Taller 1: Analítica de datos 2021-2

Profesor: Andrés Marino Álvarez Meza, Ph.D.
Departamento de ingeniería eléctrica, electrónica y computación
Universidad Nacional de Colombia - sede Manizales

1. Instrucciones

- El taller debe ser enviado al correo electrónico `amalvarezme@unal.edu.co` desde su correo institucional (no se aceptarán envios desde correos diferentes a `@unal.edu.co`) incluyendo las discusiones y desarrollos sobre celdas de texto en latex de Colaboratory y códigos en Python (celdas de código comentadas y discutidas sobre Colaboratory), referente a los ejercicios propuestos.
- Las secciones 2 y 3 deben desarrollarse de forma individual. La sección 4 debe desarrollarse en los grupos del proyecto de curso (máximo hasta 3 personas). Respecto al punto 4, enviar desde un solo correo los desarrollos, indicando los integrantes del grupo de trabajo.
- **Fecha máxima de entrega: 24 de noviembre de 2021.**

2. Conceptos básicos en ciencia de datos y preprocesamiento

- Según lo discutido en el material de apoyo introducción al aprendizaje de máquina, discuta las principales ventajas y desventajas del aprendizaje estadístico vs. aprendizaje por reglas impuestas.
- Cuáles son las tareas básicas del aprendizaje de máquina?. Realice un cuadro comparativo indicando el dominio y rango de la función de aprendizaje a encontrar en cada caso.
- Consultar y realizar los ejercicios propuestos en el cuaderno de Colab Introducción a Numpy.

3. Aprendizaje de máquina en tareas de regresión

- Consultar y realizar los ejercicios propuestos en el cuaderno de Colab Guía lado a lado en aprendizaje de máquina.
- Consultar y realizar los ejercicios propuestos en el cuaderno de Colab Regresión no lineal y búsqueda de hiperparámetros.
- Consultar y realizar los ejercicios propuestos en el cuaderno de Colab Comparación métodos de regresión.

4. Anteproyecto - aplicación en analítica de datos

- Escoger una base de datos de al menos 1000 instancias y 100 atributos, relacionada con alguna de las tareas básicas en aprendizaje de máquina (clasificación, regresíon o conglomerados). Se sugiere revisar los repositorios Kaggle, UCI machine learning, y Google Dataset search.
- Realice una análisis exploratorio utilizando Pandas de los atributos y la salida de la base de datos (histogramas, boxplot, matriz de correlación de Pearson, scatter matrix). Discuta las relaciones encontradas entre las entradas, y entre las entradas y la salida.
- Realizar un estudio del estado del arte, de al menos 3 artículos científicos (papers) de los últimos cuatro años, relacionados con la base de datos escogida, describiendo potenciales aplicaciones, desafíos en el preprocesamiento, tipos de algoritmos de aprendizaje de máquina utilizados, métricas dde desempeño, etc. Consulte en revistas Q1/Q2 (utilizar el buscador de revistas de Scimago journal ranking para corroborar la categoría de la revista). Se sugiere utilizar el buscador de bases de datos de la UNAL y Google scholar para encontrar artículos científicos relevantes.

Referencias

https://github.com/amalvarezme/AnaliticaDatos

Géron, A., (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Bishop, C. M. (2006). Pattern recognition. Machine learning, 128(9).