SP-3068: Fundamentos para el aprendizaje de máquinas (aprendizaje computacional)

Nombre del Programa: Programa de Posgrado en Ciencias Cognoscitivas
Plan de Estudios al que pertenece el curso: Maestría Académica
Modalidad: Teórico-práctico
Tipo de entorno: Presencial
Horas semanales: 3
Profesor que lo imparte: Marcelo Araya Salas PhD (marcelo.araya@ucr.ac.cr; sitio web: https://marce10.github.io/)
Sitio web del curso: https://marce10.github.io/aprendizaje_computacional_2024

Justificación

El aprendizaje computacional, también conocido como aprendizaje de máquinas o automático o "machine learning," es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y hacer predicciones para tomar decisiones basadas en datos. En lugar de seguir instrucciones explícitas para realizar una tarea, estos algoritmos identifican patrones en los datos y usan estos patrones para mejorar su desempeño en tareas específicas. El aprendizaje computacional se utiliza en una amplia variedad de aplicaciones, desde el reconocimiento de voz hasta el análisis de grandes conjuntos de datos y la automatización de procesos industriales. En este curso los estudiantes podrán conocer los fundamentos y las técnicas básicas del aprendizaje computacional para responder preguntas de investigación en Ciencias Cognoscitivas. Al iniciar el curso, se espera que las y los estudiantes conozcan los aspectos básicos de la estadística descriptiva e inferencial y el manejo básico del lenguaje de programación y análisis estadístico R.

Objetivo General
Capacitar a los estudiantes en los fundamentos, historia y diversas aplicaciones del aprendizaje de máquinas en el contexto de las Ciencias Cognoscitivas.

Objetivos Específicos
Al finalizar el curso, los estudiantes deberán ser capaces de:

- Describir y explicar los conceptos y métodos principales del aprendizaje computacional aplicados a las Ciencias Cognoscitivas.
- Diferenciar entre los tipos de aprendizaje computacional: supervisado, no supervisado, semi-supervisado y por reforzamiento.
- Comprender los fundamentos teóricos y los supuestos de técnicas como la regresión, las redes neuronales, los árboles de decisión y el análisis de conglomerados.
- Implementar, en el lenguaje de programación R, las técnicas desarrolladas a lo largo del curso.
- Seleccionar la técnica más adecuada en función del problema práctico o pregunta de investigación en las diversas áreas de las Ciencias Cognoscitivas.

Contenidos

1. **Introducción a R para el Aprendizaje Computacional**

- Instalación de R y RStudiob. Estructuras de datos básicas (vectores, matrices, listas)

- Manipulación de data frames y matrices
- Visualización básica de datos
- Estructuras de control (if, else, for, while)
- Creación de funciones personalizadas

2. **Simulación de Datos con Patrones Predefinidos**

- Generación de datos con distribuciones específicas
- Creación de datos con correlaciones
- Generación de datos categóricos y con ruido controlado

3. **Aprendizaje Supervisado: Regresión**

- Regresión lineal simple: conceptos y aplicaciones
- Regresión múltiple: modelado con múltiples variables
- Interacciones entre variables
- Regresión con variables categóricas (variables ficticias)

4. Aprendizaje Supervisado: Clasificación

- Introducción a la clasificación
- Árboles de decisión: construcción e interpretación
- Random Forest
- Máquinas de soporte vectorial (SVM)
- Evaluación de modelos de clasificación (matriz de confusión, precisión, recall, F1)

5. **Aprendizaje No Supervisado: Clustering**

- Conceptos de clustering y su importancia
- Método k-means
- Clustering jerárquico
- Evaluación de clusters (índice de Silhouette, coeficiente de Rand)

6. **Reducción de Dimensionalidad**

- Introducción a la reducción de dimensionalidad
- Análisis de Componentes Principales (PCA)
- Visualización en espacios reducidos
- Comparación con otros métodos como t-SNE

7. **Redes Neuronales y Deep Learning**

- Estructura de una red neuronal
- Entrenamiento: forward y backpropagation
- Redes neuronales convolucionales (CNNs)
- Redes neuronales recurrentes (RNNs) y LSTM
- Optimización y regularización

8. **Aprendizaje por Refuerzo**

- Conceptos y componentes del aprendizaje por refuerzo
- Diferencias con el aprendizaje supervisado
- Algoritmos básicos (Q-learning, política de aprendizaje)
- Aplicaciones en robótica y juegos

9. **Evaluación de Modelos**

- Validación cruzada
- Curvas ROC y AUC
- Análisis de errores
- Comparación y selección de modelos

10. **Regularización y Generalización**

- Conceptos de sobreajuste y subajuste
- Regularización (Lasso, Ridge, Elastic Net)
- Técnicas de ensamble (Bagging, Boosting, Stacking)
- Prácticas para evitar el sobreajuste

Métodos

Se utilizará la plataforma de Mediación Virtual de la Universidad de Costa Rica como un recurso complementario para el acceso a los *scripts*, la entrega de las tareas y la comunicación entre el profesor y el estudiantado. Las sesiones teóricas y prácticas del curso serán presenciales.

La metodología utilizada combina clases magistrales, demostraciones con software de algunos de los procedimientos de análisis de datos, y ejercicios prácticos que los y las estudiantes resolverán. Cuando se realicen los ejercicios prácticos, los resultados se discutirán en la clase para propiciar la interacción entre docente y estudiantes.

Evaluación

Tareas (60%)
Cada estudiante, de manera individual, deberá realizar 3 tareas con un valor de 20% cada una. En estas tareas, el estudiantado pondrá en práctica los conocimientos y destrezas abordados durante las sesiones presenciales del curso. Las tareas están diseñadas para evaluar la comprensión y aplicación de los conceptos teóricos y prácticos vistos en clase. Las tareas pueden incluir, pero no se limitan a:

- Implementación de algoritmos de aprendizaje computacional en R.
- Análisis de datos utilizando técnicas de visualización y preprocesamiento.
- Resolución de problemas específicos utilizando modelos predictivos y de clasificación.
- Comparación y evaluación de diferentes métodos de aprendizaje computacional.

Trabajo Final (40%)
El trabajo final tiene un valor de 40% y consiste en un proyecto integrador donde los estudiantes aplicarán todos los conocimientos adquiridos durante el curso para resolver un problema práctico. A continuación se detalla la estructura y los requisitos del trabajo final:

---

1. Selección del Problema (10%)

- Los estudiantes deben elegir un problema de investigación o un caso práctico relevante en el contexto de las Ciencias Cognoscitivas.

- El problema seleccionado debe permitir la aplicación de técnicas de aprendizaje computacional vistas en el curso.

2. Recopilación y Preprocesamiento de Datos (10%)

- Recopilación de un conjunto de datos adecuado para abordar el problema seleccionado.

- Realización de un preprocesamiento completo de los datos, incluyendo limpieza, transformación y visualización inicial.

3. Desarrollo e Implementación de Modelos (40%)

- Selección de al menos tres técnicas de aprendizaje computacional diferentes para abordar el problema.

- Implementación de los modelos seleccionados en R, con una explicación detallada de los supuestos y parámetros utilizados.

- Evaluación comparativa de los modelos implementados utilizando métricas adecuadas (e.g., precisión, recall, F1-score, RMSE).

4. Análisis y Discusión de Resultados (20%)

- Interpretación de los resultados obtenidos de cada modelo.

- Discusión sobre las ventajas y desventajas de cada técnica aplicada en el contexto del problema seleccionado.

- Identificación de posibles mejoras y futuras direcciones de investigación.

5. Presentación y Reporte Final (20%)

- Elaboración de un reporte escrito que incluya introducción, metodología, resultados, discusión y conclusiones.

- Preparación de una presentación oral de 10-15 minutos donde se expongan los principales hallazgos y se responda a preguntas del instructor y compañeros.

El trabajo final permitirá a los estudiantes demostrar su capacidad para integrar y aplicar de manera crítica y creativa los conceptos y técnicas aprendidas a lo largo del curso. Además, les proporcionará una experiencia práctica que refuerce sus habilidades en el uso de R y en la resolución de problemas reales mediante el aprendizaje computacional.

Contenidos por sesión (3 horas semanales)

Día 1: Introducción a R para el Aprendizaje Computacional - Parte 1

Introducción al lenguaje R, incluyendo su instalación, entorno de desarrollo y estructuras de datos básicas.

1. Instalación de R y RStudio.
2. Navegación en el entorno de RStudio.
3. Estructuras de datos básicas: vectores, matrices y listas.
4. Operaciones básicas y funciones en R. **Práctica con R:** Crear y manipular vectores y data frames. Uso de funciones básicas.
   Paquetes: R básico.

Día 2: Introducción a R para el Aprendizaje Computacional - Parte 2

Continuación de la introducción a R, con manipulación de datos, técnicas básicas de visualización, y control de flujo.

- Manejo de data frames y matrices.
- Filtrado, ordenamiento y resumen de datos.
- Visualización básica de datos con gráficos base.
- Creación de gráficos personalizados. **Práctica con R:** Manipulación de data frames, gráficos simples y personalizados.
  Paquetes: R básico.

Día 3: Introducción a R para el Aprendizaje Computacional - Parte 3

Profundización en estructuras de control de flujo y creación de funciones en R.

- Uso de estructuras de control: if, else, for, while.
- Creación de funciones personalizadas.
- Aplicación de funciones a estructuras de datos.
- Buenas prácticas de programación en R. **Práctica con R:** Escribir y utilizar funciones y loops para manipular datos.
  Paquetes: R básico.

Día 4: Simulación de Datos con Patrones Predefinidos

Técnicas para generar conjuntos de datos con características específicas y patrones conocidos.

- Introducción a la simulación de datos y su importancia.
- Generación de datos con distribuciones específicas (normal, uniforme, etc.).
- Creación de datos con correlaciones y estructuras de dependencia.
- Generación de datos categóricos y con ruido controlado. **Práctica con R:** Simulación de conjuntos de datos para diferentes casos de estudio.
  Paquetes: R básico.

Día 5: Aprendizaje Supervisado: Regresión Lineal Simple

Introducción a los modelos de regresión lineal simple para la predicción y análisis de relaciones entre variables.

- Conceptos básicos de regresión lineal simple.
- Ajuste de modelos y interpretación de coeficientes.
- Evaluación del modelo: R^2 y error cuadrático medio (MSE).
- Diagnóstico de supuestos y multicolinealidad. **Práctica con R:** Construcción y evaluación de modelos de regresión lineal simple.
  Paquetes: R básico.

Día 6: Aprendizaje Supervisado: Regresión Múltiple y Avanzada

Extensión de la regresión a múltiples variables y manejo de interacciones.

- Introducción a la regresión múltiple.
- Modelado con variables categóricas mediante variables ficticias.

- Interacciones entre variables y su interpretación.
- Evaluación y diagnóstico de modelos avanzados. **Práctica con R:** Análisis de regresión múltiple e interacciones.
  Paquetes: R básico.

Día 7: Aprendizaje Supervisado: Clasificación

Conceptos de clasificación y modelos básicos.

- Introducción a la clasificación y problemas de clasificación.
- Árboles de decisión: construcción e interpretación.
- Máquinas de soporte vectorial (SVM): principios básicos.
- Medidas de evaluación: matriz de confusión, precisión, recall, F1-score. **Práctica con R:** Implementación de modelos de clasificación y evaluación de desempeño.
  Paquetes: R básico.

Día 8: Aprendizaje No Supervisado: Clustering

Técnicas de agrupamiento para descubrir estructuras ocultas en los datos.

- Conceptos básicos de clustering y su importancia.
- Método k-means: algoritmos y aplicaciones.
- Clustering jerárquico: construcción de dendrogramas.
- Evaluación de clusters: índice de Silhouette y coeficiente de Rand. **Práctica con R:** Realización de análisis de clustering en datos de ejemplo.
  Paquetes: R básico.

Día 9: Reducción de Dimensionalidad

Métodos para reducir la complejidad de los datos.

- Introducción a la reducción de dimensionalidad y su necesidad.
- Análisis de Componentes Principales (PCA): teoría y aplicación.
- Visualización de datos en espacios reducidos.
- Comparación con otros métodos como t-SNE. **Práctica con R:** Aplicación de PCA y visualización de resultados.
  Paquetes: R básico.

Día 10: Redes Neuronales y Deep Learning - Introducción

Introducción a los modelos de redes neuronales.

- Estructura de una red neuronal: neuronas, capas y activación.
- Entrenamiento de redes neuronales: forward y backpropagation.
- Introducción a Deep Learning y redes neuronales profundas.
- Aplicaciones y casos de uso en el mundo real. **Práctica con R:** Construcción de una red neuronal simple.
  Paquetes: neuralnet.

Día 11: Redes Neuronales y Deep Learning - Aplicaciones Avanzadas

Exploración de redes neuronales profundas y sus aplicaciones avanzadas.

- Capas convolucionales y redes neuronales convolucionales (CNNs).
- Redes neuronales recurrentes (RNNs) y LSTM.
- Técnicas de optimización y regularización.
- Casos de estudio y aplicaciones en visión por computadora y procesamiento de lenguaje natural.
  Práctica con R: Implementación de una CNN básica y una RNN.
  Paquetes: neuralnet.

Día 12: Aprendizaje por Refuerzo

Conceptos y aplicaciones del aprendizaje por refuerzo.

- Definición y componentes del aprendizaje por refuerzo.
- Diferencias entre aprendizaje supervisado y por refuerzo.
- Algoritmos básicos: Q-learning y política de aprendizaje.
- Aplicaciones en robótica y juegos. **Práctica con R:** Simulación de un entorno de aprendizaje por refuerzo básico.
  Paquetes: R básico.

Día 13: Evaluación de Modelos

Técnicas para evaluar la precisión y efectividad de los modelos.

- Validación cruzada y división de conjuntos de datos.
- Curvas ROC y AUC: interpretación y uso.
- Análisis de errores y ajuste de modelos.
- Comparación y selección de modelos. **Práctica con R:** Evaluación y comparación de modelos con diferentes métricas.
  Paquetes: R básico.

Día 14: Regularización y Generalización

Métodos para mejorar la capacidad de generalización de los modelos.

- Conceptos de sobreajuste y subajuste.
- Regularización: Lasso, Ridge y Elastic Net.
- Técnicas de ensamble: Bagging, Boosting y Stacking.
- Prácticas para evitar el sobreajuste en modelos complejos. **Práctica con R:** Aplicación de técnicas de regularización y ensamble.
  Paquetes: R básico.

Día 15 y 16: Presentaciones del Proyecto Final

Presentacion del proyecto en formato que muestre codigo y texto.ección según el proyecto.

Bibliografía de Referencia

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An Introduction to Statistical Learning with Applications in R.** Un texto fundamental que cubre una amplia gama de técnicas de aprendizaje estadístico y computacional, con ejemplos prácticos en R. Enlace: https://www.statlearning.com

Goodfellow, I., Bengio, Y., & Courville, A. (2016). **Deep Learning.** Este libro es una referencia esencial para entender los conceptos fundamentales y avanzados de las redes neuronales y el deep learning. Enlace: https://www.deeplearningbook.org

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2021). **An Introduction to Statistical Learning.** Proporciona una introducción accesible a los métodos estadísticos y de aprendizaje computacional con aplicaciones prácticas en R. Enlace: https://www.statlearning.com

Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Un recurso exhaustivo que cubre una amplia gama de métodos de aprendizaje estadístico y computacional, incluyendo teoría y aplicaciones. Enlace: https://hastie.su.domains/ElemStatLearn/

Shalev-Shwartz, S., & Ben-David, S. (2014). **Understanding Machine Learning: From Theory to Algorithms.** Una introducción teórica sólida a los principios y algoritmos del aprendizaje computacional. Enlace: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

Murphy, K. P. (2012). **Machine Learning: A Probabilistic Perspective.** Ofrece una visión integral del aprendizaje computacional desde una perspectiva probabilística, con numerosos ejemplos y ejercicios prácticos. Enlace: https://mitpress.mit.edu/books/machine-learning

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Un libro esencial que abarca desde los fundamentos hasta las técnicas avanzadas en reconocimiento de patrones y aprendizaje computacional. Enlace: https://www.springer.com/gp/book/9780387310732

Wickham, H., & Grolemund, G. (2017). **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.** Aunque centrado en tidyverse, es útil para comprender el flujo de trabajo en R y cómo aplicar principios de limpieza y visualización de datos. Enlace: https://r4ds.had.co.nz

Alpaydin, E. (2020). **Introduction to Machine Learning.** Un recurso introductorio que cubre los conceptos clave y algoritmos en el aprendizaje computacional, accesible para estudiantes con poca o ninguna experiencia previa. Enlace: https://www.mitpress.mit.edu/books/introduction-machine-learning

Kuhn, M., & Johnson, K. (2013). **Applied Predictive Modeling**. Este libro proporciona una guía práctica para la construcción de modelos predictivos utilizando R, con un enfoque en la preparación de datos y la selección de modelos. Enlace: https://www.springer.com/gp/book/9781461468486