Asignatura: Sistemas Basados en Conocimiento y Minería de Datos (SBC)
Código: 32498
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

GUÍA DOCENTE: Sistemas Basados en Conocimiento y Minería de Datos (SBC)

Curso Académico: 2017-2018

Programa: Máster Universitario en Ingeniería Informática
Centro: Escuela Politécnica Superior
Universidad: Universidad Autónoma de Madrid

Última modificación: 2017/06/11
Estado:

1. ASIGNATURA (ID)

Sistemas Basados en Conocimiento y Minería de Datos (SBC)

1.1. Programa

Máster Universitario en Ingeniería Informática

1.2. Código asignatura

32498

1.3. Área de la asignatura

Ciencias de la Computación e Inteligencia Artificial

1.4. Tipo de asignatura

Obligatoria

1.5. Semestre

Primer semestre

1.6. Créditos

6 ECTS

1.7. Idioma de impartición

Las clases se impartirán principalmente en castellano. Algunas de las clases y seminarios pueden ser impartidas en inglés. La mayor parte del material del curso y transparencias será en inglés.

1.8. Recomendaciones / Cursos relacionados

Son recomendables conocimientos de álgebra lineal, probabilidad y estadística a un nivel introductorio, así como conocimientos básicos de programación.

Los cursos relacionados son:

- Computación a Gran Escala
- Procesamiento de información temporal
- Métodos bayesianos aplicados
- Recuperación de Información
- Minería Web

1.9. Datos del equipo docente

Nota: se debe añadir @uam.es a todas las direcciones de correo electrónico.

Dr. Manuel Sánchez-Montañés Isla (Coordinador)
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-303
Tel.: +34 914972290
e-mail: manuel.smontanes
Web: http://www.eps.uam.es/~msanchez
Horario de atención al alumno: Petición de cita previa por correo electrónico.

Dr. José R. Dorronsoro
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-358
Tel.: +34 914972329
e-mail: jose.dorronsoro
Web: http://www.eps.uam.es/~gaa
Horario de atención al alumno: Petición de cita previa por correo electrónico.

Dr. David Camacho
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-433
Tel.: +34 914972288
e-mail: david.camacho
Web: http://www.aida.ii.uam.es
Horario de atención al alumno: Petición de cita previa por correo electrónico.

Dr. Carlos Santa Cruz
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Depacho: B-343
Tel.: +34 914972337
e-mail: carlos.santacruz
Web: http://www.eps.uam.es/~santacru
Horario de atención al alumno: Petición de cita previa por correo electrónico.

1.10. Objetivos de la asignatura

El objetivo de esta asignatura es proporcionar formación al estudiante en paradigmas avanzados de aprendizaje automático y su utilización en aplicaciones prácticas. Se estudiarán las metodologías necesarias para llevar a cabo los diferentes pasos necesarios para resolver satisfactoriamente problemas reales: formalización del problema, pre-procesamiento de los datos, identificación de datos relevantes, construcción del clasificador automático, cuantificación de su rendimiento, y validación y estimación de su precisión en datos futuros.

Las competencias básicas que el estudiante adquiere en esta asignatura son:

- G4 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.
- G8 Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y multidisciplinares, siendo capaces de integrar estos conocimientos.

La competencia de tecnología específica que el estudiante adquiere en esta asignatura es:

- TI9 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.

Las cualificaciones ubicadas en el nivel de competencias transversales que el estudiante adquirirá en esta asignatura son:

- TR1 Capacidad para actualizar conocimientos habilidades y destrezas de forma autónoma, realizando un análisis crítico, análisis y síntesis de ideas nuevas y complejas abarcando niveles más integradores y pluridisciplinares.
- TR4 Capacidad para transmitir de un modo claro y sin ambigüedades a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación mas avanzada, así como los fundamentos mas relevantes sobre los que se sustentan. Capacidad para argumentar y justificar lógicamente dichas decisiones de un modo claro y sin ambigüedades, sin dejar de considerar puntos de vista alternativos o complementarios.

OBJETIVOS ESPECÍFICOS

PARTE I: Auditoria de datos y preprocesamiento	
TEMA 1.- Introducción al Aprendizaje Automático	
1.1.	Presentar una visión global del campo, mostrando áreas y aplicaciones relacionadas
1.2.	Identificar problemas de clasificación, reconocimiento de patrones, optimización, decisión y clustering
1.3.	Identificar el flujo básico de trabajo para obtener un modelo predictivo
1.4.	Validar y seleccionar los modelos más adecuados para un problema
TEMA 2.- Pre-procesamiento de los datos	
2.1.	Representar la base de datos de manera conveniente
2.2.	Identificar conceptos principales como: tipos de atributos, outliers, ruido, falsos predictores y missing values
2.3.	Aplicar el flujo básico para el preprocesamiento: limpieza de los datos, integración, transformación y reducción de datos
TEMA 3.- Reducción de la dimensión	
3.1.	Entender el concepto de la "maldición de la dimensión"
3.2.	Entender las diferencias entre extraer y seleccionar características. Entender las diferencias entre los métodos "filter" y los métodos "wrapper"
3.3.	Aplicar diferentes métodos de testeo
3.4.	Expresar las diferencias entre los dos métodos principales de extracción de características: no supervisados (guiados por la geometría) y supervisados (guiados por la tarea): PCA y LDA
3.5.	Entender los principios básicos de la reducción no lineal de la dimensión

PARTE II: Aprendizaje no supervisado	
TEMA 4.- Modelos no paramétricos	
4.1.	Expresar las diferencias entre modelos paramétricos y no paramétricos en aprendizaje no supervisado
4.2.	Aplicar el algoritmo k-means para realizar clustering en un conjunto de datos
4.3.	Aplicar el algoritmo difuso c-means para realizar clustering en un conjunto de datos
4.4.	Aplicar métodos de clustering jerárquicos usando diferentes medidas de distancia, y visualizar e interpretar los resultados usando dendrogramas
5.5	Comprender otras aproximaciones al clustering
TEMA 5.- Modelos paramétricos	
5.1.	Entender y describir el algoritmo EM
5.2.	Aplicar el algoritmo EM para ajustar una mezcla de Gaussianas a un dataset
TEMA 6.- Validación de clusters	
6.1.	Describir en términos sencillos el problema de la validación de clusters
6.2.	Usar diferentes índices de validación de clusters para validar y comparar las salidas de diferentes algoritmos de clustering
TEMA 7.- Estimación de la densidad	
7.1	Estimación paramétrica
7.2	Estimación no paramétrica

PARTE III: Aprendizaje Supervisado	
TEMA 8.- Regresión	
	Entender el concepto y suposiciones del problema de regresión. Identificar qué problemas prácticos pueden ser abordados con técnicas de regresión, y formalizarlos en consecuencia
	Describir las suposiciones, derivación, uso y limitaciones de la técnica de regresión lineal
	Describir las redes neuronales como una extensión no lineal natural de la regresión lineal e introducir las modernas redes neuronales profundas.
	Entender el concepto del sobreajuste y el dilema sesgo-varianza y sus consecuencias, y las técnicas estándar para evitarlo
	Usar Máquinas de Soporte Vectorial para solucionar problemas de regresión e interpretar los resultados (vectores de soporte, margen, etc.)
TEMA 9.- Clasificación	
	Entender los conceptos y suposiciones de MAP, ML y el test del cociente de verosimilitudes
	Describir las suposiciones, derivación, uso y limitaciones de la técnica de regresión logística
	Entender el funcionamiento de Näive Bayes
	Explicar el concepto y elementos de un árbol de decisión. Derivar una estrategia general para construir un árbol de decisión dependiendo del “criterio de división”
	Entender el fenómeno del sobreajuste en los árboles de decisión, y cómo evitarlo usando mecanismos de poda. Describir las diferencias entre pre- y post- poda
	Entender cómo las Máquinas de Soporte Vectorial se generalizan para solucionar problemas de clasificación
	Entender cómo las redes neuronales y redes neuronales profundas se generalizan para resolver problemas de clasificación
	Definir el concepto de aprendizaje de conjuntos de clasificadores, y entender las suposiciones bajo las cuales dicho aprendizaje tiene sentido. Aplicar bagging, boosting y random forest para construir conjuntos de clasificadores

Contenidos del programa

PARTE I: Auditoría de datos y preprocesamiento

1. Introducción al Aprendizaje Automático
    - Conceptos básicos (definiciones, tipos de aproximaciones)
    - Aplicaciones del Aprendizaje Automático
    - Componentes y etapas en un proyecto de Aprendizaje Automático

2. Pre-procesamiento de datos
    - Normalización y estandarización
    - Discretización
    - Procesamiento de variables especiales: fechas, series, atributos categóricos y nominales
    - Missing values y outliers

3. Reducción de la dimensión
    - La “maldición de la dimensión”
    - Selección de variables. Métodos “filter” y “wrapper”
    - Feature engineering

PARTE II: Aprendizaje no Supervisado

4. Modelos no paramétricos
    - Clasificación no supervisada (clustering)
    - Algoritmo k-means
    - Algoritmo difuso c-means
    - Clustering jerárquico

5. Modelos paramétricos
    - Clustering basado en modelos y estimación de densidad
    - Máxima Verosimilitud. El algoritmo EM
    - Mezclas de Gaussianas

6. Validación de clusters
    - Índices geométricos
    - Índices basados en información y verosimilitud

7. Estimación de la densidad
    - Paramétrica (ML, EM)
    - No paramétrica (histogramas, Ventanas de Parzen, kernels)

PARTE III: Aprendizaje Supervisado

8. Regresión
    - El modelo de regresión
    - Regresión lineal (LMS)
    - Redes neuronales para regresión y deep learning
    - Sesgo-varianza
    - Máquinas de vectores de soporte para regresión

9. Clasificación
    - Evaluación, regularización y selección del modelo. Score y matriz de confusión. ROCs
    - Árboles de clasificación
    - Criterios MAP y ML. Test del cociente de verosimilitudes. Naïve Bayes
    - Regresión logística. Redes neuronales para clasificación.
    - Máquinas de vectores de soporte para clasificación
    - Conjuntos de clasificadores: bagging, boosting y random forest

1.11. Bibliografía

1. T. Hastie, R. Tibshirani, J.H. Friedman. The Elements of Statistical Learning. Springer 2009
2. C.M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006
3. R.O. Duda, P.E. Hart. D.G. Stork; Pattern Classification; Wiley, 2000
4. T.M. Mitchell. Machine Learning. McGraw-Hill, 1997
5. Sunila Gollapud. Practical Machine Learning. Packt Publishing, 2016
6. Sebastian Raschka . Python Machine Learning. Packt Publishing , 2015

1.12. Trabajo del estudiante y evaluación

La evaluación del curso se efectuará en cuatro partes sobre cada una de las componentes de mismo y consistirá en

- Un trabajo práctico sobre el material de la componente.
- Las respuestas a un examen en formato take-home sobre el material expuesto.

Dichas evaluaciones tendrán lugar de acuerdo al siguiente calendario aproximado:

- Evaluación 1 (Preprocesado): semana del 25 de septiembre de 2017
- Evaluación 2 (No supervisado): semana del 23 de octubre de 2017
- Evaluación 3 (Regresión): semana del 20 de noviembre de 2017
- Evaluación 4: primera semana lectiva de enero de 2018

En caso de no superar alguna de estas evaluaciones, el estudiante deberá volver a efectuarla a primeros de enero.

La nota final se calculará como la media de las notas de cada componente. Cada una de estas notas se obtendrá como:

- Un 60% de la nota del trabajo práctico
- Un 40 % de la nota del cuestionario

Para cada una de las partes, es necesario alcanzar una nota mínima de 5 tanto en el trabajo práctico como en el examen.

El estudiante tiene la oportunidad, aunque haya aprobado las prácticas en la convocatoria ordinaria, de volver a entregarlas mejoradas en la convocatoria extraordinaria.