Universidad de Costa Rica
Sistema de Estudios de Posgrado
Programa de Posgrado en Estadística

SP1668 Técnicas computacionales y estadísticas de aprendizaje de máquina
PF3115 Técnicas computacionales y estadísticas de aprendizaje de máquina

INFORMACIÓN GENERAL

Modalidad:	Teórico y Práctico
Número de créditos:	4 créditos
Horas presenciales:	4 horas semanales
Horario:	*Martes, 5-9pm: 2 horas clase, 2 horas laboratorio*
Aula:	*Lab 102, Edificio ECCI*
Horas de atención de estudiantes:	2 horas semanales, con cita previa
Requisitos:	R, Python y manejo de inglés oral, escrito y lectura.
Ciclo lectivo:	*I-2020*
Profesor(a):	*Dra. Marcela Alfaro Córdoba* *Dr. Markus Eger*

PROGRAMA

Justificación

El aprendizaje de máquina se ha popularizado en los últimos años como una herramienta de predicción por excelencia. Las aplicaciones son numerosas y variadas, van desde las ciencias de la salud, pasando por modelos predictivos para el estado del tiempo y terminando por modelos que aprenden de nuestros patrones de comportamiento en línea para recomendar productos en mercadeo.

En algunas áreas de aplicación se ha utilizado este conjunto de técnicas como “caja negra”, lo que ha llevado a cometer errores de interpretación, como tratar de justificar comportamientos del modelo que no se pueden explicar. Por ello, es de suma importancia estudiar los fundamentos teóricos de los principales modelos de aprendizaje de máquina, sus similitudes y diferencias con los modelos de aprendizaje estadístico, y encontrar un punto medio en el que se comprenda primero, la diferencia entre algoritmos determinísticos y modelos para procesos aleatorios, y segundo, la teoría subyacente en el caso de los modelos estadísticos y de predicción.

Además del componente técnico de los modelos, este curso ofrece una ventana para discutir aspectos éticos de la construcción y aplicación de modelos de aprendizaje, que incluye el proceso de recolección de datos, su pre-procesamiento, el flujo de trabajo y la correcta preservación y manejo de los datos utilizados y el código, para asegurar reproducibilidad.

El área de estudio puede dividirse en tres grandes áreas, de acuerdo al tipo de modelos: aprendizaje de máquina clásico, aprendizaje estadístico clásico, y desarrollos recientes. El tipo de problema que se puede solucionar, el análisis y la interpretación correspondiente difieren entre estas tres áreas.

El curso está diseñado como curso teórico-práctico de 4 créditos y opcional compartido dentro de la maestría profesional y académica en Estadística y las maestrías y doctorado de Computación e Informática.

Objetivo general

El objetivo general del curso es que el estudiante se familiarice con los modelos de aprendizaje (de máquina y estadísticos), los problemas que se pueden estudiar dentro de esta área, y los métodos existentes. Además, discutir las implicaciones éticas de su implementación de una manera transversal.

Objetivos específicos

Al final de curso, el estudiante será capaz de:

1. Comprender lo que se entiende como modelos de aprendizaje de máquina.
2. Comprender lo que se entiende como modelos de aprendizaje estadístico.
3. Identificar el tipo de modelos de aprendizaje que se pueden aplicar dependiendo del tipo de problema.
4. Aplicar los modelos de aprendizaje de máquina a datos reales.
5. Aplicar los modelos de aprendizaje estadístico a datos reales.
6. Comprender las implicaciones éticas de la implementación de los modelos de aprendizaje.
7. Comprender la diferencia entre los modelos clásicos de aprendizaje y los de desarrollo reciente.

Descripción del curso

1. ***Introducción:*** Herramientas básicas de programación en Python. ¿Qué son modelos de aprendizaje de máquina y estadísticos?, ¿cuál es la diferencia entre los dos y cómo se relacionan con la estadística?. El flujo de trabajo con datos.
2. ***Aprendizaje de Máquina Clásico:*** Supervisado, no supervisado *y* de refuerzo. Perceptrón, Redes Neuronales, Support Vector Machine, *k-means*.
3. ***Aprendizaje Estadístico Clásico:*** Regresión Lineal y la reinterpretación de modelos supervisados y no supervisados desde la perspectiva estadística.
4. ***Desarrollos recientes*****:** Deep Learning, GANs, LSTM, Vector Models.

Metodología

Dentro de cada tema, el material se impartirá con ayuda de presentaciones, lectura de textos extraídos de libros y ejemplos prácticos. Además, se distribuirán artículos científicos relacionados con el tema para lectura y análisis; los artículos se escogerán de diferentes áreas de aplicación. Los estudiantes además realizarán proyectos dentro de cada tema y llevarán a cabo un proyecto final más comprensivo. Este curso requiere una gran cantidad de trabajo práctico por parte del estudiante.

El curso se administrará mediante la plataforma Piazza, Github y la aplicación de mensajería Slack. Los y las estudiantes se inscribirán en el curso e interactuarán con los profesores mediante estas plataformas para desarrollar sus trabajos, allí ubicarán sus trabajos y tendrán disponible el material del curso. Resulta imprescindible disponer de un acceso internet al que puedan acceder los estudiantes, hacer las consultas respectivas, comunicarse con los

profesores, recibir y enviar los materiales del curso. Se espera que cada estudiante tenga acceso a una computadora personal para desarrollar las prácticas en clase y en su hogar.

Las áreas de aplicación se adaptarán a los intereses de los estudiantes, que además tendrán la oportunidad de trabajar con datos propios. Las cuatro horas de lecciones semanales se dividirán en 2: 2 horas de clase y 2 horas de laboratorio, más tareas y un proyecto con entregas parciales durante el semestre.

Cronograma

I(4), II(4), III(4), IV(3). El número de semanas de cada tema (entre paréntesis) es un valor estimado.

Bibliografía

Libro de texto:

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.
Hastie, T., Tibshirani, R., Friedman, J. (2013) The Elements of Statistical Learning. Second Edition, Springer.

Otras referencias:

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) An Introduction to Statistical Learning with Applications in R. Springer.
Andriy Burkov. (2019) The hundred-page machine learning book.
Yaser S. Abu-Mostafa, Malik Magdon-Ismalil, Hsuan-Tien Lin. (2012) Learning from data, AMLbook.com.
Kevin P. Murphy. (2012) Machine learning - a probabilistic perspective. MIT Press.
David Barber. (2017) Bayesian Reasoning and Machine Learning. Cambridge University Press.
Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong (2019). Mathematics for Machine Learning (draft freely available), to be published by Cambridge University Press.
Bradley Efron and Trevor Hastie. (2016) Computer Age Statistical Inference: Algorithms, Evidence and Data Science. Cambridge University Press.
Yann LeCun, Yoshua Bengio & Geoffrey Hinton. (2015) Deep learning. Nature 521.
Ian Goodfellow, Yoshua Bengio & Aaron Courville. (2016) Deep learning. MIT Press.
Daniel Geng & Rishi Veerapaneni. (2018) Tricking Neural Networks: Create your own Adversarial Examples, blog post.

Evaluación

La evaluación será mediante tareas y un proyecto. El proyecto seriá definidos por lo profesores en correspondencia a las propuestas de los y las estudiantes. Para este propósito la comunicación en clase y por medios electrónicos son muy relevantes para concretar el proyecto, su alcance y desarrollo. El curso se plantea evaluar mediante la aplicación práctica de los conocimientos adquiridos a situaciones particulares.

Rubro	Porcentaje
Proyecto	35 %
Propuesta	5%
1er Reporte parcial	5%
2do Reporte parcial	5%
Presentacion final	20%
Labs	65 %
Lab 1	15%
Lab 2	10%
Lab 3	10%
Lab 4	10%
Lab 5	10%
Lab 6	10%
TOTAL	100%