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 Obligatorio de Magister

Resultados de Aprendizaje
Se quiere que el estudiante conozca las técnicas principales probabilísticas de clusterización y clasificación de datos de distintos tipos, en particular en los datos de texto libre, que pueda adaptarse a las diferentes técnicas y escenarios que aparezcan en probelmas de datos, y sea capaz de aplicarlos en un problema real con gran número de objetos y varias variables de tipos variados y complejos.

Metodología Docente	Evaluación General
Clases expositivas Aprendizaje basado en problema Laboratorio	Exposiciones Trabajo personal Tareas. Trabajo con base de datos con complejidad

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
Categóricos/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	Datos de texto libre	2

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografí
Datos de Texto libre	Disciplinas en que aparece texto libre, fificulatdaes en su tratamiento. Preprocesamiento Básico de Texto. Estructura probabilística en el lenguaje. Aplicaciones en correctores de lenguaje. Representación vectorial de texto. Reducción de dimensionalidad. Word to vec.	Capítulo 1 de V. Capítulo 11 de VI. VII.

Número	Nombre de la Unidad	Duración en Semanas
3	Clusterización	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía

(a) Analisis en componentes principales (ACP) (b) k-medias - Algoritmo estandar: algoritmo de Lloyd - Inicialización: Forgy, muestras aleatorias, k-medias++ - Relación a otras métodos estadísticos (ACP, etc.) (c) 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
4	Clasificación de datos	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
(a) Objetivos de clasificación. Errores de clasificación y medidas de precisión: Sensibilidad, especificidad, precisión, tasa de mal clasificación. (b)Clasificadores basados en distancias (k-vecinos más próximos) (c) Clasificación por Máxima verosimilitud y naïve Bayes. (d) Análisis discriminante: lineal/cuadrática/Fisher (e) Validación cruzada. Leave-one-out validación, k veces validación cruzada, conjunto de entrenamiento, conjunto de validación.	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.	Capítulos 3, 4 de II; Capítulos 2, 3 de III, IV; Capítulo 5 de IV.

Número	Nombre de la Unidad	Duración en Semanas
5	Árboles de Clasificación y Regresión	4

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
(a) Arboles, Particiones, cadenas de particiones. Probabilidades en conjuntos finitos (hojas de un árbol). (b) Funciones de Impureza: entropía de Shannon, Impureza de Gini. Propiedades: Divergencia de Kullback-Leibler e Información Mutua. (c) Clasificación por árboles. (d) Poda de árboles. Costos, función de pérdida, (e) Arboles de Regresión. (f) Bosques aleatorios/boosting	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).	Capítulos 2, 3, 4, 8, 10 de I; Capítulo 2 de II; Capítulo 7 de III; Capítulo 12 de IV.

Número	Nombre de la Unidad	Duración en Semanas
6	Clasificación de Texto	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
Aplicaciones en reconocimiento de texto	Reconocimiento de SPAM Análisis de opinión	Capítulo 13 VI.

Número	Nombre de la Unidad	Duración en Semanas
7	Hacia donde va la Ciencia de los Datos	1

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
Analisis del estado actual y desafíos en ciencia de los datos	50 años de ciencia de los datos Procesamiento de lenguaje de texto	Capítulo 13 VI. VIII y IX

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.

V. C. Manning, H. Schutze, Foundations of Statistical Natural Processing. MIT Press, 1999.

VI. C. Manning, P. Raghavan, H. Schutze. An Introduction to Information Retrieval. Cambridge University Press, 2019.

VII. C. Manning. Natural Language Processing. Stanford Coursera course.

VIII. L. Breiman. Statistical Modeling: Thetwo cultures. Statistical Science Vol. 16, No. 3, 199-231, 2001.

IX. D. Donoho. 50 years od data science. J. of Computational and Graphical Statistics. Vol 26, No. 4, 745-766, 2017.

Vigencia desde:	Otoño 2019
Elaborado por:	Jocelyn Dunstan, Andrew Hart y Servet Martínez
Revisado por:	