualización y análisis. Si el/la estudiante se limita a generar resultados sin demostrar la comprensión de los conceptos en la discusión de dichos resultados se

considerará que la práctica es insuficiente.

8.BIBLIOGRAFÍA BÁSICA

ISBN(13): 9788420540917
Título: INTRODUCCIÓN A LA MINERÍA DE DATOS (1ª)
Autor/es: Ferri Ramírez, César ; Ramírez Quintana, Mª José ; Hernández Orallo, José ;
Editorial: PEARSON

Buscarlo en libreria virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

Buscarlo en Catálogo del Patrimonio Bibliográfico

Comentarios y anexos:
El material docente del presente curso está compuesto por el texto base indicado en la bibliografía básica, por textos alternativos indicados en la bibliografía general de consulta, por los artículos referenciados en las actividades y en los epígrafes "Orientaciones", pertenecientes al desglose que se hace más adelante de cada tema por separado y, finalmente, por aquellas herramientas software indicadas en algunas de las actividades a realizar.
El texto base será el hilo conductor para el estudio de los contenidos de este curso. No obstante, dado el carácter introductorio de dicho texto, existirán algunas cuestiones que será necesario ampliar mediante la lectura de bibliografía alternativa.
Tratándose de un master orientado a la investigación, las actividades de aprendizaje se pueden estructurar tanto desde un punto de vista teórico como práctico. En el primer caso, girarán en torno al estado del arte en cada una de las materias del curso y, en el segundo caso, lo harán en relación con la búsqueda de soluciones de distintos subproblemas propios del campo de la MD.

9.BIBLIOGRAFÍA COMPLEMENTARIA

Comentarios y anexos:

Materiales y recursos de apoyo

Además de la bibliografía indicada anteriormente, los materiales de apoyo para la realización de las prácticas serán los siguientes:
De manera general, las prácticas se realizarán con el programa Weka, descargable de la dirección http://www.cs.waikato.ac.nz
Excepcionalmente, las prácticas sobre redes neuronales se realizarán con:
-SNNS (http://www-ra.informatik.uni-tuebingen.de/SNNS/) o
-JavaNNS (http://wwwra.informatik.uni-tuebingen.de/software/JavaNNS/) y
-SOMPAK (http://www.cis.hut.fi/research/som_lvq_pak.shtml).
Los ficheros con los datos de trabajo serán proporcionados por el equipo docente a través de la plataforma alF o formarán parte de la distribución del software empleado. Si no se indica que la actividad correspondiente haya de ser realizada con un conjunto de datos particular, el alumno podrá elegir un fichero de casos del repositorio de la Universidad de California Irvine http://kdd.ics.uci.edu/.
La plataforma de e-Learning aLF, proporcionará el adecuado interfaz de interacción entre el alumno y sus profesores. Esta plataforma colaborativa permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.
Se ofrecerán las herramientas necesarias para que, tanto el equipo docente como el alumnado, encuentren la manera de compaginar tanto el trabajo individual como el aprendizaje cooperativo.

BIbliografia general de consulta

-J. Han, M.Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
-H. Witten, E. Frank, Data mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann Publishers, 2005.
-B.Pyle, Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999
-C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
Además, véase la incluida en la descripción de las actividades.

10.RECURSOS DE APOYO AL ESTUDIO

Ver la sección Comentarios y anexos de Bibliografía complementaria.

11.TUTORIZACIÓN Y SEGUIMIENTO

La tutorización de los alumnos se llevará a cabo exclusivamente a través de la plataforma de e-learning Alf.

Los hoararios de guardia de los profesores son:

Luis M. Sarro Baro

Lunes, de 10:30 a 14:30

Emilio Letón Molina

Lunes, de 15:00 a 19:00

12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación de los aprendizajes se realizará mediante la corrección de las siguientes actividades, que el alumno deberá entregar en plazo. La evaluación de dichas actividades corresponde al 80% de la nota final. Recordamos de nuevo un aspecto fundamental: en las actividades no se trata de generar resultados sino de analizarlos a la luz de los conceptos fundamentales del área. Además, se tendrá en cuenta la participación del alumno en los foros, planteando cuestiones avanzadas y sobre todo generando discusión y proponiendo hipótesis de análisis sobre las prácticas. Este tipo de participación en los foros constituye el 10% de la nota final, siempre y cuando tenga lugar durante el periodo lectivo del curso. Finalmente, el 10% restante se puede obtener realizando un trabajo de investigación bibliográfica avanzada sobre algún tema acordado con el equipo docente.

Se considerará que el estudiante se ha presentado a la convocatoria sólo si ha entregado todas las prácticas, en cuyo caso se procederá a la corrección. Para aprobar la asignatura se han de haber aprobado todas y cada una de las prácticas (calificación superior a 5 sobre 10).

Actividades prácticas programadas

Tema 1

Sin actividades.

Tema 2:

Actividad 2.1: Ejercicios de simulación

El estudiante generará un conjunto de datos artificial compuesto por 100 instancias caracterizadas por una variable relevante en sentido fuerte, tres variables relevantes en sentido débil y una variable totalmente irrelevante. Esta última se puede generar mediante números aleatorios extraídos de una distribución de probabilidad uniforme o normal (gaussiana). Como indicación sugerimos extender el ejemplo XOR a tres dimensiones. A continuación, aplicará diferentes técnicas de selección de variables disponibles en weka (un mínimo de tres de filtrado, el análisis de componentes principales y la técnica de envoltura, WrapperSubsetEval, con BayesNet como clasificador y empleando todos los valores por defecto, salvo el número máximo de padres que se debe modificar a 3).

Entregables:

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 a una cara con los siguientes apartados:

1.Descripción del experimento.

2.Tabla de resultados obtenidos para las 5 aproximaciones.

3.Discusión de los resultados.

Actividad 2.2: Estudio de bibliografía avanzada

En esta actividad el alumno deberá elegir uno de los artículos del especial del Journal of Machine Learning Research sobre "Variable and Feature Selection" (http://jmlr.csail.mit.edu/papers/special/feature03.html).

La lista de artículos del número especial es la siguiente:

1.Distributional Word Clusters vs. Words for Text Categorization (Kernel Machines Section). Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, Yoad Winter.

2.Extensions to Metric Based Model Selection.Yoshua Bengio, Nicolas Chapados.

3.Dimensionality Reduction via Sparse Support Vector Machines. Jinbo Bi, Kristin Bennett, Mark Embrechts, Curt Breneman, Minghu Song.

4.Benefitting from the Variables that Variable Selection Discards. Rich Caruana, Virginia R. de Sa.

5.A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification. Inderjit S. Dhillon, Subramanyam Mallela, Rahul Kumar.

6.An Extensive Empirical Study of Feature Selection Metrics for Text Classification. George Forman.

7.Sufficient Dimensionality Reduction. Amir Globerson, Naftali Tishby.ç

8.Grafting: Fast, Incremental Feature Selection by Gradient Descent in Function Space. Simon Perkins, Kevin Lacker, James Theiler.

9.Variable Selection Using SVM based Criteria. Alain Rakotomamonjy.

10.Overfitting in Making Comparisons Between Variable Selection Methods. Juha Reunanen.

11.MLPs (Mono Layer Polynomials and Multi Layer Perceptrons) for Nonlinear Modeling. Isabelle Rivals, Léon Personnaz.

12.Ranking a Random Feature for Variable and Feature Selection. Hervé Stoppiglia, Gérard Dreyfus, Rémi Dubois, Yacine Oussar.

13.Feature Extraction by Non Parametric Mutual Information Maximization.Kari Torkkola.

14.Use of the Zero Norm with Linear Models and Kernel Methods. Jason Weston, André Elisseef, Bernhard Schölkopf, Mike Tipping.

Algunos de los artículos de la lista anterior presuponen conocimientos sobre técnicas que se describirán en detalle en el tema 3. Los alumnos que opten por este tipo de artículos deberán hacer el esfuerzo adicional de adelantarse al temario y estudiar la técnica en cuestión antes de comentar el artículo. Por ello, recomendamos una lectura de todos los abstracts o resúmenes y una selección cuidadosa del artículo sobre el que tratará el entregable.

Entregables:

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 a una cara con los siguientes apartados:

1.Una justificación breve sobre los motivos para la elección del artículo.

2.Un resumen de la aportación novedosa frente a trabajos anteriores citados en el propio artículo. ¿Qué ventajas comparativas presenta la contribución?

3.Un estudio sobre el ámbito de aplicabilidad de las conclusiones obtenidas (para qué tipo de datos/algoritmos está especialmente indicado, limitaciones, en qué situaciones está contraindicado...).

4.Un estudio de la bibliografía reciente del autor y el área. El estudiante puede hacer el estudio comenzando con una búsqueda por autor en el servidor citeseer (http://citeseer.ist.psu.edu/). Con los resultados, deberá realizar una selección de publicaciones relacionadas con el tema de la selección de atributos y, en particular, con la aproximación elegida, y analizar su impacto medido por el número de citaciones. Finalmente, el informe deberá recoger publicaciones de otros autores relacionadas con el artículo original, de publicaciones de relevancia y los mayores índices de citación encontrados.

Bibliografía asociada:

-Sistemas Basados en el Conocimiento II. Introducción a la Neurocomputación. Disponible en: http://www.ia.uned.es/asignaturas/sbc2/sbc2/libro/book.pdf

-JMLR Special Issue on Variable and Feature Selection. Artículos disponibles en http://jmlr.csail.mit.edu/papers/special/feature03.html

-Kohavi, R. & John, G.H., Wrappers for Feature Subset Selection (1997). Disponible en: http://citeseer.ist.psu.edu/13663.html

-Para las definiciones estadísticas comunes o de teoría de la información (información mutua, ganacia de información o entropía cruzada) se pueden consultar las entradas correspondientes de la enciclopedia matemática on-line Mathworld http://mathworld.wolfram.com de la wikipedia, http://en.wikipedia.org/

Tema 3

Se propone realizar un conjunto de actividades prácticas relacionadas con la resolución de diferentes tipos de problemas de minería de datos. El alumno se familiarizará así con el uso de las distintas técnicas estudiadas en este tema. Para ello, se utilizará Weka, un entorno que proporciona una interfaz gráfica desde la cual se puede acceder a una colección de algoritmos estándares de aprendizaje automático para tareas de data mining. Además, soporta también herramientas para procesado y visualización de datos. Finalmente, una característica destacable de Weka es que es de uso libre y código abierto (open source) bajo licencia GNU y está desarrollada enteramente en Java (multiplataforma). El conjunto de prácticas a realizar están contenidas en un documento accesible y descargable desde el curso de la asignatura ubicado en la plataforma aLF y giran en torno a los siguientes contenidos:

Actividad 3.1: Redes Neuronales I. Clasificación

Actividad 3.2: Redes Neuronales II. Mapas Autoorganizados.

Actividad 3.3: Máquinas de Vectores Soporte.

Actividad 3.4:Clustering: Algoritmo K-medias.

Tema 4

Actividad 4.1: Ejercicios de simulación

El estudiante utilizará weka para generar 10 particiones de 10 bloques del conjunto de datos "iris.arff" proporcionado junto con el software de la Universidad de Waikato. Para cada partición, deberá realizar un experimento de validación cruzada con un clasificador basado en redes bayesianas y otro en árboles de decisión, y deberá ordenar los resultados de mayor a menor en una lista. Deberá promediar los resultados de cada experimento y, con las dos listas ordenadas de los promedios (una para los clasificadores bayesianos y otra para los árboles de decisión), deberá realizar un test t de Student que determine si existen diferencias estadísticas entre los resultados obtenidos.

Entregables:

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 a una cara con los siguientes apartados:

1.Descripción del experimento.

2.Tablas ordenadas de cada uno de los 10 experimentos de validación cruzada para cada clasificador.

3.Valores promediados de la tabla anterior.

4.Cálculo de la tasa media de error y su varianza para cada clasificador y resultado del test de Student.

La distribución t de Student se puede obtener de muchas fuentes. En particular, el estudiante puede hallarla implementada en la librería gsl de GNU para c/c++.

Actividad 4.2: Estudio de bibliografía avanzada

En esta actividad el alumno debe leer el texto "ROC graphs: Practical considerations for Researchers". En él se expone una aproximación alternativa/complementaria a la forma habitual de evaluar los modelos (a través de la tasa de errores de clasificación, la suma cuadrática de los errores de regresión o medidas equivalentes) denominada AUC (Area Under Curve). La curva a la que hace referencia el nombre es la Receiver Operating Characteristic Curve y el mismo artículo expone sus fundamentos (procedentes de Teoría de la Señal) y la forma de calcularla.

Entregables:

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 a una cara con los siguientes apartados:

1.Un resumen de los principios del análisis AUC/ROC

2.Un análisis de las diferencias con el método clásico de estimar el error de clasificación/regresión. Ventajas/inconvenientes de cada aproximación.

3.Un estudio de la bibliografía reciente del autor y el área. El estudiante puede hacer el estudio comenzando con una búsqueda por autor en el servidor citeseer. Con los resultados, deberá realizar una selección de publicaciones relacionadas con el tema de los análisis AUC/ROC y analizar su impacto medido por el número de citaciones. Finalmente, el informe deberá recoger publicaciones de otros autores relacionadas con el artículo original, de publicaciones de relevancia y los mayores índices de citación encontrados.

Bibliografía asociada

-Sistemas Basados en el Conocimiento II. Introducción a la Neurocomputación. Disponible en: http://www.ia.uned.es/asignaturas/sbc2/sbc2/libro/book.pdf

-Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for researchers. Tech Report HPL-2003-4, HP Laboratories. Disponible en: http://www.hpl.hp.com/personal/Tom Fawcett/papers/ROC101. pdf

-Bouckaert, R. (2004). Estimating Replicability of Classifier Learning Experiments, ICML, Disponible en: http://www.aicml.cs.ualberta.ca/_ban_04/icml/pages/papers/61.pdf

-Para el test pareado de Student se puede consultar el texto .Estadística. Modelos y Métodos"de Daniel Peña Sánchez de Rivera. o las entradas correspondientes de la enciclopedia matemática on-line Mathworld http://mathworld.wolfram.com/Pairedt-Test.html o de la wikipedia, http://en.wikipedia.org/wiki/Student's_t-test.

Actividad 2.3: Ejercicio de simulación

* Se recomienda que la realización de esta actividad se lleve a cabo tras haber estudiado los temas 3 y 4.

Local Linear Embedding y Diffusion Maps son técnicas de reducción de la dimensionalidad (compresión de datos) alternativas al Análisis de Componentes Principales. En esta práctica, vamos a aplicar dichas técnicas a datos de muy alta dimensionalidad para proyectar las instancias en espacios de pocas dimensiones. Vamos a evaluar el resultado de dicha reducción en un problema de regresión.

El estudiante deberá aplicar la técnica de Componentes Principales, Local Linear Embedding y Diffusion Maps al conjunto de datos denominado Kurucz. Dicho conjunto de datos contiene espectros sintéticos de estrellas obtenidos para diferentes temperaturas. El problema consiste en utilizar dichos datos para entrenar un modelo de regresión que prediga temperaturas a partir de espectros. El estudiante tendrá que comparar (mediante validación cruzada) la validez de dichos modelos entrenados a partir de la variables seleccionadas con las tres técnicas.

Los procesos de selección de variables deberán realizarlos mediante el programa R. Se puede encontrar una descripción de las técnicas Local Linear Embedding y Diffusion Maps en las lecciones 14 y 15 del curso de *Cosma Shalizi* del departamento de Estadística de la Universidad Carnegie Mellon. Dichas lecciones se encuentran en la carpeta de la actividad.

Para la aplicación del LLE, el estudiante encontrará el código R en la misma lección 14 antes citada. Para la aplicación de los Mapas de difusión, el estudiante puede emplear la función diffuse del paquete diffusionMap

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 a una cara con los siguientes apartados:

1. Descripción de las técnicas LLE y Diffusion Maps. La descripción (que no podrá ser copiada textualmente de ninguna fuente) deberá demostrar que el alumno ha entendido los fundamentos y principios de las metodologías.
2. Gráficas que representen la temperatura de un modelo en función de dos variables. Si, por ejemplo, tomamos el análisis de componentes principales, habría que representar la primera componente frente a la segunda componente y dibujar cada punto (estrella) utilizando un código de color para la temperatura. La escala de color debe formar parte de la gráfica. Sólo habrá que representar todos los pares posibles de las tres primeras variables nuevas.
3. Gráficas que representen la temperatura de un modelo en función de dos variables de metodologías distintas. Por ejemplo, la primera componente de LLE frente a la primera componente de los mapas de difusión. Habrá que dibujar cada punto (estrella) utilizando un código de color para la temperatura. La escala de color debe formar parte de la gráfica. Restrínjanse a las tres primeras variables nuevas.
4. Resumir y discutir los resultados de experimentos de validación cruzada de una regresión respecto a la temperatura. Es decir, el estudiante tendrá que entrenar un modelo de regresión mediante Máquinas de Vectores Soporte por cada conjunto de variables: componentes principales, LLE y Diffusion Maps. Para ello, deberá buscar el kernel (y el conjunto de parámetros que lo describen) óptimo. Tendrá que evaluar la funcionalidad en cada caso y comparar los distintos resultados. Para construir el modelo el estudiante podrá utilizar weka.

Tema 5

Actividad 5.1. [OPTATIVA] La metodología CRISP-DM

Visitar la página web relativa al proyecto CRISP-DM. Descargar y leer el documento relativo al modelo y guía de referencia de este estándar.

http://www.crisp-dm.org/index.htm

Entregables:

El alumno deberá realizar un conjunto de transparencias (tipo Powerpoint) en el que se resuma los fundamentos y las distintas fases de esta metodología.

Actividad 5.2. MD y escalabilidad: estudio de bibliografía avanzada

Realizar un análisis de cuáles de los algoritmos de minerías de datos estudiados a lo largo de este curso escalan bien a medida que se incrementa el volumen de datos.

Entregables:

El estudiante deberá entregar un trabajo de entre 3 y 6 páginas A4 en donde se realice un análisis y resumen del comportamiento de distintos algoritmos ante el problema de la escalabilidad (ver referencia [Han et al-96] como punto de partida,) y de las distintas estrategias utilizadas en el campo de la minería de datos para abordarlo de forma eficiente (utilizar la referencia [Provost&Kolluri-99] como punto de partida).

Actividad 5.3. [OPTATIVA] Minería de datos distribuida: Estudio de bibliografía avanzada

La mayoría de las técnicas de minería de datos vistas a lo largo de este curso aplican a ficheros de datos planos o bases de datos relacionales. Sin embargo, tal y como se ha estudiado en el presente tema, debido a la existencia de datos heterogéneos, de múltiples fuentes o almacenes de datos, y de la interconectividad con la web, ha cobrado recientemente importancia una nueva aproximación: la minería de datos distribuida.

Entregables:

Aunque este tipo de minería es un campo relativamente nuevo, se propone hacer una búsqueda bibliográfica sobre tipos de arquitecturas utilizadas para abordar la minería de datos distribuida y sobre las distintas técnicas que ésta utiliza. Realizar un informe sobre el estado actual del tema. Un punto de partida podría ser la referencia [Park&Kargupta-02]. También dispone en http://www.cs.umbc.edu/hillol/DDMBIB/ de un repositorio de bibliografía relacionada con este tema.

Actividad 5.4 [OPTATIVA]. Difusión y uso de la MD: Estudio de bibliografía avanzada

Un asunto importante a la hora de utilizar la información resultante de aplicar un programa de minería de datos es el de cómo integrar sus salidas en otro tipo de herramientas. Por ejemplo, cómo hacer un uso eficiente de los patrones o modelos aprendidos durante el proceso de minería en herramientas de toma de decisión. Según lo estudiado en este tema, existen distintas estrategias que abordan esta cuestión (reglas de actividad (triggers), integración de los modelos aprendidos en el sistema de gestión de base de datos, la utilización de estándares para el intercambio de modelos o el uso de protocolos

basados en XML).

Entregables:

Se propone al alumno analizar en más profundidad alguna de estas soluciones y crear un documento de 3 a 6 páginas A4 en el que se recoja sus características, su operativa, ámbito de aplicación y grado de aceptación.

Actividad 5.5. [OPTATIVA] Aplicaciones de la MD

La formación de un especialista en minería de datos debería no sólo atender a la evolución de sus distintos aspectos teóricos, sino complementarla continuamente con la consulta de ejemplos de aplicación. El conocimiento de lo ya solucionado puede ser de gran ayuda a la hora de abordar nuevos problemas en contextos similares. El alumno puede consultar alguno de los siguientes libros, donde se recopila información de soluciones de problemas abordados mediante minería de datos en diferentes campos.

-CRM y marketing [Berry&Linof-00]

-Telecomunicaciones [Mattison-97]

-Aplicaciones de ingeniería y científicas [Grossman et al-01]

-Medicina [Krzysztof-01, IBM-01]

-Finanzas, gubernamentales, seguros, etc. [Klösgen&Zytkow-02]

-Evidentemente, el abanico de referencias es muchísimo más extenso. Sólo en Internet se puede bucear en un amplio repertorio de trabajos publicados que están relacionados con aspectos prácticos y aplicados de la minería de datos.

Actividad 5.6. [OPTATIVA]La MD: cuestiones éticas y legales

En el contexto de las cuestiones éticas y legales surgidas por el potencial buen uso o mal uso de la minería de datos, se propone al alumno que dé un vistazo a las dos directivas encargadas de regular, tanto a nivel nacional como europeo, el tema de la protección de datos personales. Se recogen aquí dos enlaces desde donde puede consultarse:

-La Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal:

-http://civil.udg.es/normacivil/estatal/persona/PF/Lo15-99.htm

-La Directiva 95/46/EC, del Parlamento Europeo, de 23 de noviembre de 1995, conocida como European Data Protection Directive:

-http://www.cdt.org/privacy/eudirective/EU_Directive_.html

Bibliografía asociada:

[Berry&Linof-00] Berry, M., Linoff, G., Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley, 2000.

[Grossman et al-01] Grossman, R., Kamath, C., Kegelmeyer, W., Kumar, V., Namburu, R. (eds.). Data Mining for Scientific and Engineering Applications, Kluwer, September, 2001.

[Han et al-96] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu. DBMiner: A System for Mining Knowledge in Large Relational Databases, in E. Simoudis, J. Han, U. Fayyad, (eds.). Proc. Intl. Conf. on Data Mining and Knowledge Discovery, pp. 250-255, AAAI Press, 1996.

[IBM-01] IBM Redbooks Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data, IBM Corp, 2001.

[Klösgen&Zytkow-02] W. Kloesgen, JM Zytkow (Eds.), Handbook of data Mining and Knowledge Discovery. Oxford University Press, 2002.

[Krzysztof-01] Krzysztof J. (ed.), Medical Data Mining and Knowledge Discovery. Physica-Verlag, Springer, New York, 2001.

[Mattison-97] R. Mattison, Data Warehousing and Data Mining for Telecommunications. Artech House Computer Science Library, 1997

[Provost&Kolluri-99] F. Provost, V. Kolluri. A survey of methods for scaling up inductive algorithms. Data Mining and Knowledge Discovery, 3(2), pp. 131-169, 1999.

[Park& Kargupta-02] B. Park and H. Kargupta. Distributed Data Mining: Algorithms, Systems, and Applications. In Nong Ye, editor, Data Mining Handbook, pages 341-358. IEA, 2002

13.COLABORADORES DOCENTES

Véase equipo docente.