

IX ESCUELA DE VERANO - 2020

1. Nombre del curso:

MACHINE LEARNING. APLICACIONES EN DATA MINING Y DEEP LEARNING.

Facultad de Informática.

2. Docentes a cargo:

- **Dr. Franco Ronchetti.** Investigador asistente de la CIC-PBA. Profesor de postgrado en la Facultad de Informática, UNLP.

- **Dr. Alejandro Rosete Suárez.** Profesor de la Universidad Tecnológica de la Habana CUJAE.

3. Fundamentación:

El Aprendizaje Automático (Machine Learning) es el área de la Inteligencia Artificial que estudia cómo encontrar modelos estadísticos complejos para resolver problemas de difícil solución analítica. En otras palabras, es la ciencia que hace que las computadoras “aprendan” sin ser explícitamente programadas. En la actualidad, prácticamente todo producto de software posee un componente “inteligente”, lo que hace a esta disciplina una herramienta sumamente importante.

La Visión por Computadora es una subárea del Aprendizaje Automático aplicado a imágenes digitales. La aparición de modelos profundos (Deep Learning) en los últimos años, así como hardware específico, permitió una mejora considerable en el área. Las redes neuronales profundas son una extensión de las redes neuronales artificiales tradicionales. En la actualidad estos modelos son el componente principal de cualquier sistema inteligente, ya sea para reconocimiento de imágenes, texto, audio o video.

4. Objetivos:

En este curso se desarrollarán los fundamentos teóricos del aprendizaje automático y aprendizaje profundo. Por medio del curso el alumno podrá:

- Adquirir los conceptos básicos del aprendizaje automático, sus fundamentos teóricos y una aproximación a sus bases formales.
- Aprender diferentes técnicas de preprocesamiento de datos necesarias para llevar a cabo un proceso de aprendizaje automático.
- Conocer los principales modelos y algoritmos de aprendizaje computacional, de minería de datos para uso general, así como los diseñados específicamente para problemas de Visión por Computadora.
- Conocer metodologías que permitan seleccionar el modelo apropiado a los casos prácticos que se le presenten. Aplicar diferentes métricas que permitan un análisis de los algoritmos desarrollados.
- Internalizar los conceptos por medio de la práctica con herramientas de libre disponibilidad establecidas en el mundo académico como *knime* y entornos Python.

5. Perfil del estudiante:

Se espera que los alumnos posean conocimientos básicos de programación y de álgebra lineal.

6. Contenidos:

- Introducción. Los orígenes del aprendizaje automático. Usos del aprendizaje automático. Aciertos y límites. Aspectos éticos. Almacenamiento y estructura de datos. Conjuntos de entrenamiento y evaluación. Generalización. Sobreajuste y subajuste. Tipos de algoritmos de aprendizaje.

- Introducción a la Minería de Datos. Alcances, tareas y métodos. Herramientas y metodologías. Relación con las fuentes de información y los almacenes de datos. Inteligencia de negocios pasiva y activa.
- Vista minable. Relación con las metas de Minería. Visualización y preprocesamiento. Filtrados. Carga y modificación de datos. Dependencia con los métodos a aplicar para descubrimiento. Normalización de los datos. Atributos nominales y numéricos.
- Aprendizaje no supervisado. Datos no etiquetados. Clustering. Medidas de distancia y de conectividad. Proceso de agrupamiento. Clustering partitivo. Algoritmos K-medias y Modelos de Mixturas Gaussianas.
- Inducción de árboles de clasificación. Algoritmo ID3 y sus extensiones: C4.5, J48. Parámetros que controlan los árboles a obtener. Relación entre árboles y reglas. Relación entre categorización y resultados. Obtención de Reglas de Asociación. Partición vs cobertura. Parámetros que controlan las reglas a obtener. *Knime* como herramienta para la minería de datos. *Knime* para el preprocesamiento y la obtención de reglas y árboles.
- Aprendizaje supervisado. Regresión Lineal. Función de error. Descenso del gradiente. Regresión logística. Clasificación binaria vs. Clasificación multiclase.
- Performance del modelo. Conjuntos de entrenamiento y evaluación. Matriz de confusión. Métricas *Accuracy*, *Precision*, *recall* y *F-measure*. Visualización de curvas de entrenamiento.
- Redes Neuronales. Descripción de la arquitectura feedforward. Algoritmo de entrenamiento backpropagation. Funciones de activación. Resolución de problemas de clasificación y predicción. Capacidad de generalización de la red.

- Visión por Computadora. Introducción al aprendizaje Profundo (Deep Learning). Lenguajes tensoriales. Tipos de capas. Procesamiento digital de imágenes. Espacios de color. Operaciones sobre imágenes. Capas convolucionales y de pooling. Visualización de la red. Aplicaciones en Python.

7. Metodología:

En cada jornada de clase se dividirá el tiempo en dos partes. En la primera mitad se explicarán los conceptos teóricos correspondientes al tema de estudio. En la segunda mitad se harán actividades prácticas con modalidad taller en sala de PC, con supervisión del docente a cargo. El objetivo de las actividades será la utilización de las técnicas presentadas, junto con diferentes conjuntos de datos de repositorios conocidos, con el fin de resolver problemas de regresión, clasificación binaria, clasificación multiclase y minería de datos en general. Estas actividades se harán utilizando el lenguaje la herramienta Knime y el lenguaje Python, junto con librerías actuales que ayuden a la comprensión de los temas abordados. Se guiará a los alumnos en la resolución de los problemas, ayudándolos con la interpretación de los datos y las técnicas utilizadas.

Una vez finalizadas las actividades se discutirán los resultados de manera grupal con el fin de establecer criterios en común para la resolución de las actividades presentadas en el día.

En el último día se reservará la mitad del tiempo para la evaluación final escrita.

8. Forma de evaluación y fecha límite de presentación:

La evaluación será escrita y se tomará el último día de clase, con una duración máxima de 2hs. En el caso de los alumnos que desapruében dicha evaluación, se pedirá un trabajo práctico a desarrollar que involucre diversos temas abordados en el curso y que permitan la evaluación de los

conocimientos adquiridos. El plazo límite para este trabajo será de 2 meses luego de la evaluación escrita.

9. Bibliografía:

Bibliografía obligatoria

1. Stephen Marsland. Machine Learning: An Algorithmic Perspective – 2nd Ed. CRC Press. 2015.
2. Michael A. Nielsen. “Neural Networks and Deep Learning”, Determination Press, 2015.
3. François Chollet. “Deep Learning with Python”, Manning, 2017.
4. Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques, (Fourth Edition). Morgan Kaufmann. 2017.

Bibliografía complementaria

1. Jason Bell. Machine Learning: Hands-On for Developers and Technical Professionals. Wiley. 2015.
2. Sarah Guido, Andreas Müller. Introduction to Machine Learning with Python. O'Reilly Media. 2016.
3. Christopher Bishop. Pattern Recognition and Machine Learning. Springer-Verlag. 2006.
4. Brett Lantz. Machine Learning with R. Packt Publishing. 2015.
5. David Barber. Bayesian Reasoning and Machine Learning. Cambridge University Press. 2015.
6. Hernández Orallo, Ramírez Quintana, Ferri Ramírez. Introducción a la Minería de Datos. Editorial Pearson – Prentice Hall. 2004



7. Ian Goodfellow. Yoshua Bengio y Aaron Courville, "Deep Learning", MIT Press, 2016
8. Tom M. Mitchell. Machine Learning. WCB McGraw-Hill, 1997.
9. Hernández Orallo, Ramírez Quintana, Ferri Ramírez. Introducción a la Minería de Datos. Prentice Hall. 2004.