

PROGRAMA DE CURSO

Código	Nombre			
MA5406	Probabilidad y Estadística en el Análisis de Datos			
Nombre en Inglés				
Probability and Statistics in the Analysis of Data				
SCT	Unidades Docentes	Horas de Cátedra	Horas Docencia Auxiliar	Horas de Trabajo Personal
	10.0	3.0	1.5	5.5
Requisitos			Carácter del Curso	
MA3403 Probabilidad y Estadística O MA3401 Probabilidades			Electivo de Carrera, Magister y Doctorado	
Resultados de Aprendizaje				
Se quiere que el estudiante conozca las técnicas principales probabilísticas de clusterización y clasificación de datos de distintos tipos y que sea capaz de aplicarlos en un problema real con gran número de objetos y varios variables.				

Metodología Docente	Evaluación General
Clases expositivas Aprendizaje basado en problema Laboratorio Examen final	Control Trabajo personal Tareas, etc.

Unidades Temáticas

Número	Nombre de la Unidad	Duración en Semanas
1	Tipos de datos	1
Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
Categoricos/ordinales/intervalos/razones. Vectores de características	Que conoce los distintos tipos de datos, cuando y como usarlos para representar características en el mundo real.	Capítulo 2 de IV.

Número	Nombre de la Unidad	Duración en Semanas
2	Clusterización	3
Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
(a) k-medias - Algoritmo estandar: algoritmo de Lloyd - Inicialización: Forgy, muestras aleatorias, k-medias++ - Relación a otras métodos estadísticos (ACP, etc.) (b) Clusterización jerárquica - Distancias: Euclideana, Manhattan, máxima, Mahalanobis, Hamming, etc. - linkage: single (mínimo), promedio (media), completo (máximo), centroide, etc. - Métodos: aglomerativos, divisivos	Conocimiento de métricas tipo Euclideana, conocimiento de algoritmos y su rendimientos, capaz de aplicar métodos de clusterización a conjuntos reales de datos	Capítulo 9 de III.

Número	Nombre de la Unidad	Duración en Semanas
3	Clasificación de datos	5
Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía

<p>(a) Objetivos de clasificación. Errores de clasificación y medidas de precisión: Sensibilidad, especificidad, precisión, tasa de mal clasificación.</p> <p>(b) Clasificadores basados en distancias (k-vecinos más próximos)</p> <p>(c) Clasificación por Máxima verosimilitud y naïve Bayes.</p> <p>(d) Análisis discriminante: lineal/cuadrática/Fisher</p> <p>(e) Validación cruzada. Leave-one-out validación, k veces validación cruzada, conjunto de entrenamiento, conjunto de validación.</p>	<p>Entender clasificación como manera de separar el espacio de características en conjuntos disjuntos por hiperplanos lineales o no lineales; saber como entrenar en validar un clasificador usando conjuntos de entrenamiento y validación y luego como aplicar el clasificador resultante a datos nuevos.</p>	<p>Capítulos 3, 4 de II; Capítulos 2, 3 de III, IV; Capítulo 5 de IV.</p>
--	---	---

Número	Nombre de la Unidad	Duración en Semanas
4	Árboles de Clasificación y Regresión	6
Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
<p>(a) Árboles, Particiones, cadenas de particiones. Probabilidades en conjuntos finitos (hojas de un árbol).</p> <p>(b) Funciones de Impureza: entropía de Shannon, Impureza de Gini. Propiedades: Divergencia de Kullback-Leibler e Información Mutua.</p> <p>(c) Clasificación por árboles.</p> <p>(d) Poda de árboles. Costos, función de pérdida,</p> <p>(e) Árboles de Regresión.</p>	<p>Entender la estructura de árboles en general, creación de preguntas por división de nodos, el mecanismo de reducción de impureza, el por qué se debe podar, las ventajas y debilidades de árboles de decisión. Ser capaz de aplicar árboles de decisión a conjuntos de datos complejo (por ejemplo, en bioinformática).</p>	<p>Capítulos 2, 3, 4, 8, 10 de I; Capítulo 2 de II; Capítulo 7 de III; Capítulo 12 de IV.</p>

(f) Bosques aleatorios/boosting		
---------------------------------	--	--

Bibliografía General	
I. Breiman, Friedamn, Olshen y Stone. Classification and regression trees. Chapman & Hall, 1984.	
II. Murphy, K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.	
III. Ripley, B.D. Pattern Recognition and Neural Networks. Cambridge University Press, 1996.	
IV. Witten, I.H., Frank, E. and Hall, M.A. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., 2005.	

Vigencia desde:	Otoño 2017
Elaborado por:	Andrew Hart y Servet Martínez
Revisado por:	