



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
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: 2016-2017

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

Última modificación: 2016/05/10
Estado:



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

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.



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

Los cursos relacionados son:

- 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>

Dr. Luis F. Lago Fernández

Departamento de Ingeniería Informática

Escuela Politécnica Superior

Depacho: B-307

Tel.: +34 914972211

e-mail: luis.lago

Web: <http://www.eps.uam.es/~lflago>

Dra. Ana González Marcos

Departamento de Ingeniería Informática

Escuela Politécnica Superior

Despacho: B-332

Tel.: +34 914972234

e-mail: ana.marcos

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.



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

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 multidisciplinarios, 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 pluridisciplinarios.
- 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: Introducción	
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 Supervisado	
TEMA 4.- Clasificación y regresión	
4.1.	Entender el concepto y suposiciones del aprendizaje supervisado.
4.2.	Identificar qué problemas prácticos pueden ser abordados con técnicas de aprendizaje supervisado, y formalizarlos en consecuencia.
4.3.	Describir las suposiciones, derivación, uso y limitaciones de las técnicas de regresión lineal y logística.
4.4.	Explicar el concepto del sobreajuste y las técnicas estándar para evitarlo.
4.5.	Describir problemas de clasificación y de regresión como casos particulares en el marco GLM.
TEMA 5.- Teoría Bayesiana de la Decisión	
5.1.	Entender los conceptos y suposiciones de la Teoría Bayesiana de la Decisión. Probabilidades a priori y a posteriori, costes.
5.2.	Aplicarlos a obtener funciones discriminantes y el test del cociente de verosimilitudes. Relacionarlas con los criterios MAP y ML.
5.3.	Aplicar esta teoría para derivar diferentes técnicas de aprendizaje supervisado: algoritmos de aprendizaje generativos, análisis discriminante Gaussiano, Naïve Bayes.
TEMA 6.- Estimación de la densidad de probabilidad	
6.1.	Entender el concepto de estimar la densidad de probabilidad y su relación con el aprendizaje supervisado.
6.2.	Entender el concepto de estimación paramétrica de la densidad usando ML, y derivar la solución para el caso Gaussiano.
6.3.	Usar diferentes técnicas de estimación no paramétrica de la densidad: histogramas, ventanas de Parzen, kernels.
TEMA 7.- Árboles de decisión	
7.1.	Explicar el concepto y elementos de un árbol de decisión.
7.2.	Derivar una estrategia general para construir un árbol de decisión dependiendo del “criterio de división”.
7.3.	Entender y formalizar los requisitos que un criterio de división apropiado debe de satisfacer, y revisar las diferencias prácticas entre diferentes criterios.
7.4.	Entender el fenómeno del sobreajuste en los árboles de decisión, y cómo evitarlo usando mecanismos de poda.
7.5.	Describir las diferencias entre pre- y post- poda.



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

TEMA 8.- Conjuntos de clasificadores	
8.1.	Definir el concepto de aprendizaje de conjuntos de clasificadores, y entender las suposiciones bajo las cuales dicho aprendizaje tiene sentido.
8.2.	Entender el dilema sesgo-varianza y sus consecuencias.
8.3.	Aplicar bagging y boosting para construir conjuntos de clasificadores.
TEMA 9.- Máquinas de Soporte Vectorial	
9.1.	Entender el concepto de dimensión VC y calcularla para ejemplos sencillos.
9.2.	Explicar de una manera intuitiva por qué los separadores lineales que maximizan el margen son óptimos, y adquirir los conocimientos básicos para seguir profundizando en el area de Máquinas de Soporte Vectorial.
9.3.	Usar Máquinas de Soporte Vectorial para solucionar problemas de clasificación, e interpretar los resultados (vectores de soporte, margen, frontera de decisión, etc.).

PARTE III: Aprendizaje no supervisado	
TEMA 10.- Modelos paramétricos	
10.1.	Entender y describir el algoritmo EM.
10.2.	Aplicar el algoritmo EM para ajustar una mezcla de Gaussianas a un conjunto de datos.
10.3.	Validar los diferentes ajustes usando AIC y BIC.
TEMA 11.- Modelos no paramétricos	
11.1.	Expresar las diferencias entre modelos paramétricos y no paramétricos en aprendizaje no supervisado.
11.2.	Aplicar el algoritmo k-means para realizar clustering en un conjunto de datos.
11.3.	Aplicar el algoritmo difuso c-means para realizar clustering en un conjunto de datos.
11.4.	Aplicar métodos de clustering jerárquicos usando diferentes medidas de distancia, y visualizar e interpretar los resultados usando dendogramas.
11.5.	Comprender otras aproximaciones al clustering.
TEMA 12.- Validación de clusters	
12.1.	Describir en términos sencillos el problema de la validación de clusters.
12.2.	Usar diferentes índices de validación de clusters para validar y comparar las salidas de diferentes algoritmos de clustering.



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

1.11. Contenidos del programa

PARTE I: Introducción

1. Introducción al Aprendizaje Automático
 - Conceptos básicos (definiciones, tipos de aproximaciones)
 - Aplicaciones del Aprendizaje Automático
 - Componentes y etapas en un proyecto de Aprendizaje Automático
 - Evaluación, regularización y selección del modelo (medidas de error, matriz de confusión, análisis ROC, PR, validación)
 - Introducción al software de Aprendizaje Automático Weka
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”)
 - Construcción de variables (LDA, PCA, métodos espectrales)

PART II: Aprendizaje Supervisado

4. Clasificación y regresión
 - Regresión lineal (LMS)
 - Regresión logística y clasificación
 - Sobreajuste
 - Modelos Lineales Generalizados
5. Teoría Bayesiana de la Decisión
 - Test del cociente de verosimilitudes
 - Riesgo de Bayes
 - Criterios MAP y ML
 - Algoritmos de aprendizaje generativos, análisis discriminante Gaussiano
 - Naïve Bayes
6. Estimación de la densidad
 - Paramétrica (ML, EM)
 - No paramétrica (histogramas, Ventanas de Parzen, kernels)

7. Árboles de decisión

- Definiciones
- Construcción de árboles de decisión. Criterios de división (error esperado, índice de Gini, ganancia de información)
- Poda

8. Conjuntos de clasificadores

- Suposiciones de precisión y diversidad. Aprendizaje de conjuntos de clasificadores
- Problemas estadísticos, numéricos y de representación de clasificadores individuales
- Bagging y boosting
- El dilemma sesgo-varianza

9. Máquinas de Soporte Vectorial

- Dimensión de Vapnik-Chervonenkis
- Minimización del riesgo estructural
- Clasificador de margen máximo, hiperplano separador óptimo
- Casos separables y no separables
- Problemas no lineales. Teorema de Cover. Kernels
- El algoritmo SMO

PARTE III: Aprendizaje no Supervisado

10. Modelos paramétricos

- Clustering basado en modelos y estimación de densidad
- Máxima Verosimilitud. El algoritmo EM
- Mezclas de Gaussianas
- Criterios de validación. AIC y BIC, validación cruzada

11. Modelos no paramétricos

- Clasificación no supervisada (clustering)
- Algoritmo k-means
- Algoritmo difuso c-means
- Clustering jerárquico

12. Validación de clusters

- Índices geométricos
- Índices basados en información y verosimilitud
- Índices basados en neg-entropía



Asignatura: Machine Learning: theory and applications (ML)
Código: 32419
Institución: Escuela Politécnica Superior
Programa: Máster Universitario en Ingeniería Informática
Nivel: Máster
Tipo: Obligatoria
ECTS: 6

1.12. 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

1.13. Trabajo del estudiante y evaluación

El curso involucra clases teóricas, prácticas de laboratorio, y un examen final.

- En el periodo ordinario de evaluación, esta será realizada de acuerdo al siguiente esquema:
 - 50 % Prácticas de laboratorio
 - 50 % Examen final

Es necesario alcanzar una nota mínima de 5 tanto en el examen final como en la nota media de prácticas para aprobar el curso.

En caso de que el estudiante no apruebe la asignatura en la convocatoria ordinaria, las notas de las partes individuales se conservan para la convocatoria extraordinaria.

- En el periodo extraordinario de evaluación, esta será realizada de acuerdo al siguiente esquema:
 - 50 % Prácticas de laboratorio
 - 50 % Examen final

Es necesario alcanzar una nota mínima de 5 tanto en el examen final como en la nota media de prácticas para aprobar el curso.

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